Herausgeber

M. HEIZMANN
T. LÄNGLE
F. PUENTE LEÓN

FORUM BILDVERARBEITUNG 2016

KIT Scientific Publishing

M. Heizmann | T. Längle | F. Puente León (Hrsg.)

FORUM BILDVERARBEITUNG 2016

FORUM BILDVERARBEITUNG 2016

Herausgegeben von
M. Heizmann, T. Längle und F. Puente León

KIT Scientific Publishing

Impressum

Scientific Publishing

Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe

KIT Scientific Publishing is a registered trademark of Karlsruhe Institute of Technology. Reprint using the book cover is not allowed.

www.ksp.kit.edu

Print on Demand 2016

ISBN 978-3-7315-0587-7
DOI 10.5445/KSP/1000059899

Vorwort

Bildverarbeitung spielt in vielen Bereichen der Technik zur schnellen und berührungslosen Datenerfassung eine Schlüsselrolle. Beispielsweise in der Qualitätssicherung industrieller Produktionsprozesse, in der Robotik und zur Fahrerassistenz haben sich Bildverarbeitungssysteme einen unverzichtbaren Platz erobert. Diese Entwicklung wird unterstützt durch die Verfügbarkeit qualitativ hochwertiger und günstiger Sensorsysteme sowie durch die Zunahme der Leistungsfähigkeit von Rechnersystemen.

Das „Forum Bildverarbeitung“ hat das Ziel, über aktuelle Trends zu allgemeinen und robusten Lösungen in der Bildverarbeitung zu berichten und zum fachlichen Austausch zwischen den Teilnehmern beizutragen. Die Beiträge wurden vom Programmausschuss ausgewählt. Sie umfassen die folgenden Schwerpunkte:

- Bildgewinnung,
- Simulation,
- Klassifikation und Bewertung,
- 3D-Erfassung,
- Mensch und Medizin sowie
- Industrielle Anwendungen.

Das Forum Bildverarbeitung möchte einen Beitrag zur Weiterentwicklung der Bildverarbeitung als wichtige und zukunftsträchtige Fachdisziplin leisten. Es richtet sich an Fachleute, die sich in der industriellen Entwicklung, in der Forschung oder der Lehre mit Bildverarbeitungssystemen befassen, und bietet eine Plattform für den Wissens- und Erfahrungsaustausch zwischen Wissenschaftlern und Anwendern.

Dezember 2016 M. Heizmann, T. Längle und F. Puente León

Wissenschaftliche Leitung

Prof. Dr.-Ing. M. Heizmann	Karlsruher Institut für Technologie
Prof. Dr.-Ing. T. Längle	Fraunhofer-IOSB Karlsruhe
Prof. Dr.-Ing. F. Puente León	Karlsruher Institut für Technologie

Programmausschuss

Prof. Dr. C. Bach	NTB, CH-Buchs
Prof. Dr.-Ing. J. Beyerer	Fraunhofer IOSB Karlsruhe
Dr. rer. nat. J. Burke	Fraunhofer IOSB Karlsruhe
Prof. Dr. K. Donner	Universität Passau
Dr. rer. nat. J. Eggert	Honda Research Institute, Offenbach
Prof. Dr. A. Heinrich	Hochschule Aalen
Prof. Dr. B. Jähne	Universität Heidelberg
Dr.-Ing. M. Kruse	ITK Engineering AG, Rülzheim
Dipl.-Ing. M. Maurer	Vitronic Dr.-Ing. Stein GmbH
Prof. Dr. R. Neubecker	Hochschule Darmstadt
Prof. Dr. W. Osten	Universität Stuttgart
Prof. Dr. F. Salazar Bloise	Universidad Politécnica de Madrid
Dipl.-Ing. M. Stelzl	Schott AG, Mainz
Prof. Dr.-Ing. C. Stiller	Karlsruher Institut für Technologie
Prof. Dr.-Ing. R. Tutsch	Technische Universität Braunschweig
Dr.-Ing. S. Werling	Fraunhofer IOSB Karlsruhe
Dipl.-Ing. S. Wienand	ISRA VISION AG, Darmstadt
Dr.-Ing. V. Willert	Technische Universität Darmstadt

Inhaltsverzeichnis

Klassifikation und Bewertung

3D-Erfassung

Mensch und Medizin

Industrielle Anwendungen

A fractal calibration pattern for improved camera calibration

Hendrik Siedelmann, Maximilian Diebold, Marcel Gutsche,
Hamza Aziz-Ahmad and Bernd Jähne

Heidelberg University, Heidelberg Collaboratory for Image Processing (HCI),
Berliner Str. 43, 69120 Heidelberg

Abstract Camera calibration, crucial for computer vision tasks, often relies on planar calibration targets to calibrate the camera parameters. This work explores a planar, fractal, self-identifying calibration pattern, which provides a high density of calibration points for a large range of magnification factors. An evaluation on ground truth data shows the target provides very high accuracy over a wide range of conditions.

Keywords Camera calibration, high accuracy, passive target, image geometry, image measurement, fiducial marker, self-identifying, image localization.

1 Introduction

Common methods for calibrating camera systems utilize passive targets with checkerboard patterns, which can be fabricated by printing. Such targets are easy to detect by a large range of available software, like the OpenCV library [1].

In computer vision tasks it is already possible to locate image features with an accuracy down to a few hundredths of a pixel, for example in light-field measurements [2]. However, the camera calibration, critical in describing the relation between the world and the observed features, commonly achieves an accuracy not much below a tenth of a pixel, also visible in our evaluation, see figure 4(b). Thus a lot of the accuracy cannot be exploited due to the lack of a suitable calibration.

Regarding the calibration target we can find several reasons for the limited accuracy. If a pattern is used without identifying marks, it is

impossible to determine pattern coordinates, if parts of the pattern are outside the field of view of the camera. This forces the user to place the calibration pattern away from the borders and therefore induces a bias towards the center of the imaging area. To avoid this it is necessary to use a self-identifying marker pattern [3,4].

The size of the calibration markers also plays a key role in the accuracy of the localization. If too few calibration points are available per view, then calibration quality suffers due to over fitting. However, if the calibration markers become too small, it becomes impossible to properly detect them. As the optimal size depends on the camera magnification, it is not possible to create a pattern with a single optimal density.

This work introduces a fractal calibration target, published under an open source license [5], which addresses this problem by providing calibration markers at several scales, allowing the adaption to the camera magnification at the time of detection. This is a first step towards a more sophisticated camera calibration, lifting accuracy to the next level suitable for highly accurate measurements.

2 Related work

The following overview is limited to the subject of calibration point localization. For passive targets there are two main types of localization features, checkerboard corners and circular markers.

For the checkerboard localization three main principles have been established. The first method uses line intersection, where lines are refined individually and the intersection of two lines defines the calibration points [3]. The second approach uses functional descriptions of saddle points [6–8], where the local neighborhood is approximated by using a 2D polynomial in which the corner point can be derived analytically. In comparison to line intersection, the saddle point method is more suited to smaller neighborhoods and at higher resolution achieves either no improvements [8] or even worse results [7].

The third method makes use of orthogonal gradients as implemented for example by OpenCV [1]. It exploits the property that the vector from the desired corner point to any edge pixel is orthogonal to the gradient vector in this pixel. This provides better results compared to line inter-

section as shown by Atcheson et al. [3]. Note that line intersection also suffers from distortion bias [7].

For circular markers the modeling of the transformation from a perfect circle to the observed image plays the largest role. Mallon et al. [7] evaluate localization performance on real and simulated images, under blur, noise, perspective and distortion. They conclude that circular patterns are dominated by perspective and distortion bias and recommend, for circular patterns, the usage of small (10 px) circles and the correction according to the distortion model. Consequently Datta et al. [9] as well as Douxchamps and Chihara [10] use an iterative refinement process which alternates between updating the camera model and refinement of the calibration marker image location. While this gives high quality results this approach inherently links the camera calibration with the marker refinement which makes the process more slow and inflexible compared to a two step approach where marker refinement and camera calibration are separate.

3 Design

Regarding the marker size, there are two possible directions. Either few large, highly accurate markers, as used by Douxchamps and Chihara [10], or many small markers of individually low accuracy. We argue that the underlying measurement accuracy is similar, as it depends solely on the quantity and magnitude of image gradients.

The localization of larger markers is dominated by perspective and radial distortion [7], which require a more complex localization approach for accurate results. While perspective bias is easy to compensate for, even without doing a full calibration, to correct distortion bias the marker localization has to incorporate the radial distortion [7, 9, 10], which ties marker refinement directly into the calibration. Thus it is impossible to separate marker detection and camera calibration which also makes it difficult to adapt new calibration methods to such a combined calibration scheme. This means that small features are better suited for a generic calibration pattern, as the individual calibration features are not affected much by bias. In addition, smaller markers can be placed closer to the image borders, reducing center bias.

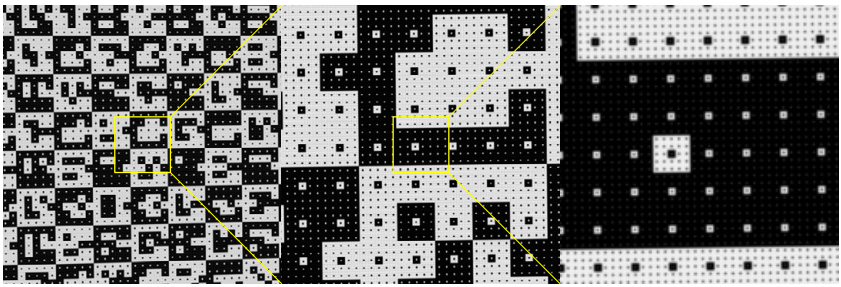


Figure 1: A view of the fractal calibration target. From left to right the magnification is increased by a factor of 5 at each step. Note how individual calibration dots resolve to square features when increasing the magnification.

3.1 Fractal layout

For the reasons exposed in section 3 we try to minimize the calibration point size. However, a pattern which is optimal in one given situation can become completely useless if the magnification is decreased, because closely spaced features will blend into each other. As the effective scale at which target features are projected onto the image sensor changes with distance, angle and radial distortion, it would therefore be necessary to capture a range of targets, each optimal at a different scale, and later fuse the calibration information from multiple targets.

Our solution is the adoption of a fractal scheme which operates on multiple scales of calibration points. This scheme allows to use features which are within some fixed bound of the maximum calibration point density for the given magnification. Figure 1 shows the calibration pattern at several magnifications.

The used calibration points are individual square dots. Due to the fractal nature of the pattern the finest resolvable calibration points are always at the pixel scale where several independent degradations, like physical pixel aperture, lens aberrations, diffraction, etc., render the calibration point as a Gaussian like 2D distribution. When a calibration point becomes large enough that the non-Gaussian characteristics come to bear, it is already possible to resolve the next recursion layer, making more complex refinement methods unnecessary. In addition, this elegantly avoids the bias problems of large markers, as the fractal nature

of the calibration pattern always allows the detection of relatively small markers, with a size close to the resolution limit.

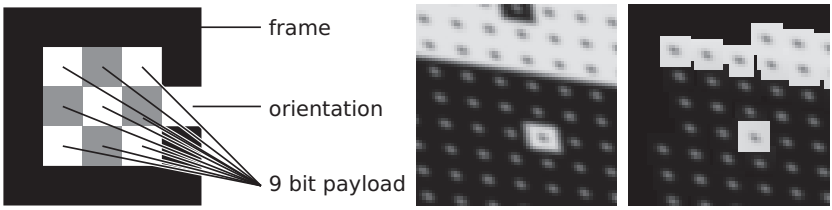
A new layer of calibration points is generated from a coarser layer by resizing the pattern with a factor of 5, using nearest neighbor interpolation. In the scaled pattern new calibration points can be inserted at every fifth pixel by simply inverting the respective pixel, see figure 1 for a visualization.

4 Implementation

The detection of the fractal pattern is split into two parts, the actual detection using the payload for identification and the fractal refinement. Detection takes place in scale space, using an x-corner detector followed by brute force iteration of possible markers using neighboring corner candidates.

4.1 Identification

It is important that parts of the pattern are allowed to be out of view, to give the freedom to cover the whole imaging area with the calibration target, and place calibration points close to edges and corners. Our method builds on square blocks with uniform borders and a payload of 9 bit in the center, see figure 2(a). The orientation is fixed using an opening in the border and markers are arranged in a checkerboard manner. This means that borders are either black or white. The coding scheme combines the payload of black and white markers to provide 18 bits of payload for addressing and always requires the detection of at least two neighboring black and white markers. The address is encoded using an XOR mask to provide a bit of randomization and to keep the overall distribution of bits more uniform. In contrast to other fiducial marker systems, like for example CALTag [3], the payload within the markers does not provide error detection or correction. Instead the address of neighboring markers is compared to check whether they are consistent, which provides very robust error detection but no error correction. This means the marker grid can have a maximum size of 512×512 markers which results in 262144 markers overall. For more details on the identification scheme please see [11].



(a) Marker schematic (b) Input image (c) Fitted points

Figure 2: (a) shows the schematic of a single marker. The constant color border with the hole at one side allows detection and fixes the orientation. In (b) the top left corner of a calibration image is shown and the fitted calibration points are visualized (c). Only the calibration points, not the marker borders are detected in the recursion scheme, which explains the jagged edges. Note how the recursion scheme is able to detect calibration points less than 10 pixels from the image border.

4.2 Fractal refinement

The fractal refinement starts after markers have successfully been detected. Expected positions of the first layer of calibration points are estimated from their respective marker using a perspective transform. Those positions are then refined using a least squares solver which fits a rotated 2D anisotropic Gaussian with a linear gradient as background. Individual pixels are used as samples in the fit, with a weighting relative to the initially estimated position. This step is repeated once more to improve accuracy by using the updated center position.

The recursion works by repeating this procedure until the size of the calibration points become too small for a meaningful fit, which is less than 4 pixels across. Various heuristics are used to verify the validity of the fit, like rejecting calibration points for which either the residual of the fit is too large, or for which the expected accuracy is low due to low contrast, steep background gradient, too small or too large width of the Gaussian, or due to saturation.

4.3 Bayer pattern mode

Because individual pixels are directly used as data points in the fit, it is possible to leave out individual samples. For color images acquired using a color filter array, demosaicing can be avoided by processing colors independently of each other, simply leaving out the pixels which belong to a different color channel.

5 Evaluation

The evaluation is based on rendered images corrupted with Gaussian blur, additive Gaussian noise, uneven illumination, reduced contrast and radial distortion. We compare classic checkerboard detection with subpixel refinement using the implementation from OpenCV with our fractal calibration pattern. The detected calibration points are then used for either a full camera calibration, again using the implementation from OpenCV, or alternatively to solve only for the camera pose under known camera intrinsics.

5.1 Error metric

In the literature cited in section 2, there are two principal methods of evaluating calibration patterns, the root mean square (rms) error of the calibration model and the rms of calibration points against ground truth, called true pixel error (TPE) by Douxchamps and Chihara [10]. The calibration rms is a poor choice, as it will not detect bias but simply incorporate it into the model. The TPE works well for example for the evaluation of different refinement methods of corner points, but it cannot be used for a meaningful comparison of calibration targets where the number of calibration points is not roughly the same. The reason is that more calibration points of similar localization performance obviously lead to better calibration results which is not reflected in the TPE.

A further method, which is often seen, is to compare triangulated world coordinates of calibration points or the extrinsics of the calibrated cameras. While these are valid metrics for some applications, they will also miss problematic areas like the corners of the image, if there were no calibration points present for that area. This is especially problem-

atic, because the calibrated model will be bad specifically in those areas where no calibration points were acquired.

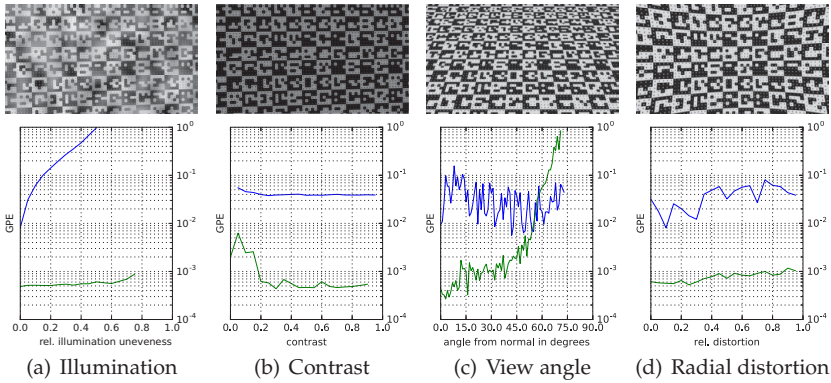
However, with ground truth data, it is possible to compare the projection defined by the fitted camera model with the known ground truth projection. This gives an error measure which relates to the actual camera calibration and incorporates the error distribution and bias of the used calibration pattern, as well as averaging from multiple markers. We compare the camera models in image space, by projecting each pixel into the scene, at the depth of the target, using the ground truth model and then projecting them back into image space with the calibrated model. The root means squared difference between the original pixel coordinate and the projection of the calibrated model defines the error metric, which we call Groundtruth Pixel Error (GPE). The GPE measures an error in pixel units, stating the localization accuracy of the tested pattern in the context of the tested calibration model. The metric gives a score averaged over the whole image, including the edges and corners, which are the most problematic areas in the context of camera calibration, both because they exhibit the strongest distortion and because it is difficult to place calibration markers on the edges of the image.

5.2 Ground truth data

The evaluation images were rendered with the cycles renderer from Blender [12] using 1024 samples per pixel, with Gaussian anti aliasing at a resolution of 1920 x 1080. For distortion simulation the images were rendered with the resolution increased by a factor of four and scaled down after performing the distortion in order to reduce the influence of the interpolation.

5.3 Comparison

For reference, a checkerboard with 12×6 squares is detected using the checkerboard detection from OpenCV with subpixel refinement over an area of 21 pixels. When evaluating the performance under blur and noise we noticed that under most circumstances the refinement was dramatically improved by performing a Gaussian blur with a sigma of 2 before refinement, which is therefore used in the whole comparison.



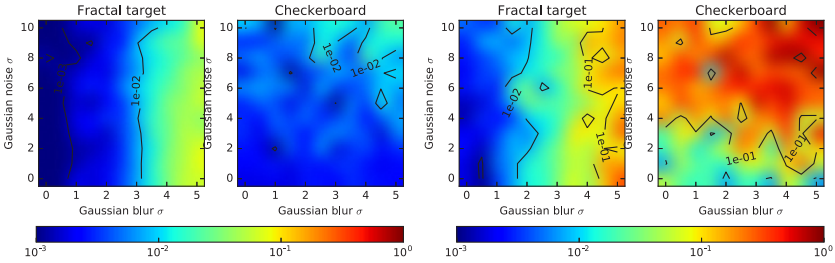
(a) Illumination (b) Contrast (c) View angle (d) Radial distortion

Figure 3: Performance comparison under different conditions. The top row shows an example of the respective deterioration, the bottom row the evaluation result. The blue line shows the checkerboard detection, the green line the recursive target. The only case where the checkerboard detection surpasses the recursive target is under shallow angles more than 55 degrees from the target normal.

5.4 Results

In most cases the results of the fractal target surpass the accuracy of the checkerboard pattern by one to two orders of magnitude. In general the calibration is very robust to noise, reduced contrast, uneven illumination and radial distortion, see figure 3 and 4. Two areas where the method is not as robust and where it is eventually surpassed by the checkerboard detection are detection under shallow angles and under strong blur. The reasons and possible enhancements are discussed in section 6.

One reason for the introduction of the new target was the reduction of center bias due to the difficulty to observe calibration points in the corner. To evaluate center bias, the GPE is measured for the four corners and the center of the images, see figure 4. While the bias is not completely removed it is greatly reduced with the corner results having an GPE approxmately five times that of the center, while for the checkerboard this factor is between 10 to 100.



(a) Center GPE (b) Corner GPE

Figure 4: Comparison of center and corner quality. The plots show the reduction of center bias by our dense target. The improvements in the corner are much more pronounced compared to the center, with more than two orders of magnitude improvement in some areas. This effect can be attributed to the fact that the recursive target always covers the corners which is not the case for the checkerboard target.

6 Discussion

While the recursive target delivers most of the anticipated advantages, with highly improved general accuracy, it fails to completely remove center bias and it performs badly under strong blur and shallow view angles.

Regarding the center bias, it seems a dense target is not enough to completely solve the problem. A possible solution could be to introduce a weighting scheme into the calibration method which weights corner samples stronger, however more work is required to estimate good weighting factors and the influence on the overall calibration results.

The second shortcoming, low robustness under shallow view angles and strong blur, is a direct result of the desired characteristic of uncoupling the pattern detection from the calibration process. As Mallon and Whelan point out [7], most patterns need to incorporate the calibration within the pattern refinement to remove perspective and distortion bias from the pattern localization. While the checkerboard refinement does not suffer from this drawback, the 2D Gaussian fits used in our target do have this problem. The workaround is to use very small calibration

points where this bias is very small, which works for most cases as the fractal structure ensures that the smallest calibration points can be used independently from magnification. This fails only if the smallest scale cannot be used for another reason than the image scale, which are blur and strong perspective distortion, which happens at shallow view angles.

Reliable correction of this bias needs the calibration information which would tie the pattern detection to the camera calibration. However this would reduce the universality of the calibration pattern which in the current form can be used completely independent of the camera calibration.

7 Conclusions

For computer vision tasks that require a high accuracy, the calibration target can be a limiting factor in the camera calibration. As we have shown in this work, the use of a fractal calibration target can dramatically improve accuracy and reduce center bias. The target was evaluated and is quite robust under a range of imaging conditions. With an accuracy between a hundredth and a thousandth of a pixel, the target provides an excellent basis for the development of new calibration models, as deficiencies in the camera model are much more apparent with more precise raw calibration data, as provided by the fractal target. Due to the small features size the pattern detection can be completely decoupled from the camera calibration, which simplifies development of new calibration methods. To support further research our implementation is available under an open source license [5].

8 Acknowledgment

The work was carried out during a research cooperation between the Computational Imaging Group at the Stuttgart Technology Centre of Sony Europe Limited and the Heidelberg Collaboratory for Image Processing (HCI). We would like to thank in particular Alexander Gatto from Sony Stuttgart for his feedback and fruitful discussions.

References

1. Itseez, "Open source computer vision library," https://github.com/itseez/opencv, 2015.
2. M. Diebold, O. Blum, M. Gutsche, S. Wanner, C. Garbe, H. Baker, and B. Jähne, "Light-field camera design for high-accuracy depth estimation," in *SPIE Optical Metrology*. International Society for Optics and Photonics, 2015, pp. 952 803–952 803.
3. B. Atcheson, F. Heide, and W. Heidrich, "CALTag: High Precision Fiducial Markers for Camera Calibration," in *Vision, Modeling, and Visualization (2010)*, R. Koch, A. Kolb, and C. Rezk-Salama, Eds. The Eurographics Association, 2010.
4. S. Daftry, M. Maurer, A. Wendel, and H. Bischof, "Flexible and user-centric camera calibration using planar fiducial markers." in *BMVC*, 2013.
5. H. Siedelmann, "A high density fractal calibration pattern," http://hci-repo.iwr.uni-heidelberg.de/hsiedelm/hdmarker, 2016.
6. L. Lucchese and S. K. Mitra, "Using saddle points for subpixel feature detection in camera calibration targets." in *APCCAS (2)*. IEEE, 2002, pp. 191–195.
7. J. Mallon and P. F. Whelan, "Which pattern? biasing aspects of planar calibration patterns and detection methods," *Pattern recognition letters*, vol. 28, no. 8, pp. 921–930, 2007.
8. S. Placht, P. Fürsattel, E. A. Mengue, H. G. Hofmann, C. Schaller, M. Balda, and E. Angelopoulou, "Rochade: Robust checkerboard advanced detection for camera calibration." in *ECCV (4)*, ser. Lecture Notes in Computer Science, D. J. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8692. Springer, 2014, pp. 766–779.
9. A. Datta, J.-S. Kim, and T. Kanade, "Accurate camera calibration using iterative refinement of control points," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1201–1208.
10. D. Douxchamps and K. Chihara, "High-accuracy and robust localization of large control markers for geometric camera calibration." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 376–383, 2009.
11. H. Siedelmann, "Recording, Compression and Representation of Dense Light Fields," Diplomarbeit, Universität Stuttgart, Fakultät Informatik, Elektrotechnik und Informationstechnik, Germany, Januar 2015.
12. Blender Online Community, *Blender – a 3D modelling and rendering package*, Blender Foundation, Blender Institute, Amsterdam, 2016. [Online]. Available: http://www.blender.org

Aufbau einer Kamera mit programmierbarer Apertur durch Nutzung eines transmissiven Flüssigkristalldisplays

Construction of a programmable aperture camera using a transmissive liquid crystal display

Thomas Nürnberg und Fernando Puente León

Karlsruher Institut für Technologie
Institut für Industrielle Informationstechnik
Hertzstraße 16, 76187 Karlsruhe

Zusammenfassung Kameras mit codierten Blenden kommen in zahlreichen Gebieten zur Anwendung. Die hohe Anzahl sowie die Abhängigkeit der optimalen Systemparameter von den Umgebungsbedingungen erfordern die Entwicklung flexibler Kamerasysteme. Dieser Beitrag beschreibt den Aufbau einer Kamera mit programmierbarer Apertur. Dazu wird ein transparentes Flüssigkristalldisplay in ein Objektiv integriert. Erste Untersuchungen zeigen, dass damit nach einer Korrektur von Störanteilen ein gutes Abbildungsverhalten erreicht werden kann.

Schlagwörter Bildverarbeitung, Computational Imaging, Blendencodierung.

Abstract Cameras with coded apertures are applicable in many different scenarios. The high amount of applications and the dependency of the optimal system parameters on the ambient conditions require the development of flexible camera systems. This article presents the construction of a programmable aperture camera. We integrate a transmissive liquid crystal display into the objective. A first analysis indicates promising imaging capabilities, after the extraneous light has been compensated.

Keywords Image processing, computational imaging, coded apertures.

1 Einleitung

Licht im Raum kann allgemein durch die plenoptische Funktion beschrieben werden, die von insgesamt sieben Dimensionen abhängt: den Raumkoordinaten x, y und z, den Winkeln der Ausbreitungsrichtung ϕ und θ, der Wellenlänge λ sowie der Zeit t. Eine klassische Kamera erfasst davon lediglich zwei Ortsdimensionen und wenige Wellenlängenbereiche zu einem Zeitpunkt t. Daher geht der derzeitige Trend bei der Entwicklung neuartiger Kameras anstelle des bloßen Vorantreibens des Auflösungsvermögens der bereits erfassbaren Dimensionen immer mehr dazu über, die Information anderer Dimensionen zu erfassen. Diese neuartigen Ansätze werden unter dem Begriff *Computational Imaging* zusammengefasst. Bei all diesen Ansätzen findet noch vor der Digitalisierung des Bildsignals durch den Sensor eine optische Signalverarbeitung durch eine problembezogene Anpassung der Kameraoptik oder die Einbringung zusätzlicher Komponenten in den Lichtweg statt. *Computational Imaging* kann also als Erweiterung der Signalverarbeitungskette auf die physikalische Domäne angesehen werden.

Die Blende regelt den Lichteinfall in eine Kamera und ist damit von maßgeblicher Bedeutung für das abbildende System. Ihre Form definiert die verschiebungsvariante Punktverschmierungsfunktion (PSF, engl. *Point Spread Function*) und legt damit das Abbildungsverhalten der Kamera fest. Damit ist die Blende eine naheliegende Einflussgröße für neuartige Ansätze des *Computational Imagings*.

Neben der Optimierung des Abbildungsverhaltens bei gewünschten Ortsfrequenzen sind codierte Blenden für zahlreiche Anwendungen geeignet. Durch die Wahl einer möglichst breitbandigen Übertragungsfunktion lässt sich Unschärfe im Bild durch eine Entfaltung kompensieren. Eine im Spektralbereich eindeutig identifizierbare Blendenform erlaubt die Extraktion von Tiefeninformation auf Grundlage zweier nacheinander aufgenommener Bilder [1] oder sogar eines Einzelbildes [2–4]. Durch eine Blende mit impulskammförmigem Spektrum [4] oder durch ein zeitliches Multiplexing [5] der Blendenform lässt sich das vierdimensionale Lichtfeld, also Orts- und Winkelinformation, erfassen. Eine Codierung der Blendenform in Zeitrichtung ermöglicht die Reduktion von Bewegungsunschärfe [6].

Der Umfang dieser – keineswegs vollständigen – Auswahl an Anwendungsmöglichkeiten codierter Blenden zeigt bereits, dass für einen

universellen Einsatz Kameras mit programmierbaren Blendenformen notwendig sind. Dadurch wird sowohl die Anpassung der Kamerafunktion auf die geforderte Anwendung als auch die Anpassung der Form an die konkret vorliegenden Umgebungsbedingungen, wie das Störszenario, ermöglicht. Abbildung 1 zeigt den vereinfachten Strahlengang einer solchen Kamera mit der programmierbaren Aperturblende hinter der Linse.

Nach einer Übersicht über bisherige Ansätze zur wechselbaren oder programmierbaren Realisierung einer Aperturblende in Abschnitt 2 wird in Abschnitt 3 der Aufbau und die Inbetriebnahme einer *Programmable Aperture Camera* mit einem Flüssigkristalldisplay beschrieben. Die Ergebnisse werden abschließend in Abschnitt 4 vorgestellt.

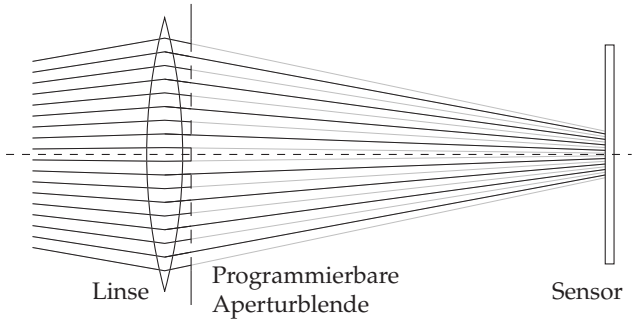


Abbildung 1: Idealisierter Strahlengang einer Kamera mit programmierbarer Apertur mit einer Linse.

2 Stand der Technik

Um die Blendenform einer Kamera zu ändern, kann die Blende mechanisch gewechselt werden. Dies ist jedoch entweder zeitaufwendig oder erfordert zusätzliche Wechselvorrichtungen [5]. Darüber hinaus ist eine derartige Realisierung zum einen anfällig für Verschmutzungen und zum anderen für geringfügige Änderungen der Blendenposition, die eine Neukalibrierung der Kamera notwendig machen. Bei der Nutzung eines fest montierten, programmierbaren, optischen Elements treten die beschriebenen Nachteile nicht auf.

Fujikake et al. [7] entfernen mit einem auf Flüssigkristallen (LC, engl. *Liquid Crystal*) basierenden adaptiven Polarisationsfilter vor der Kamera an dielektrischen Oberflächen reflektierte und damit polarisierte Lichtanteile. Nayar und Branzoi [8] dämpfen das einfallende Licht mit einem LC-basierten *Spatial Light Modulator* zur Aufnahme eines *High Dynamic Range*-Bildes. Zomet und Nayar [9] nutzen mehrere LC-Displays zur dynamischen Wahl der Sichtfelds einer linsenlosen Kamera. Zur Bestimmung des Lichtfelds aus einer Bildserie verwenden Liang et al. [5] ein Folienband, das mit unterschiedlichen Blendenformen bedruckt ist. Das Band wird durch einen Spalt zwischen Objektiv und Kamera geführt. Die Autoren beschreiben außerdem einen Prototyp mit einem LC-Display als Blende, das jedoch nur eine Auflösung von 7×7 Pixeln aufweist. Manami et al. [10] nutzen ein reflektierendes LC-Display in Kombination mit einem halbdurchlässigen Spiegel, um eine programmierbare Apertur zu realisieren. Durch die Umlenkung des Lichts senkrecht zur Einfallsrichtung ist jedoch keine kompakte Bauform möglich. Nagahara et al. [11] nutzen denselben Aufbau mit einer LCoS-Apertur (engl. *Liquid Crystal on Silicon*).

3 Aufbau einer Kamera mit programmierbarer Apertur

In diesem Abschnitt wird der Aufbau sowie die Inbetriebnahme einer Kamera mit programmierbarer Apertur beschrieben.

3.1 Objektivumbau

Als programmierbare Blende wird das transmissive LC-Display aus Abbildung 2 verwendet. Es handelt sich um ein DOGM128S-6-Display von Electronic Assembly, dessen Eigenschaften in Tabelle 1 angegeben sind.

Aufgrund der Funktionsweise eines LC-Displays wird die Intensität des einfallenden – im Allgemeinen unpolarisierten – Lichts allein durch die Polarisationsfilter um $\eta_{\mathrm{Pol}} = 50\,\%$ gedämpft. Durch die nichtideale Transmission des Flüssigkristalls und die Verdrahtung der Pixel wird der Wirkungsgrad des Displays weiter verringert ($\eta_{\mathrm{LCD}} = \eta_{\mathrm{Pol}} \cdot \eta_{\mathrm{LC}} \cdot \eta_{\mathrm{Füll}}$). Mit einem experimentell bestimmten Transmissionsgrad im lichtdurchlässigen Zustand (LCD hell) von $\eta_{\mathrm{LCD,hell}} = 26{,}7\,\%$ ergibt sich der Transmissionsgrad des Flüssigkristalls der programmier-

baren Blende zu $\eta_{\mathrm{LC,hell}} = 57{,}4\ \%$. Aufgrund dieser starken Dämpfung sollte das Objektiv mit einer besonders lichtempfindlichen Kamera verwendet werden. Im lichtundurchlässigen Zustand (LCD dunkel) gelangen immer noch $\eta_{\mathrm{LCD,dunkel}} = 3{,}0\ \%$ des auftreffenden Lichts durch das Display. Dieser Störanteil muss nachträglich entfernt werden.

Das Display wird in ein handelsübliches AF Nikkor 50 mm 1:1,8D-Objektiv eingebaut. Durch die Verwendung eines Objektivs mit Festbrennweite und ohne Motor zum Autofokus wird der Einbau erleichtert, da der grundlegende Aufbau klarer ist und wenige bewegliche Komponenten verbaut sind. Aufgrund der großen Blendenöffnung kann die wirksame Fläche des Displays optimal ausgenutzt werden und Blendenformen lassen sich mit feinerer örtlicher Auflösung realisieren.

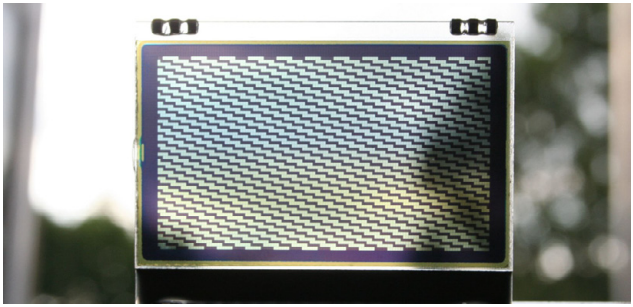


Abbildung 2: Transmissives Flüssigkristalldisplay EA DOGM128S-6 vor dem Einbau in das Objektiv.

Tabelle 1: Parameter des EA DOGM128S-6-Displays.

Technologie	FSTN negativ transmissiv
Auflösung	128×64
Pixelgröße	$375\ \mu\mathrm{m} \times 435\ \mu\mathrm{m}$
Wertequantisierung	1 bit
Füllfaktor	93 %
Dicke	2 mm

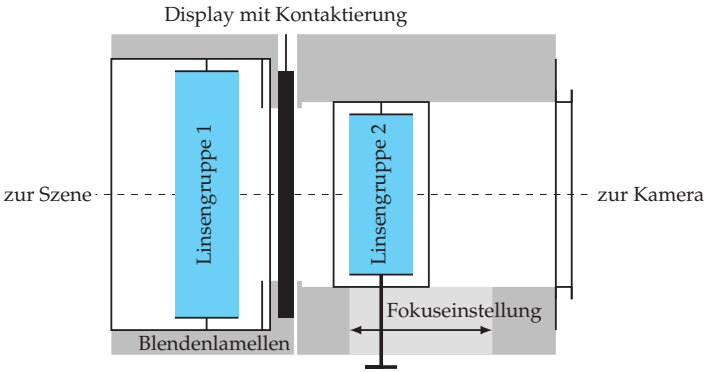


Abbildung 3: Schematischer Aufbau des umgebauten Objektivs. Die zweite Linsengruppe kann entlang der optischen Achse verschoben werden, um die Fokuseinstellung zu ändern.

Darüber hinaus kann durch eine größere Blendenöffnung mehr Licht auf den Sensor fallen, wodurch die Dämpfung durch das LC-Display teilweise kompensiert werden kann. Das Objektiv zeichnet sich außerdem durch einen vergleichsweise geringen Preis aus.

Der schematische Aufbau des umgebauten Objektivs ist in Abbildung 3 dargestellt. An der Stelle, an der sich die klassische Blende befindet, wird das Objektiv aufgeteilt. Die Lamellen der klassischen Blende werden fixiert. Das Display wird in einem Einlass vor der bildseitigen Linsengruppe, also so nah wie möglich an der ursprünglichen mechanischen Blende, platziert und durch einen Spalt kontaktiert. Die beiden Objektivkomponenten werden durch zwei Kunststoffmanschetten fixiert, die mit zwei Schrauben fest miteinander verbunden werden. An der sensorseitigen Manschette wird der Metallring des F-Bajonetts befestigt, wodurch sich das Objektiv an gängige Kameras anschließen lässt. Die elektrischen Kontakte des Objektivs am Anschluss werden entfernt, da sie nach dem Umbau keine Funktion mehr erfüllen. Abbildung 4 zeigt das umgebaute Objektiv.

Die sensorseitigen Linsengruppen werden zur Einstellung des Fokus beweglich angebracht. Die Abstände zum Metallring des F-Bajonetts sollten dabei entsprechend dem ursprünglichen Objektiv gewählt werden, um denselben Fokusbereich zu erreichen.

Abbildung 4: Umgebautes Objektiv mit F-Bajonett (links) und Blick entlang optischer Achse mit angesteuertem LC-Display (rechts).

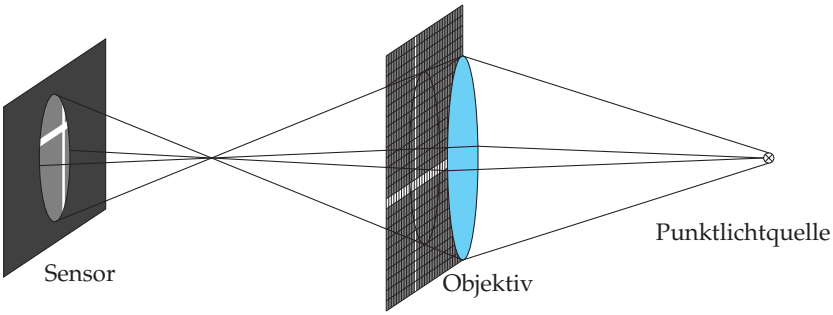


Abbildung 5: Schema der Anordnung zur Bestimmung der nutzbaren Displayfläche und der Punktverschmierungsfunktion.

3.2 Kalibrierung

Nach dem Einbau des Displays ist eine Bestimmung der Position des Displays im Lichtweg des Objektivs notwendig. Dazu wird zunächst, wie in Abbildung 5 dargestellt, die PSF aufgenommen, indem eine punktförmige Lichtquelle außerhalb des Fokus abgebildet wird. Zur Bestimmung der Pixel, die sich tatsächlich im Lichtweg befinden, werden horizontale und vertikale, lichtdurchlässige Streifen auf dem Display angezeigt und nacheinander verschoben.

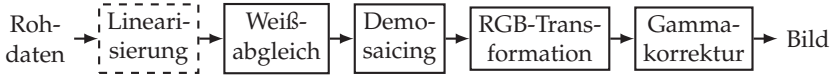


Abbildung 6: Signalverbeitung der Sensorrohdaten bis zum darstellbaren Bild für Farbkameras mit Bayer-Sensor (nach Sumner [13]). Abhängig von der verwendeten Kamera ist der gestichelte Schritt nicht notwendig.

3.3 Bildkorrekturen

Das LC-Display erreicht im dunklen Zustand keine vollständige Abschattung, wie etwa bei Verwendung einer mechanischen Blende. Fremdlicht kann zwischen den Pixeln oder aufgrund von Beugungseffekten an den Verdrahtungen zum Sensor gelangen. Darüber hinaus ist die Dämpfung lichtundurchlässig geschalteter Pixel nur unvollständig. Daher ist den Aufnahmen stets ein Störanteil aus dem Fremdlicht überlagert. Zur Kompensation kann der Störanteil separat aufgenommen werden, indem dieselbe Szene bei vollständig dunklem Display abgebildet wird. Anschließend wird das so gewonnene Störbild vom ursprünglichen Bild abgezogen [5]. Da die gemessenen Intensitäten des Sensors proportional zur Bestrahlungsstärke sind [12], bleibt bei dieser Kompensation die Energie des Nutzsignals erhalten. Allerdings ist der Prototyp wegen der zwei benötigten Aufnahmen nur zur Abbildung statischer Szenen geeignet.

Bei der Kompensation des Störanteils muss sichergestellt sein, dass die Intensität des Bildes linear mit den gemessenen Intensitäten am Sensor zusammenhängen. In der Signalverarbeitungskette der Rohdaten bis zum darstellbaren Bild aus Abbildung 6 muss die Kompensation also nach der Linearisierung und vor der Gammakorrektur geschehen.

Oftmals liefern Farbkameras die Daten für den Weißabgleich. Für den Fall, dass das verwendete LC-Display im dunklen Zustand eine starke spektrale Varianz der Dämpfung aufweist, muss der Weißabgleich nach der Kompensation gegebenenfalls nochmals durchgeführt werden.

4 Ergebnisse

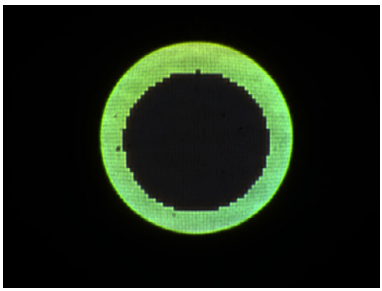
Zur Erprobung wurde das Objektiv mit einer handelsüblichen Spiegelreflexkamera verbunden, wodurch die Kamera zu einer Kamera mit

(a) Aufnahme mit Objektivprototyp. (b) Aufnahme mit Referenzobjektiv.

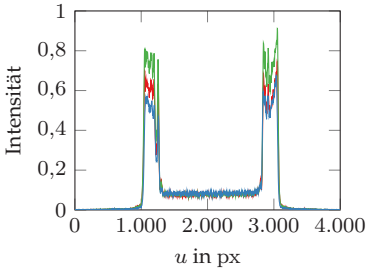
Abbildung 7: Vergleich des Objektivprototyps mit Referenzobjektiv.

programmierbarer Apertur wird. Auch nach dem Umbau ist das Objektiv weiterhin in der Lage, Szenen mit zufriedenstellender Qualität abzubilden, wie der Vergleich der Aufnahmen des Prototyps mit denen eines Referenzobjektivs aus Abbildung 7 verdeutlicht. Die erreichte Bildschärfe des Referenzobjektivs ist geringfügig besser als die des Prototyps. An der PSF, die im linken Bildrand durch die unscharfe Abbildung einer Punktlichtquelle sichtbar ist, sind leichte Verzerrungen und Strukturen durch die Verkabelung des Displays zu erkennen. Die Abbildungsfehler (Koma) aufgrund des schrägen Lichteinfalls am Bildrand sind gleich stark ausgeprägt

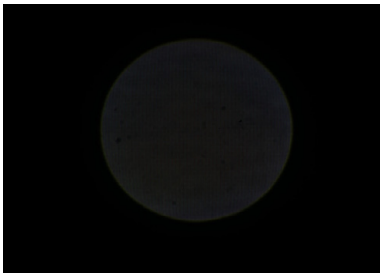
Der mit dem Objektiv erreichbare Fokusbereich wurde bestimmt, indem, bei gleichbleibender Fokuseinstellung und Beleuchtung, Bildserien eines Kalibriermusters in unterschiedlichen Entfernungen aufgenommen wurden. Anhand des Gradientenbetrags der Kanten im Bild können die Entfernungen ermittelt werden, die bei der nächsten bzw. der fernsten Objektiveinstellung scharf abgebildet werden. Der so ermittelte Entfernungsbereich des Prototyps reicht von 18 cm bis 1,8 m. Dieser geringe Einstellungsbereich ist auf eine zu große Bildweite des umgebauten Objektivs zurückzuführen, was in einem nächsten Prototyp durch eine Verkürzung der bildseitigen Kunststoffmanschette verbessert werden könnte.



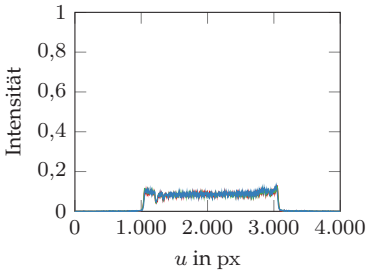
(a) Aufnahme mit überlagertem Störanteil.



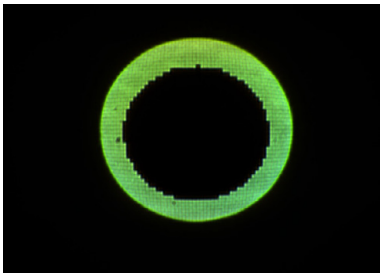
(b) Intensitätsverlauf der Aufnahme.



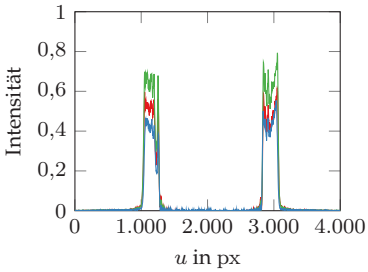
(c) Aufnahme des Störanteils.



(d) Intensitätsverlauf des Störanteils.



(e) Kompensierte Aufnahme.



(f) Intensitätsverlauf ohne Störanteil.

Abbildung 8: Veranschaulichung der Entfernung des Störanteils anhand der Aufnahme der PSF des Objektivs.

Abbildung 8(a) zeigt eine Aufnahme der PSF der Prototyps entsprechend der Anordnung aus Abbildung 5. Das auf dem LC-Display angezeigte ringförmige Blendenmuster ist unmittelbar in der aufgenommenen PSF erkennbar. In der Bildmitte und insbesondere im Intensitätsverlauf des horizontalen Querschnitts aus 8(b) lässt sich der Störanteil im dunklen Displaybereich erkennen. Die Abbildungen 8(c) und 8(d) veranschaulichen den mit derselben Anordnung isoliert aufgenommenen Störanteil, der in der kompensierten Aufnahme 8(e) und 8(f) subtrahiert wurde.

Der Störanteil der dunklen Displayflächen wirkt sich gleichmäßig auf die drei Farbkanäle aus (Abbildung 8(d)). Daher wirkt sich dessen Kompensation nur geringfügig auf die Farbbalance der Aufnahme aus. Folglich kann auf einen erneuten Weißabgleich verzichtet werden.

5 Zusammenfassung

Eine Kamera mit programmierbarer Apertur erlaubt eine flexible und anwendungsspezifische Wahl der Blendenform und damit der Übertragungsfunktion der optischen Abbildung. Beim in diesem Beitrag beschriebenen Prototyp wurde die Blende durch ein transmissives LC-Display, das in ein Objektiv integriert wurde, realisiert. Während sich störende Lichteinteile durch eine separate Aufnahme kompensieren lassen, wird durch die baubedingte Dämpfung des LC-Displays, auch im lichtdurchlässigen Zustand, nur eine verringerte Lichtintensität am Sensor erreicht. Auch nach dem Einbau des Displays konnten die guten Abbildungseigenschaften des Objektivs erhalten bleiben, lediglich der Fokusbereich sollte durch eine erneute Anfertigung der bildseitigen Objektivfassung angepasst werden.

In fortführenden Untersuchungen bleibt zu klären, welche Ergebnisse sich in konkreten Anwendungen, wie beispielsweise der Tiefenextraktion, mit dem Prototyp erzielen lassen.

Literatur

1. A. Levin, „Analyzing depth from coded aperture sets“, in *Proceedings of the 11th European conference on Computer vision: Part I*. Springer-Verlag, 2010, S. 214–227.

2. A. Levin, R. Fergus, F. Durand und W. T. Freeman, „Image and depth from a conventional camera with a coded aperture“, *ACM Transactions on Graphics*, Vol. 26, Nr. 3, 2007.
3. T. Nürnberg, C. Zimmermann und F. Puente León, „Simulationsgestützte Optimierung einer Computational-Kamera zur dichten Tiefenschätzung“, *tm – Technisches Messen*, Vol. 83, Nr. 9, S. 511–520, 2016.
4. A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan und J. Tumblin, „Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing“, *ACM Transactions on Graphics*, Vol. 26, Nr. 3, S. 69–1–69–12, 2007.
5. C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu und H. H. Chen, „Programmable aperture photography: Multiplexed light field acquisition“, *ACM Transactions on Graphics*, Vol. 27, Nr. 3, S. 55:1–55:10, 2008.
6. R. Raskar, A. Agrawal und J. Tumblin, „Coded exposure photography: Motion deblurring using fluttered shutter“, *ACM Trans. Graph.*, Vol. 25, Nr. 3, S. 795–804, 2006.
7. H. Fujikake, K. Takizawa, T. Aida, T. Negishi und M. Kobayashi, „Video camera system using liquid-crystal polarizing filter to reduce reflected light“, *IEEE Transactions on Broadcasting*, Vol. 44, Nr. 4, S. 419–426, 1998.
8. S. K. Nayar und V. Branzoi, „Adaptive dynamic range imaging: Optical control of pixelexposures over space and time“, in *IEEE International Conference on Computer Vision (ICCV)*, Vol. 2, 2003, S. 1168–1175.
9. A. Zomet und S. K. Nayar, „Lensless imaging with a controllable aperture“, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, 2006, S. 339–346.
10. H. Mannami, R. Sagawa, Y. Mukaigawa, T. Echigo und Y. Yagi, „Adaptive dynamic range camera with reflective liquid crystal“, *Journal of Visual Communication and Image Representation*, Vol. 18, Nr. 5, S. 359–365, 2007.
11. H. Nagahara, S. Kuthirummal, C. Zhou und S. K. Nayar, „Flexible depth of field photography“, in *European Conference on Computer Vision (ECCV)*, 2008.
12. J. Beyerer, F. Puente León und C. Frese, *Automatische Sichtprüfung: Grundlagen, Methoden und Praxis der Bildgewinnung und Bildauswertung*, 2. Aufl. Berlin Heidelberg: Springer, 2016.
13. R. Sumner, *Processing RAW Images in MATLAB*, University of California, Santa Cruz, 2014.

Individualisierung optischer Messtechnik für In-Line-Applikationen basierend auf additiven Fertigungstechnologien

Individualized optical metrology for in-line applications based on additive manufacturing

Andre Sigel, Manuel Rank, Phillipe Maillard, Yannick Bauckhage, Sven Pekrul, Markus Merkel und Andreas Heinrich

Hochschule Aalen, Zentrum für optische Technologien, Beethovenstraße 1, 73430 Aalen

Zusammenfassung Im Bereich der additiven Fertigung konnten in den letzten Jahren rasante Fortschritte beobachtet werden. Diese Fortschritte ermöglichen die Fertigung von transmissiven und reflektiven Optikelementen im 3D-Druck. Der Fokus liegt hierbei auf schwer zu realisierenden Freiformoptiken, die spezifisch an messtechnische Problemstellungen angepasst werden. Dieser Artikel liefert eine Übersicht an aktuellen Entwicklungen von additiv gefertigten, angepassten Sensorsystemen, die für Anwendungen in der In-Line-Fertigungskontrolle dimensioniert wurden.

Schlagwörter Additive Fertigung, 3D-Metalldruck, Freiformoptiken, strukturierte Beleuchtung, Optiksimulation.

Abstract In the field of additive manufacture rapid progress could be observed in recent years. Based on this progress transmissive and reflective optical elements can now be manufactured with 3D printers. Hereby the manufacturing process is focused on with classical technologies difficult to manufacture freeform optics, which are optimized for specific metrology tasks. This article summarizes recent developments of additive manufactured, optimized metrology systems designed for in-line inspection use.

Keywords Additive manufacturing, 3D metal printing, freeform optics, structured illumination, optics simulation.

1 Einleitung

In modernen Produktionsstraßen findet die Qualitätsüberprüfung der gefertigten (identischen) Bauteile mit Hilfe der sog. In-Line-Messtechnik statt. Hierbei handelt es sich um automatisierte Messeinheiten, die z. B. die Form eines Bauteils (x-, y-, z-Maße) mit Hilfe optischer Sensoren erfassen. Dabei kommen meist Standardlösungen wie z. B. Lasertriangulationssensoren, Fotogrammmetrie- oder Streifenprojektionssysteme zum Einsatz. Diese generellen Systeme sind jedoch nicht auf die individuellen Eigenschaften des momentan zu vermessenden Bauteils hin optimiert. So bedeutet dies im Falle der Geometrievermessung, dass das optische Abtastsignal (z. B. eine Laserlinie) nicht auf die Form des zu vermessenden Bauteils angepasst ist. Dies hat den Nachteil, dass z. B. Hinterschnitte vom Sensor wegen einer Schattenbildung nicht erfasst werden können. Auch die Messunsicherheit des messtechnischen Systems weist durch Deformationen des Abtastsignals aufgrund der Bauteilform Schwankungen auf. Ideal wäre es in diesem Fall, wenn die Laserlinie der kompletten Kontur des Bauteils ohne Unterbrechung mit konstanter Signalstärke folgen könnte. Dies bedeutet, dass das optische Abtastsignal individuell auf die Geometrie des zu vermessenden Objekts angepasst werden müsste. Die Konsequenz daraus ist allerdings, dass die optischen Komponenten jeweils neu individuell an die jeweilige Messaufgabe anzupassen sind, was unwirtschaftlich bzw. oftmals unmöglich ist.

Die wirtschaftliche Realisierung von individualisierten Komponenten in komplexer Form mit Losgröße 1 [1, 2] ist mit Hilfe der additiven Fertigung möglich. Die Entwicklung additiver Fertigungsmethoden ist in den letzten Jahren rasant vorangeschritten. Dabei fokussieren die meisten Arbeiten auf der Realisierung mechanischer Bauteile. Aber die additive Fertigungstechnologie bietet auch ein hohes Potenzial im Bereich der Optik, da mit ihr neue Designfreiheitsgrade und damit komplett neue Lösungsansätze möglich sind.

2 Additive Fertigungstechnologien in der Optik

Im Bereich der reflektiven Optiken werden additive Verfahren genutzt, bei denen Metallpulver im fokussierten Laserstrahl im sog. Selective La-

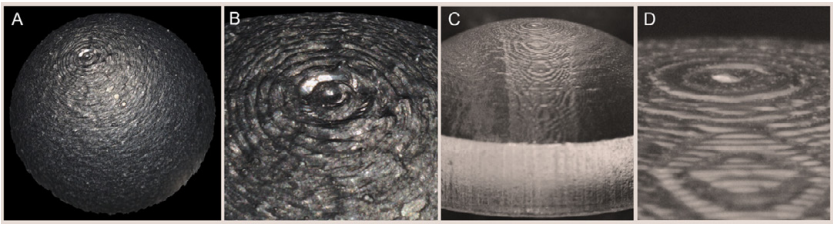


Abbildung 1: (A) Mit Keyence Mikroskop VR-3100 aufgenommene 3D-Darstellung eines SLM-Aluminiumprobekörpers (Drucker SLM Solutions 280HL) (B) Schichtenstrucktur des SLM-Probekörpers in 38facher Vergrößerung; Schichtdicke von 50 μm (C) mit Dunkelfeldbeleuchtung aufgenommene Schichtenstruktur einer durch MJM gefertigten Kunstharzoptik (Drucker Keyence Agilista 3000) (D) Schichtenstrucktur der MJM-Optik in 50facher Vergrößerung; Schichtdicke von 25 μm

ser Melting (SLM) schichtweise aufgeschmolzen wird. Ein Standardmaterial beim SLM-Druck ist AlSi10Mg. Neben Aluminium können jedoch auch Stahl-, Titan-, Kobalt- und Nickellegierungen gedruckt werden. In Abb. 1 (A) und (B) wird die Schichtenstruktur eines Aluminiumprobekörpers gezeigt.

Erste Untersuchungen [3–6] zeigen, dass die Fertigung von Spiegeloptiken mit dieser Technologie möglich ist. Durch unzureichende Rauheits- und Reflexionseigenschaften sind jedoch Nachbearbeitungsschritte nötig, um z. B. durch Sandstrahlen, Fräsen und Polieren [3–5] oder durch Diamantdrehen und Magnetorheological Finishing (MRF) [6] Oberflächen mit optischer Güte zu erzielen. Die Untersuchungen zeigen, dass die polierten Druckmaterialien zwar zum Teil geringere Reflexionswerte als herkömmliche Werkstoffe aufweisen [3, 4], jedoch können Hochpräzisionsflächen mit Rauheiten im Bereich 6-8 nm (RMS) gefertigt werden [6].

Transmissive Optiken werden durch additive Fertigungstechnologien auf Basis von Polymermaterialien gefertigt. Genutzte Verfahren sind Multi Jet Modeling (MJM) [7–10] und Stereolithographie (SLA) [9–11].

Beim MJM wird ein Polymer durch einen Druckkopf vergleichbar zum Tintenstrahldrucker lokal aufgebracht und schichtweise ausgehärtet. In Abb. 1 (C) und (D) wird die Schichtenstruktur einer durch den MJM-Drucker Keyence Agilista 3000 gedruckten Optik

gezeigt. Durch wasserlösliche Materialien können bei diesem Verfahren Stützelemente vermieden werden, indem ein abwaschbares Stützmaterial gedruckt wird.

Beim SLA-Verfahren wird das UV-aktive Material mit einem Projektor oder einem Laser direkt bestrahlt und so schichtweise ausgehärtet. Da nur ein Material genutzt werden kann, sind Stützstrukturen aus dem gleichen Material notwendig, deren Anschlussstellen an die Optik später durch einen Nachbearbeitungsvorgang entfernt werden müssen.

Auch bei der Fertigung transmissiver Optiken durch additive Technologien sind Nachbearbeitungsschritte notwendig. Zunächst wird die Rauheit der additiv gefertigten Optiken durch eine Politur reduziert. Es werden entweder Standardpoliturmaschinen [8] oder Roboterpoliturstationen [9,11] genutzt. Im Anschluss an den Politurvorgang kann die Oberfläche der Optikelemente noch weiter verbessert werden, indem eine an den Brechungsindex angepasste Schicht aufgebracht wird [9,10]. Die Kunststoffoptiken können entweder mit Klarlack oder mit Epoxidharz beschichtet werden. Beschichtungstechnologien sind das Tauch-, Spray-, und Rotationsbeschichten [9]. Durch Nachbearbeitungsschritte können minimale Rauheitswerte von 12 nm (RMS) erreicht werden [8].

3 Vermessung von Hinterschnitten mit reflektiven Optiken aus additiver Fertigung

In einem ersten Applikationsbeispiel für additiv gefertigte Sensorelemente in der Qualitätskontrolle wird ein System zur Vermessung von Hinterschnitten von Spritzgussteilen dargestellt. Der Innenraum eines Pumpenstators (Exzenterschneckenpumpe, Fa. Seepex, siehe Abb. 2 (A)) soll durch einen individualisierten Lichtschnittsensor (Sensorkonzept nach [12]) auf 3D-Defekte kontrolliert werden. Bei dieser Pumpe wird zum Zweck des Flüssigkeitsvortriebs eine bis zu 4 m lange, spiralförmige Innenkontur mit einem Innendurchmesser von 3 cm benötigt. Herkömmliche Sensorsysteme sind erstens zu groß, um sie in das Pumpenteil einzuführen, und zweitens nicht an die schwankende Querschnittsfläche der Spirale angepasst. Das Projektvorhaben wird durch die Baden-Württemberg Stiftung gGmbH finanziert.

Das realisierte Sensorsystem 1 wird nach Abb. 2 (B) in den Pumpenstator 2 eingeführt. Während des Einführvorgangs werden Punktedaten

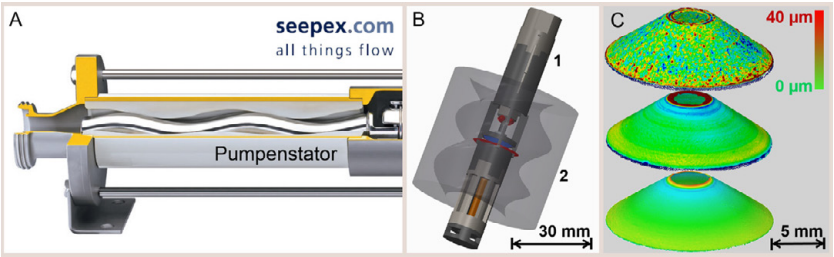


Abbildung 2: (A) Schnittbild eines zu vermessenden Pumpenstators der Fa. Seepex (B) Sensorkonzept mit Sensorelement und Prüfobjekt (C) additiv gefertigtes Optikelement in der Prozesskette mit farbig kodierten Oberflächenabweichungen: SLM-Druck (oben), geschliffen (Mitte), poliert (unten); aufgenommen mit Keyence Mikroskop VR-3100

von der innen liegenden Oberfläche generiert.

Zur Realisierung des miniaturisierten Sensorsystems wurde eine 16 mm durchmessende, konische Aluminiumoptik im SLM-Druck gefertigt und bis auf optische Güte nachbearbeitet. Die Optikelemente in der Prozesskette bestehend aus additiver Fertigung im SLM-Druck (Ra: 13 μm) sowie je einem Schleif- (Ra: 2 μm) und einem Poliervorgang (Ra: 15 nm) durch eine automatische Polierstation werden in Abb. 2 (C) gezeigt. Oberflächenabweichungen werden in der Abbildung farbig dargestellt.

Das Sensorsystem besteht aus einer Beleuchtungs- und einer Abbildungseinheit. In Abb. 3 (A) und (B) werden das Sensorsetup und Kernelemente dargestellt. Es wird eine rote Laserdiode (1) mit einem diffraktiv optischem Element (DOE) genutzt, um zunächst eine Abstrahlcharakteristik in Form eines Hohlkegels zu generieren. Durch die additiv gefertigte Optik (2) wird die Strahlung zu einer Lichtschnittebene (3) modelliert und das Bild der Laserlinie durch Miniaturkameras (4) erfasst. Mit einem Auswertealgorithmus wird die erfasste Laserlinie zu Punktedaten umgerechnet.

Das in Abb. 3 (C) gezeigte Kameramodul beinhaltet vier USB-CMOS-Kameras mit einer VGA-Auflösung von 640x480 Pixeln und einem Öffnungswinkel von 70°. Kamera und Optik haben jeweils einen Durchmesser von nur 3 mm. Der Sensor mit der durch die Beleuchtungskomponente generierten Laserlinie wird in Abb. 3 (D) gezeigt.

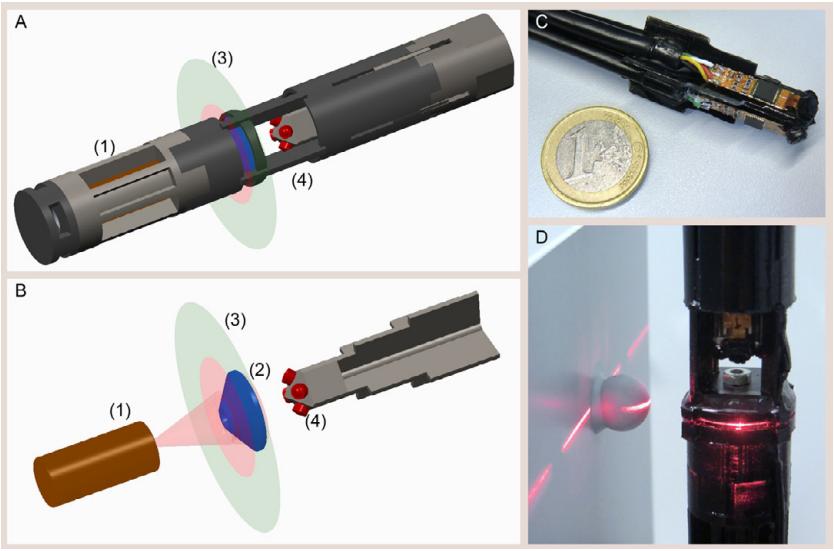


Abbildung 3: (A) CAD-Entwurf eines Hinterschnittsensors (B) Kernelemente des Hinterschnittsensors (C) Kameramodul des Hinterschnittsensors mit Größenvergleich (D) Sensorsystem im Betrieb: Vermessung eines sphärischen Kalibrierobjektes

Zur 3D-Datengewinnung durch ein Lichtschnittverfahren können Algorithmen genutzt werden, die auf intrinsischen Kameraparametern basieren (Bildhauptpunkt, Brennweite, Verzerrungskoeffizient, Skalierungsfaktor der Pixel) [13]. Diese können jedoch bei den genutzten Minikameras aufgrund der ungenügenden Abbildungsqualität nicht mit ausreichender Genauigkeit bestimmt werden. Aus diesem Grund wurde ein robustes Ersatzverfahren konzipiert, um auch ohne innere Kameraparameter 3D-Punktedaten zu generieren.

Bei dem Verfahren werden zunächst Verzeichnung und perspektivische Verzerrung durch ein Schachbrettmuster in der Lichtschnittebene ermittelt und kompensiert. Die metrische Kalibrierung wird anschließend durch einen Least Square Fit mit einem 3D-Kalibrierobjekt realisiert. Um die Kamerapositionen der 4 Miniaturkameras im Weltkoordinatensystem zu bestimmen, wird ein Referenzobjekt mit 12 Kugeln

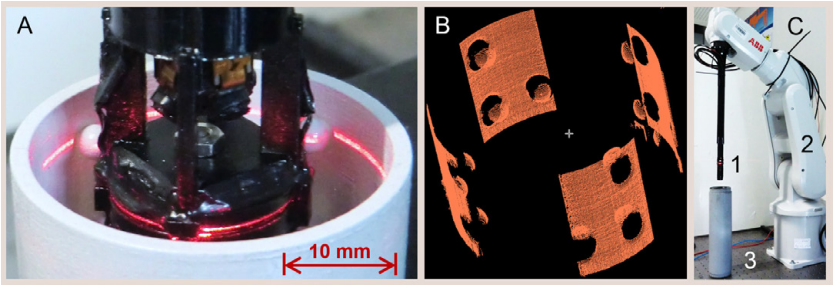


Abbildung 4: (A) Erfassung eines 3D-Targets zur Bestimmung von Kameraposen mit dem Hinterschnittsensor (B) durch den Sensor generierte 3D-Ausgabe des 3D-Targets nach der Kalibrierung (C) Integration eines Hinterschnittsensors in die Roboterkinematik ABB IRB 120

(Abb. 4 (A)) erfasst. Über eine Objekterkennung in der Punktewolke (Abb. 4 (B)) werden im nächsten Schritt die jeweils 3 Kugelpositionen innerhalb der 4 Kamerakoordinatensysteme bestimmt. Durch bekannte Kugelpositionen des Referenzobjektes ermöglicht dies eine Ausrichtung im Weltkoordinatensystem.

Der Sensor wurde letztendlich in eine Roboterkinematik integriert (Abb. 4 (C)). Das Sensorelement 1 wird durch den Roboter 2 in den Pumpenstator 3 eingeführt. Der Sensor kann freilaufend während einer Linearbewegung genutzt werden. Alternativ besteht auch die

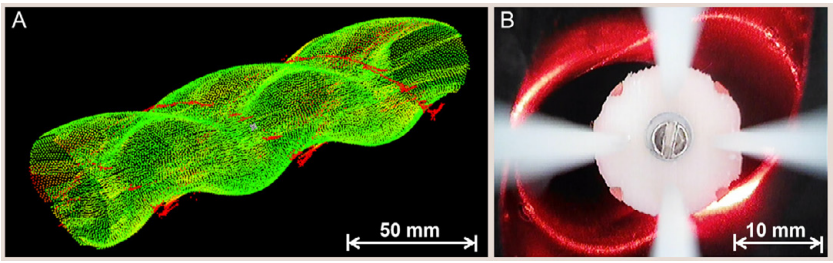


Abbildung 5: (A) Durch den Hinterschnittsensor generierte 12 cm lange Punktewolke der Leitapplikation (Pumpenstator) mit farbig kodierten Punkteabweichungen zum CAD-Modell (B) Einsicht in den Pumpenstator während eines Messvorgangs durch eine Miniatur-VGA-Kamera

Möglichkeit, den Sensor über ein LabView-Programm automatisiert an definierte Positionen zu fahren und Bildaufnahmen über den Roboter extern zu triggern.

In Abb. 5 (A) wird die durch den Sensor generierte Punkteausgabe der Leitapplikation dargestellt. Abweichungen zum CAD-Modell des Pumpenstators sind rot markiert. Es ist zu erkennen, dass entlang des Querschnittminimums ungültige Punktedaten generiert werden. Die Ursache für diese Abweichungen besteht in einer inhomogenen Ausleuchtung der 3D-Kontur (Abb. 5 (B)) durch das Optikelement.

Das bisher genutzte, additiv im SLM-Verfahren gefertigte Optikelement ist für einen zylindrischen Querschnitt optimiert. Um den Sensor an den elliptischen Querschnitt der Leitapplikation anzupassen, wird eine Modulation der Laserleistung in der Lichtschnittebene über Krümmungsradien einer Freiformoptik notwendig. In laufenden Arbeiten werden Freiformelemente simuliert und die Fertigung im SLM-Verfahren erprobt.

4 Vermessung von 3D-Oberflächen durch transmissive Elemente aus additiver Fertigung

In einem zweiten Applikationsbeispiel wird die Möglichkeit der Beleuchtung komplexer Geometrien mittels transmissiver Optiken am Beispiel der Vermessung eines Tannenbaums einer Turbinenschaufel erläutert (Abb. 6 (A)). Zum Zweck der Lichtschnittauswertung wird eine homogene, linienförmige Ausleuchtung des Tannenbaums ohne Ab-

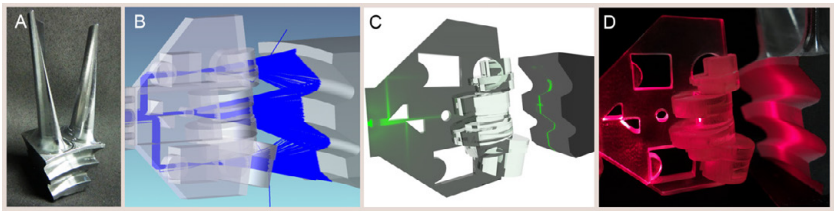


Abbildung 6: (A) Tannenbaum einer Turbinenschaufel (B) Zemax-Simulation des Beleuchtungsstrahlengangs einer Freiformoptik zur Generierung einer Laserlinie (C) photorealistisches Rendern eines virtuellen Kamerabildes durch die Software LightTools (D) additiv gefertigtes Optikelement in der Evaluierung

schattungseffekte benötigt. Durch die additive Fertigung der Beleuchtungseinheit können für jeden Teilabschnitt der Geometrie speziell optimierte Optiken entworfen und zu einer Einheit kombiniert werden, um eine ununterbrochene Linie konstanter Bestrahlungsstärke zu generieren. Der Vorteil der additiven Fertigung der Beleuchtungseinheit liegt darin, dass Geometrien ohne die Limitierungen herkömmlicher Fertigungsverfahren, wie bspw. Rotationssymmetrien oder fließende Übergänge, dimensioniert werden können.

Zur Simulation der komplexen optischen Elemente werden Optiksimulationsprogramme, sog. Raytracer, genutzt, um den Strahlengang innerhalb der Optik (Zemax, Abb. 6 (B)) durch eine auf Zielsetzungen (Laserlinienbreite) basierende Entscheidungsfunktion zu optimieren. Zur Anpassung der Beleuchtungseinheit an die Anforderungen des Sensorsystems kann durch Vorgabe von Objektiv und Sensor ein virtuelles Kamerabild gerendert werden (LightTools, Abb. 6 (C)). Basierend auf den Simulationsergebnissen werden die Optikelemente anschließend additiv gefertigt und evaluiert (Drucker Keyence Agilista 3000, Abb. 6 (D)).

5 Strukturierte Beleuchtung durch additiv gefertigte Optikelemente

Um ein definiertes Abtastsignal auf einer komplexen Geometrie zu generieren, werden optische Elemente benötigt, die durch refraktive Effekte Strahlung einer Quelle (z. B. einer Laserdiode) bündeln. Refraktive Optiken zur Generierung eines linienförmigen Abtastsignals können erstens durch Krümmungen von optischen Funktionsflächen erzeugt werden. Die zweite Möglichkeit besteht darin, simultan mehrere Materialien mit unterschiedlichen Brechungsindices im additiven Fertigungsprozess zu kombinieren [7] und Grenzflächen im Material zu erzeugen.

In Abb. 7 (A) bis (D) werden zwei Ansätze gezeigt, um durch einen Lichtleiter eine linienförmige Struktur zu generieren. Allein durch die Wahl des umgebenden Materials (Luft in (A) und (B), Polymer in (C) und (D)) und damit des Brechungsindexgradienten an einer Grenzfläche können Reflexionseigenschaften modifiziert und eine linienförmige Struktur generiert werden. Die Dimensionierung solcher Elemente basiert auf Simulationsergebnissen (Abb. 7 (E) und (F)), die im

Praxisversuch (Abb. 7 (G)) bestätigt werden. In Abb. 7 (H) und (I) wird ein weiteres Beispiel dargestellt, um zum Zweck der Neigungsmessung eine punktförmige Abstrahlcharakteristik zu generieren.

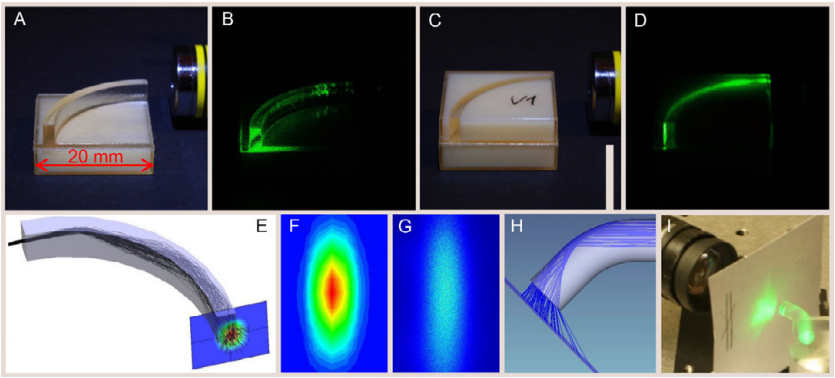


Abbildung 7: (A) und (B) Additiv gefertigter Lichtleiter zur Liniengenerierung (Drucker Keyence Agilista 3000) (C) und (D) Additiv gefertigter und mit Supportmaterial umgebener Lichtleiter zur Liniengenerierung (Drucker Keyence Agilista 3000) (E) und (F) LightTools-Simulation zum Lichtleiter aus C (G) Experimentell ermittelte Abstrahlcharakteristik des Lichtleiters aus C (H) Zemax-Simulation eines Lichtleiters zur Neigungsmessung (I) Lichtleiter aus H in der Evaluierung

6 Simulation von Freiformoptiken und 3D-Rendering

Die optische Simulation bietet nicht nur die Möglichkeit, den Beleuchtungsstrahlengang im Optikelement zu verfolgen, sondern zusätzlich auch das Detektorsystem bestehend aus Kamerasensoren und Objektiven in die Simulation zu integrieren. Die so erzeugten Daten können genutzt werden, um vor Aufbau des Systems bereits Kamerabilder auf Simulationsbasis zu rendern und die Auswertung durch eine entsprechende Bildverarbeitungssoftware zu testen.

Ein Beispiel für eine Applikation, die auf Simulationen mit gerenderten Kamerabildern beruht, wird in Abb. 8 gezeigt. In dem Beispiel wird ein additiv gefertigtes 3D-Display dargestellt (Abb. 8 (A), vgl. [1]). Das 24 mm lange Display beinhaltet ein Lichtleiterarray bestehend aus 1600

Lichtleitern mit einem Durchmesser von je 0.5 mm. Im Beispiel wird das Display durch einen Projektor (acer K132, Abb. 8 (B)) einseitig beleuchtet. Unter einem Winkel von 90° ist durch die Lichtleiter ein spiegelverkehrtes Abbild der Projektion zu erkennen.

Um ein solches Display zu dimensionieren, wurden zunächst Lichtleiter in der strahlenoptischen Simulation erprobt (LightTools, Abb. 8 (C)). Anschließend wurde durch Wahl von Kamera und Optik ein virtuelles Kamerabild gerendert (LightTools, Abb. 8 (D)). Dieses ermöglicht, schon vor der Fertigung den Kontrast durch das Display abzuschätzen.

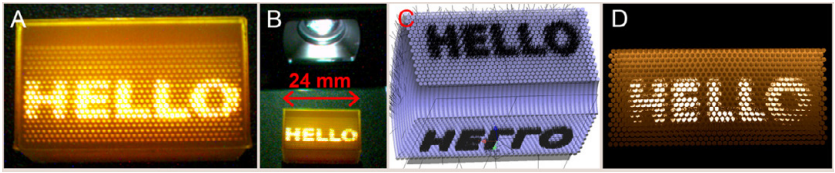


Abbildung 8: (A) Darstellung eines additiv gefertigten 3D-Displays bestehend aus 1600 Lichtleitern (Drucker Keyence Agilista 3000) (B) Display-Setup bestehend aus Projektionseinheit acer K132 und additiv gefertigtem Lichtleiterarray (C) 3D-Display in der strahlenoptischen Simulation (D) virtuelles Kamerabild durch Rendern der Simulationsdaten

7 Schlussfolgerung

Die angeführten Beispiele zeigen, dass in bisher schwer realisierbaren Applikationen durch additiv gefertigte Freiformoptiken Abtastsignale generiert werden können, um bei einem Minimum an Intensitätsschwankungen und Abschattungseffekten eine gewünschte Strukturierung zu erzielen. Durch diese kompakten, funktionalisierten Optikelemente ergeben sich neue Möglichkeiten im Bereich miniaturisierter Sensorik und automatisierter Qualitätskontrolle. Da ein erhöhter Aufwand bei der Nachbearbeitung additiv gefertigter Optikelemente und bei der Simulation von Freiformdesigns besteht, ist die Bildgewinnung durch individualisierte Optikelemente vor allem dann sinnvoll, wenn durch komplexe Oberflächenstrukturen des Prüfteils oder durch limitierte Sensorgrößen spezialisierte Detektorkonzepte notwendig werden.

Literatur

1. K. D. D. Willis, E. Brockmeyer, S. E. Hudson und I. Poupyrev, „Printed optics: 3D printing of embedded optical elements for interactive devices“, *Proceedings of the 25th annual ACM symposium on User interface software and technology*, S. 589–598, 2012.
2. E. Brockmeyer, I. Poupyrev und S. Hudson, „Papillon: Designing curved display surfaces with printed optics“, *Proceedings of the 26th annual ACM symposium on User interface software and technology*, S. 457–462, 2013.
3. R. Lachmayer, R. B. Lippert und T. Fahlbusch, *3D-Druck beleuchtet – Additive Manufacturing auf dem Weg in die Anwendung*. Hannover: Springer Vieweg, 2016.
4. R. Lachmayer, A. Wolf und G. Kloppenburg, „Rapid prototyping of reflectors for vehicle lighting using laser activated remote phosphor“, *SPIE*, Vol. 9383, Nr. 10.1117/12.2078791, 2015.
5. R. Lachmayer, G. Kloppenburg und A. Wolf, „Additive manufacturing of optical components“, *SPIE Newsroom*, Nr. 10.1117/2.1201511.006233, 2015.
6. M. Sweeney, M. Acreman, T. Vettese, R. Myatt und M. Thompson, „Application and testing of additive manufacturing for mirrors and precision structures“, *SPIE*, Vol. 9574, Nr. 10.1117/12.2189202, 2015.
7. A. Heinrich, Y. Bauckhage, S. Pekrul und M. Rank, „3d printed light pipes for advanced illumination“, *DGaO proceedings*, Vol. 117. Tagung, 2016.
8. A. Suckow, F. M. Shariff und A. Heinrich, „Additive manufacturing – design and fabrication of a fisheye lens“, *DGaO proceedings*, Vol. 117. Tagung, 2016.
9. A. Heinrich, M. Rank, P. Maillard, A. Suckow, Y. Bauckhage, P. Rößler, Y. Lang, F. Shariff und S. Pekrul, „Additive manufacturing of optical components“, *Advanced Optical Technologies*, Vol. 5, Nr. 4, 2016.
10. P. Maillard und A. Heinrich, „3D printed freeform optical sensors for metrology application“, *SPIE*, Vol. 9628, Nr. 10.1117/12.2191280, 2015.
11. A. Heinrich, P. Maillard, A. Suckow, A. Grzesiak, P. Sorg und U. Berger, „Additive manufacturing – a new approach for individualized optical shape metrology“, *SPIE*, Vol. 9525, Nr. 10.1117/12.2183168, 2015.
12. A. Heinrich, B. Sorg, A. Grzesiak und U. Berger, „An optical sensor for shape metrology based on additive manufacturing“, *DGaO proceedings*, Vol. 116. Tagung, 2015.
13. X. Jiang und H. Bunke, *Dreidimensionales Computersehen – Gewinnung und Analyse von Tiefenbildern*. Bern: Springer Verlag, 1997.

Automated surface inspection of small customer-specific optical elements

Alexander Schöch[1], Patric Perez[1], Sabine Linz-Dittrich[2],
Carlo Bach[1] and Carsten Ziolek[2]
{alexander.schoech, patric.perez, sabine.linz,
carlo.bach, carsten.ziolek}@ntb.ch

[1] NTB Interstaatliche Hochschule für Technik Buchs, Institute for Production Metrology, Materials and Optics, Machine Vision Group
Werdenbergstrasse 4, CH-9471 Buchs
[2] NTB Interstaatliche Hochschule für Technik Buchs, Institute for Production Metrology, Materials and Optics, Technical Optics Group
Werdenbergstrasse 4, CH-9471 Buchs

Abstract Surface imperfections on basic optical elements are typically assessed by manual visual inspection according to the standard DIN ISO 14997. In this article, a specifically designed machine vision setup to mimic a human tester's inspection process is proposed. It consists of multiple cameras and LED light sources. Both are arranged on the surface of a hemisphere with the optical element to be inspected at its center. By enabling individual LED sources on the hemisphere, any movement during acquisition can be omitted. Thus, the system is capable of acquiring a sparse pseudo BRDF[1] representation of imperfections in less than one second. It is shown by experiments that this representation allows to discriminate between certain imperfections.

Keywords Quality control, automated surface inspection, optical elements, optical industry.

1 Introduction

As in all manufacturing industries, producers of basic optical elements aim to deliver defect-free products. The production of, e. g., lenses, mirrors or prisms becomes more and more challenging due to continuously

[1] Bidirectional Reflectance Distribution Function

decreasing geometrical dimensions and tolerances, e. g., for application in medical devices. Moreover, customer specific requirements may result in small batch sizes. As a consequence, demands on quality control such as surface inspection are steadily increasing in terms of accuracy and flexibility. Even if surface imperfections may be solely of cosmetic interest [1–3], i. e., they do not significantly impair optical performance, customers often demand for high quality nevertheless.

Since surface inspection is nowadays typically done by human operators, more skilled personnel becomes necessary to retain quality. By the decreasing size of imperfections and varying customer-specific geometries, the time until fatigue may be reduced. Clearly, manual inspection is inevitably accompanied with variation in the results due to subjective perception [1–3]. Due to these shortcomings, several attempts for automated systems were made which are shortly reviewed in the next section.

2 Research context

A preceded survey by the authors indicates that quality control of basic optical components is typically done by means of manual visual inspection nowadays. During observation at six Swiss industrial optics manufacturers, it was ascertained that experienced testers detect and classify imperfections of lateral extensions down to $16\,\mu\mathrm{m}^2$. They achieve such accuracy by use of various magnification utensils, a dedicated comparison artifact and proper variation of viewpoint and light source position w. r. t. the surface.

The definition of imperfections and test methods are addressed in the standards DIN ISO 10110-7 [4] and DIN ISO 14997 [5] respectively. Alternatively, an American MIL standard exists. The MIL-O-13830A uses a different approach and cannot be compared directly to the ISO counterpart [1]. This article is focused on testing methods regarding the ISO standard.

The DIN ISO 10110-7 defines a variety of different surface imperfections. “Edge chips”, “Long scratches”, “Coating blemishes” and “Localized surface imperfections” which combine several regional defects and particles (orange boxes in figure 1). These imperfections can be tol-

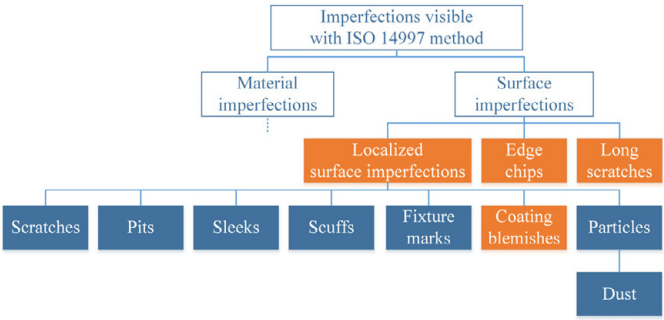


Figure 1: Classification of surface imperfections, based on DIN ISO 10110-7:2009-06 [4].

erated in technical drawings, what typically is written as follows:

$$5/N \times A; \mathrm{C}N' \times A'; \mathrm{L}N'' \times A''; \mathrm{E}A'''$$

The "5" indicates the description of a surface imperfection tolerance. The four sections separated by a semicolon describe in order "Localized surface imperfections", "Coating blemishes", "Long scratches", "Edge chips". This enumeration corresponds to the orange boxes in figure 1. The different N's are the maximal number of occurrences of the respective imperfections. And the A's represent an imperfections *grade*, defined by the maximal square root of its area in millimetre.

A dedicated mention of automated inspection systems is included in the draft version of the upcoming E DIN ISO 14997:2016-08 [6]. However, concerning automated systems, it does not assess boundaries or measurement guidelines. The main statement is a need of agreement between manufacturer and customer on a testing method.

Regarding automated systems, Turchette and Turner [3] propose a darkfield[3] microscope setup and image processing methods to specifically detect imperfections on laser resonator optics. They compare their results to manual visual inspection and cavity ringdown data and conclude a sufficient agreement. Clearly their method was applied on planar optics only, i. e., laser mirrors.

[3] Darkfield is a light setting where only scattered light is captured. Specular reflections are avoided by principle.

Liu et al. [7] propose a similar darkfield microscope setup and employ stitching techniques to achieve an expanded measurement area. They assess their results on photolithographic scratches and report errors about 3.5 %. Experiments are again limited to planar optics.

To the knowledge of the authors, two commercial products to evaluate surface imperfections on optical elements are currently available.

The SavvyInspector [8, 9] from Savvy Optics Corp. employs a darkfield setup and a matrix sensor, allowing for a $1\,\mathrm{mm} \times 1\,\mathrm{mm}$ inspection area. Besides its applicability for flat parts, also "mild concave" surfaces are advertised for. The system was originally designed to assess conformance to the American standard MIL-O-13830A but provides also an ISO 10110-7 mode. For this, the minimal specified grade is $5/1 \times 0.025$ for a single imperfection and imperfection classes are limited to general defects and coating imperfections.

Recently, the Dioptic ARGOS system was presented [2]. It employs a line sensor and a rotation stage. A holistic representation of an optical element's surface is constructed by rotating the element during the acquisition phase and transforming the line sensor data accordingly. Their darkfield setup enables assessment of curved elements such as lenses by preliminary manual adjustment. It prevents the occurrence of specular reflections from curved surfaces and provides high resolution images up to 256 Megapixel. As minimal quantifiable grade, $5/1 \times 0.0025$ for single imperfections is stated and a repeatability between $0.4\,\mu\mathrm{m}$ and $1.3\,\mu\mathrm{m}$ based on 30 repetitions is reported. Etzold et al. [2] indicate that the system discriminates digs, scratches and edge chips.

The need for quality inspection of transparent parts is obviously not limited to classical optical components but also found in, e. g., inspection of car headlamp lenses. Martinez et al. [10, 11] propose a dedicated lighting setup which is able to "move" a light source. By this method, a darkfield configuration is generated which yields sufficient image contrast during measurement.

It is worth noting that all reviewed approaches employ a darkfield setup and are mostly limited to planar optics. Neubecker & Hon [1] show that the signal contrast and the signal-to-background-noise ratio are potentially better than in brightfield conditions. However, they state that the direction of the maximal radiance, scattered by an imperfection, is almost unpredictable in a darkfield setup. This in turn could be accounted for by introducing movement of light sources and/or cameras

or alternatively by multiple light sources and/or cameras. Neubecker & Hon [1] argue that mechanical actuators are slow and require more maintenance which is why a system capable to capture a sample with one shot is preferable.

3 Research goals

This work is a first step to advance current methods w. r. t.

1. Evaluation of camera positions and light scenario heuristics to enable holistic inspection on curved surfaces without mechanical movement.

2. Classification of different imperfection and particle types. In a first step, specification possibilities from the DIN ISO 10110-7 shall be covered (orange boxes in figure 1). As further target, all of the imperfections indicated in figure 1 shall be discriminated.

4 Acquisition system

As a starting point, first experiments were based on the manual inspection process which is in use at all surveyed optics manufacturers. This manual method of searching for imperfections is not trivial to reproduce in an automated system. A lot of movements, viewing angles and lighting conditions are based on inspectors experience. As expert knowledge grows, the operators can discriminate between imperfection categories and even link a defect to a manufacturing step.

However, the main light setting is a white light darkfield setting. First instincts lead to a diffuse light dome. This setup has already the potential to make a lot of typical defects visible and shows the main crux of the task: while it is feasible to find defects on a surface already in the field of view and in focus with appropriate light settings, it is, however, hard to detect imperfections on the whole curved surface of an optical component [1,2].

4.1 Universal capture setup

In order to test for a multitude of camera poses and keep the experiments economically feasible, a surrogate dome was designed that allows to emulate different camera positions. As shown in figure 2, the dome can be rotated around two orthogonal axes – this is equivalent to rotations of the camera around the domes center and was preferred due to practical reasons. Clearly, camera positions are only feasible where the line of sight is not obstructed by the hemispheres geometry, i. e., at the north pole and between strands.

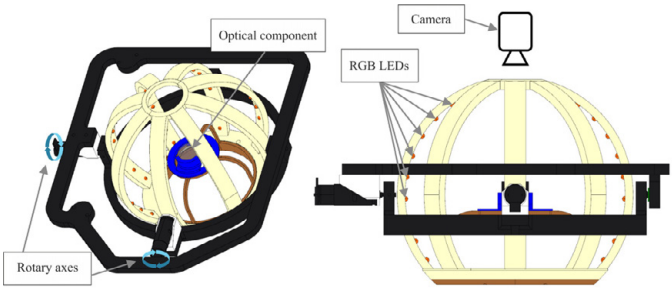


Figure 2: CAD drawing of the surrogate setup. RGB LED positions are indicated by orange spheres. Left: Isometric view. Right: Front view.

The dome consists of eight strands, with ten uniformly spaced LEDs attached to each of them (figure 2). Each RGB LED can be controlled separately in intensity and color. This allows for discrete changes of the illumination angle, creating a brightfield or darkfield setting, as well as lighting scenarios with multiple sources enabled. To address the at times highly curved surface of convex or concave lenses, we needed to decide between either a multiple camera configuration or a rotation stage for our experimental setup. A rotation stage allows to look at each position on the sample surface with different angles and hence allows us to control the angle between illumination and camera-axis continuously. Regarding this correction possibility for the present surface normal, we decided for the rotation stage.

A major part of the dome was realized by means of additive manufacturing (figure 3). Further details are given in table 1.

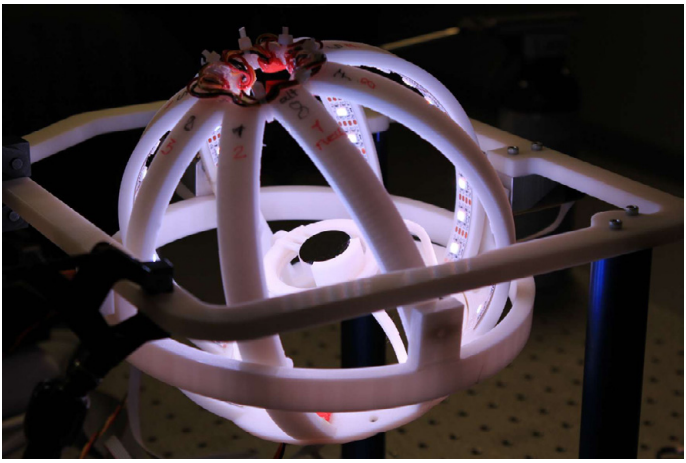


Figure 3: Photograph of the surrogate setup, slightly tilted with all LEDs enabled and planar optical component inserted.

4.2 First experiments

The described setup allows us to see all typical imperfections. To test the performance on small defects we made a comparison between some imperfection classes in our dome and a microscopy capture. With the current equipment we are able to detect defects down to a minimal grade of $5/1 \times 0.016$ for a single defect, as depicted in figure 4.

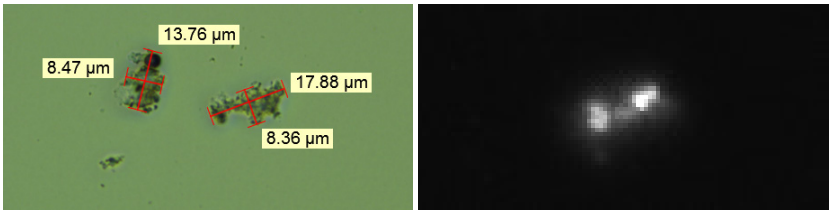


Figure 4: An example of a dig of size $5/1 \times 0.016$ on a planar mirror. On the left under a microscope and on the right with our dome.

A second typical and straightforward case is the scratch. It appears often as a set of digs or holes on a line. This example is shown in figure 5.

Table 1: Setup specifications.

Camera	**AV MAKO U-503B**		
Pixelsize	2.2 µm	Resolution	2592 px × 1944 px
Sensor type	8 bit Mono CMOS	Sensor size	5.7 mm × 4.3 mm
Connection	USB 3.0	Frame rate	14 fps

Lens	**SILL TZM 4425/1,0-C** Telecentric		
Magnification	×1.0	Max. Distortion	0.1 %
Working distance	107 mm	Connection	C-Mount

Lighting	**Flex RGB LED Strip**		
LED Chip	WS2822S	Chip format	5 mm × 5 mm
Control	DMX512		

It is not significantly harder to find but a classification has to correctly link these single defects to a larger structure as they are evaluated as a whole scratch and not as single digs.

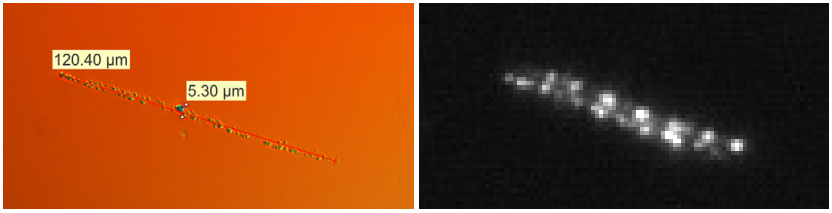


Figure 5: This example shows a scratch with a width of about 5 µm. On the left under a microscope and on the right with our dome.

One of the hard to image imperfection types is the scuff. Scuffs are typically very small trenches with a continuous profile. They can be very small in width and vary extremely in length. One example is shown in figure 6. Its width is about a micrometre but the length is more than two millimetres, it stretches over the whole convex lens. The images show a typical behaviour of those damages. They are only visible if illuminated from a very narrow sector of our dome. This explains the missing intersecting scuff from the microscopy image on the dome capture.

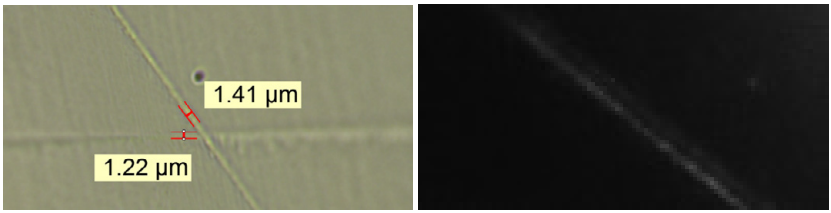


Figure 6: This is an example of our current limit. On the left a microscopy image where two scuffs can easily be seen. The horizontal scuff is only partly in focus because of the high curvature of the sample lens. The right image shows one of the scuffs under the dome with a single LED activated.

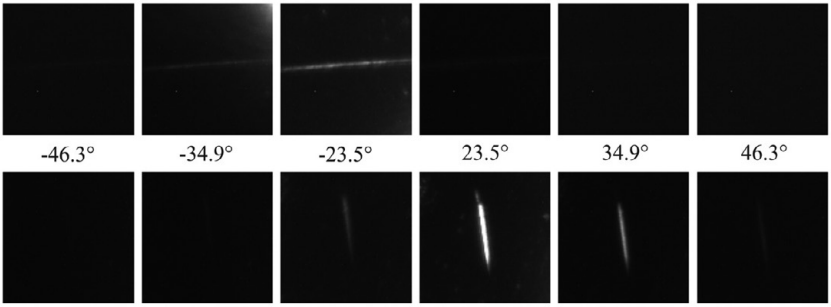


Figure 7: The two picture rows show different scuffs with a fixed camera position and changing light angle. Light is shining in a perpendicular azimuth angle to the scuffs axis but with changing polar angle as indicated.

From several examples and expert knowledge of the inspectors we know about some differences in appearance of different imperfection types. Some are quite hard to formulate quantitatively. But one particular rule is the visibility of scuffs. Experts describe the visibility of these as only a flashing if the illumination angle is perfect. This can easily be replicated. In the two image-collections in figure 7 we visualise this behaviour. After testing multiple samples, we can express this by some rules. Scuffs are best visible if illuminated in a nearly brightfield condition, where the light source is preferably only tilted around the scuffs direction-axis.

5 Conclusions

The current surrogate setup allows to mimic the manual inspection process according to the DIN ISO 14996 for surface imperfections in an automated way.

To discriminate between scuffs and other types of imperfections we capture a set of images with different LED settings and formulate a sparse pseudo BRDF. We call this pseudo BRDF since the captured light intensity is evaluated and not the ratio between irradiance and reflected radiance. This, however, is sufficient because the aim is a relative comparison among defects.

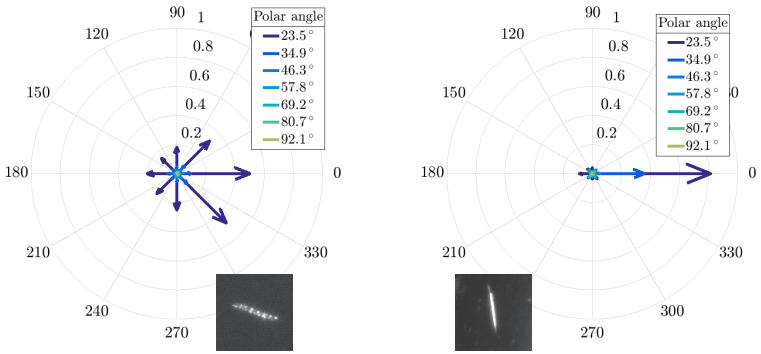


Figure 8: Comparison between a sparse pseudo BRDF of the scratch from figure 5 (left) and the lower scuff from figure 7 (right).

As a prominent example, figure 8 shows a scratch and a scuff with their correspondent pseudo BRDF evaluations. They seem similar in visible area on the samples but have a significantly different reflection behaviour. Besides the given example, several experiments indicate that a pseudo BRDF representation of imperfections allows for discrimination between scuffs and scratches.

6 Outlook

The gained understanding of most defect types under different illumination and camera angles leads to possible designs for a motionless setup. Essentially, mechanical motion shall be replaced by employing multiple cameras at appropriate positions. To test a broader variety of customer-specific geometries, a flexible fixture for optical elements is to be manufactured.

One of the main needs for classification is to find a possibility to discriminate between particles such as dust and surface imperfections. To achieve this, classification approaches such as from Li et al. [12] are to be tested with the system. Further promising possibilities to discriminate imperfections include the evaluation of the defect surface's topography by means of multi view approaches [13].

References

1. R. Neubecker and J. E. Hon, "Automatic inspection for surface imperfections: requirements, potentials and limits," in *Third European Seminar on Precision Optics Manufacturing*, ser. SPIE Proceedings, R. Rascher, O. Fähnle, C. Wünsche, and C. Schopf, Eds. SPIE, 2016, p. 1000907.
2. F. Etzold, D. Kiefhaber, A. F. Warken, P. Würtz, J. Hon, and J.-M. Asfour, "A novel approach towards standardizing surface quality inspection," in *Third European Seminar on Precision Optics Manufacturing*, ser. SPIE Proceedings, R. Rascher, O. Fähnle, C. Wünsche, and C. Schopf, Eds. SPIE, 2016, p. 1000908.
3. Q. Turchette and T. Turner, "Developing a more useful surface quality metric for laser optics," in *SPIE LASE*, ser. SPIE Proceedings, W. A. Clarkson, N. Hodgson, and R. Shori, Eds. SPIE, 2011, p. 791213.
4. DIN ISO 10110-7, "Optics and photonics – Preparation of drawings for optical elements and systems – Part 7: Surface imperfection tolerances (ISO 10110-7:2008)," 2009.
5. DIN ISO 14997, "Optics and photonics – Test methods for surface imperfections of optical elements (ISO 14997:2011)," 2013.
6. E DIN ISO 14997, "Optics and photonics – Test methods for surface imperfections of optical elements (ISO/DIS 14997:2016)," 2016.
7. D. Liu, Y. Yang, L. Wang, Y. Zhuo, C. Lu, L. Yang, and R. Li, "Microscopic scattering imaging measurement and digital evaluation system of defects

for fine optical surface," *Optics Communications*, vol. 278, no. 2, pp. 240–246, 2007.

8. D. M. Aikens, "The Truth About Scratch And Dig," in *Optical fabrication and testing*, ser. OSA technical digest (CD). OSA, The Optical Society, 2010.
9. ——, "Objective Measurement of Scratch and Dig," in *Applied industrial optics*, ser. OSA technical digest (online). OSA, The Optical Society, 2012.
10. S. S. Martínez, G. J. Ortega, G. J. García, and S. A. García, "A machine vision system for defect characterization on transparent parts with non-plane surfaces," *Machine Vision and Applications*, vol. 23, no. 1, pp. 1–13, 2012.
11. S. S. Martínez, J. G. Ortega, J. G. García, A. S. García, and E. E. Estévez, "An industrial vision system for surface quality inspection of transparent parts," *The International Journal of Advanced Manufacturing Technology*, vol. 68, no. 5-8, pp. 1123–1136, 2013.
12. L. Li, D. Liu, P. Cao, S. Xie, Y. Li, Y. Chen, and Y. Yang, "Automated discrimination between digs and dust particles on optical surfaces with dark-field scattering microscopy," *Applied Optics*, vol. 53, no. 23, pp. 5131–5140, 2014.
13. R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, 2nd ed. Cambridge, UK and New York: Cambridge University Press, 2003.

Combining synthetic image acquisition and machine learning: accelerated design and deployment of sorting systems

Max-Gerd Retzlaff[1,3], Matthias Richter[2,3],
Thomas Längle[3], Jürgen Beyerer[2,3] and Carsten Dachsbacher[1]

[1] Karlsruhe Institute of Technology (KIT), IVD, Computer Graphics Group
[2] Karlsruhe Institute of Technology (KIT), IAR, Vision and Fusion Laboratory
[3] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB)

Abstract Machine learning methods can automate the design of large parts of an image processing pipeline in automated optical inspection (AOI) systems. However, these methods typically require an annotated sample of the objects under inspection, and creating such samples is still a manual and labor-intensive process. Synthetic image acquisition (SIA) can fill the gap to automate this step. SIA joins a physically-based image synthesis pipeline and procedural modeling techniques to recreate a physical image acquisition process. We show that, when the hardware parameters of a system are known, SIA can be used to train a classifier, which can then be used for the physical system. Time-consuming manual acquisition and labeling of a training sample is no longer necessary. Evaluations in the domain of glass recycling demonstrate that the SIA approach performs on par with a classifier that was trained using a manually collected training set.

Keywords Optical inspection, computer graphics, image synthesis, procedural modeling, pattern recognition, classification, glass sorting.

1 Introduction

Sorting systems, as any automated optical inspection (AOI) system, must solve two major tasks: image acquisition and image processing.

The image acquisition has to be designed in a way that ensures that the subsequent image processing and classification deliver consistent and accurate results. System parameters include, but are not limited to, illumination, camera including lenses, filters, and image sensor, and their geometric arrangement. The image processing requires definition of discriminative feature descriptors and classification rules that derive a sorting decision from the features. Because the number of free parameters for both hardware and software is very large, the design of a AOI is a time consuming and costly process.

When the feature descriptors rely on color information, a suitable selection of a illumination and camera might be surprisingly complicated, especially with LEDs or fluorescent lamps that exhibit non-uniform emission spectra (cf. Fig. 5(a)). Both camera sensor and light source have a significant impact on the color signal. The result of a poor choice can be that some materials with different absorption spectra might not be distinguishable in the resulting RGB images (cf. Fig. 1).

This complication can also appear when the camera or the light source in existing sorting setup is replaced, or even when the system is put into a different environment. If the color signal changes too much, the modification may also invalidate the previously defined classification rules, requiring further recalibration work.

1.1 Motivation

As outlined in the abstract, the combination of synthetic image acquisition (SIA), detailed in section 2.2, and machine learning allows automating large parts of the development and deployment of an automated optical inspection (AOI) system.

Not only does SIA replace the time-consuming manual acquisition and labeling of a training sample, the synthetic training sample may also feature statistical variations that are difficult to include otherwise, like varying environmental lighting conditions during different times of day or seasons of the year. A classifier trained on such data can compensate for this variation, generalize over these nuisance factors, and therefore show more consistent performance.

Of course, this advantage is diminished by the fact that instead a realistic model of the objects has to be developed. However, this drawback is more than compensated by two important properties of SIA.

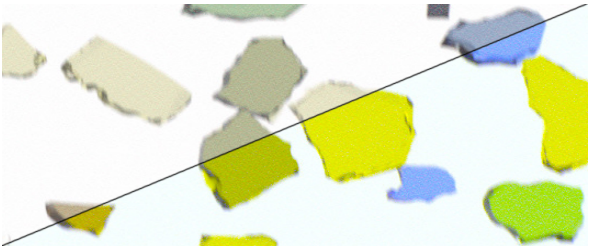


Figure 1: Images generated with spectral rendering in strong transmitted light, simulating the raw sensor response of an *e2v ELiiXA UC4/UC8* camera (top) or the human color sensitivity (CIE 1931 standard colorimetric observer) (bottom).

First, huge synthetic data sets that are infeasible to collect and annotate manually can be generated automatically, which in turn usually leads to better classifiers [1]. Second, the synthetic sample can easily be regenerated, for example, when the system's hardware components or the environmental conditions are altered. It is not necessary to manually collect a new dataset every time the setup is changed.

Furthermore, performance characteristics of the classifier can give valuable feedback on the design of the sorting system itself. The classification performance of many different hardware configurations can be compared automatically to quickly narrow down on promising parameters that can be tested in a physical prototype.

1.2 Related work

Given the benefits, it comes to no surprise that this area has seen a lot of research in the past. In 1995, Tarabanis et al. [2] surveyed a wealth of methods for automatic sensor planning in computer vision systems. They focus on methods for which the goal is to determine good configurations (position, orientation, etc.) for sensors and illumination in object feature detection, object recognition and localization, and scene reconstruction tasks. Here, however, we are not interested in optimization of parameters, but rather whether the virtual evaluation of a given set of parameters is a valid optimization goal in the first place.

Nilsson et al. [3] pursue a similar goal. They consider a system to detect welding joints between two metal sheets. They use a commercial

robotics simulation software to simulate the camera image of an inspection robot, and compare the simulation to a real system. In their use case, they find high consistency between the camera image and the simulation. Reiner [4] comes to the same conclusion in a similar work.

Closely related to our work, Irgenfried et al. [5] estimate parameters for object segmentation algorithms where the ground truth was generated by rendering 3D models. In evaluation using real-world images they find that the estimated parameters work well, even when the 3D models do not have a lot of details. But nevertheless, they stress the importance of physically-based rendering for robust, transferable results. This aspect is even more important for our setting, in which classification rules learned from synthetic images are applied to the real world.

1.3 Contributions

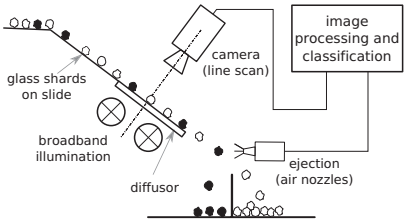
We demonstrate our approach by the example of glass recycling. We give an overview over the synthetic image acquisition method and classification algorithms. Images of glass shards of varying color are generated and a classifier is trained using these images. We evaluate the performance of the classifier with synthetic images as well as with images recorded using a matching physical system, and discuss the results.

2 Methods

In this section, we briefly describe the acquisition of the real-world and synthetic image datasets, as well as the classification pipeline.

2.1 Image acquisition system

A schematic of the physical sorting system used in this work is shown in Figure 2(a). Glass shards of different color enter the system on a slide, on which they accelerate until they reach the inspection stage. Here, an RGB line scan camera records the shards as they pass over a broadband illumination behind a diffusor. The line image is subjected to a shading correction, that modifies the color value of each pixel such that a reference image would appear as uniform white. An occlusion-free image of the light source is used as reference image. A full RGB image is constructed by stacking consecutive line images, estimating the speed of



(a) Schematic of an AOI system for glass sorting.

(b) Lab setup of the *HR Fine* image acquisition system at Fraunhofer IOSB.

Figure 2: Schematic and lab setup of an automatic glass sorting system.

the glass shards so that the image is not be distorted along the direction of travel of the shards.

2.2 Synthetic image acquisition

In [6], we described an image generation pipeline for the evaluation and optimization of measuring and AOI systems. Physically-based image synthesis methods, a research direction in computer graphics, are capable of simulating optical measuring systems in their entirety. In addition, so called procedural modeling techniques can be used to quickly generate large sets of virtual samples and scenes thereof that comprise the same variety as physical testing objects and real scenes, e. g., if digitized sample data is not available or difficult to acquire.

We use these procedural modeling techniques to generate large sets of virtual glass shards and appropriate image synthesis techniques to model the realistic image formation of a virtual recreation of the physical image acquisition system, considering light sources, materials, complex lens systems (optionally), and sensor properties; refer to Figure 3 for an example. This approach also allows us to generate a wealth of annotation information for each object in the image.

The resulting synthetic images can be used to evaluate and improve complex measuring systems and automated optical inspection (AOI) systems independent of a physical realization. In this paper, we use the synthetic images to improve the classification of a glass sorting system.

Figure 3: A synthetic image of procedurally generated glass shards produced by a Monte Carlo rendering framework simulating a real lens system.

2.3 Image processing

Using histogram back-projection, a mask of back- and foreground is created, and connected component analysis (CCA) is used to locate individual shards. For the synthetic images, back-projection and CCA are not necessary, as exact segmentation information is generated alongside the virtual camera image. In either case, the mask images are morphologically eroded to discard the fringes of the shards in classification.

Pattern recognition As the hue of a shard is the primary classification criterion, the images are converted into the HSL colorspace before classification. The saturation and luminosity channel also carry discriminative information, e. g., a brown shard appears darker than a clear one.

We evaluated two classification pipelines. The first pipeline, which acts as the baseline, describes each shard by the first color moment of the foreground pixels, i. e., $\mathbf{m} = \frac{1}{|\mathcal{F}|} \sum_{\mathbf{x} \in \mathcal{F}} \mathbf{s}(\mathbf{x})$, where $\mathbf{m} \in \mathbb{R}^3$ denotes the color moment, $\mathcal{F}$ denotes the set of foreground pixels $\mathbf{x}$, and $\mathbf{s}(\mathbf{x})$ denotes the color of the pixel at $\mathbf{x}$. The color moment is then classified using a support vector machine (SVM) classifier with a Gaussian kernel.

The second pipeline uses the bag of visual words descriptors of the color. Briefly, the HSL space is quantized using using K-means clustering of the foreground pixels. Each of the resulting K Voronoi regions is identified with its center $\boldsymbol{\mu}_k$. An unseen shard image is described by the count static of cluster memberships over the foreground pixels, that is,

$$\mathbf{w} = (w_1, \ldots, w_K) \quad \text{where} \quad w_k = \frac{1}{|\mathcal{F}|} \sum_{\mathbf{x} \in \mathcal{F}} \delta^k_{\arg\min_i \|\mathbf{s}(\mathbf{x}) - \boldsymbol{\mu}_i\|}.$$

These descriptors are then classified using a linear SVM classifier. This approach is explained and explored in more detail in [7].

3 Experiments

3.1 Sample of glass shards

We collected a sample of 2,516 glass shards from a batch of glass waste from a producer of glass sorting machines and project partner of Fraunhofer IOSB. We manually sorted these glass shards into three classes of 368 *brown*, 1,368 *green*, and 780 *clear* individual shards.

3.2 Acquisition of datasets

Recorded images To obtain the dataset we captured images with a lab setup of a *HR Fine* image acquisition system by Fraunhofer IOSB, shown in Figure 2(b). After image preprocessing and object detection as described in section 2.3, we discarded shards with less than 20 foreground pixels after erosion. This left the dataset with 367 *brown*, 1,366 *green*, and 769 *clear* samples.

Synthetic images We measured the absorption spectra of several glass shards representing the different glass-types in the VIS spectrum (380 nm to 780 nm). For each glass type at least three individual shards were measured, resulting in a total of 23 measurements. Figure 4 shows representative spectra of six glass subtypes of the three classes.

The absorption spectra, and therefore colors, of the main glass classes *brown*, *green*, and *clear* can differ considerably. Bottle glass, usually consisting largely of recycled waste glass, is available in many shades from brown to green. The line between the three classes is not at all well defined. Furthermore, the green and blueish green shard absorb red light (620 nm to 720 nm), but the yellowish green shard does not. Its absorption spectrum shows more resemblance to the brown than the green shards. Overall, such shards may be difficult to classify correctly.

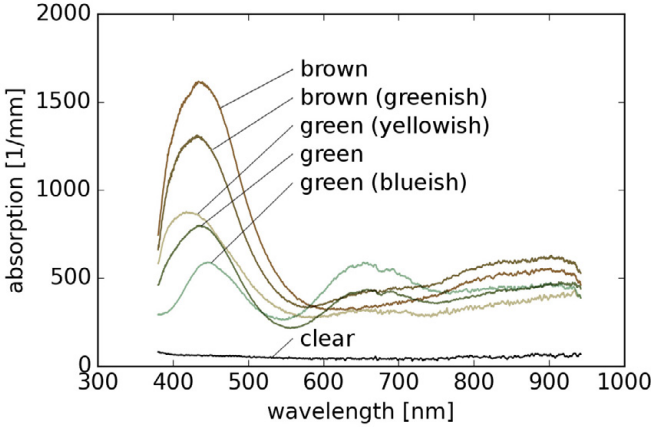
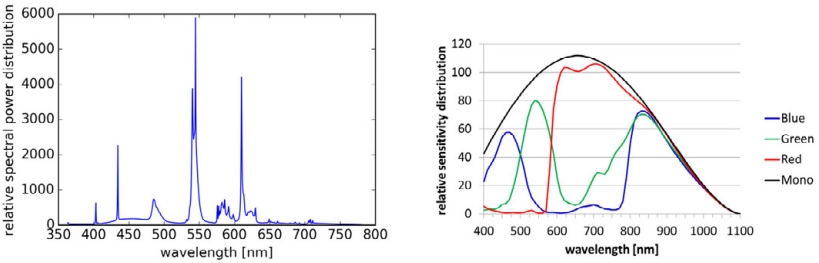


Figure 4: A collection of spectra of measured glass types (colors taken from pictures of the measured glass shards themselves, except for clear glass).

In addition to the glass spectra, we measured the emission spectrum of the fluorescent light source of the *HR Fine* system (Fig. 5(a)) and acquired the raw spectral response (Fig. 5(b)) of the image sensor of the *e2v ELiiXA UC4/UC8* camera used in the system. These measurements allowed to virtually recreate the measuring situation of the *HR Fine* system. Our existing procedural glass shards models, described in [8], were used to generate virtual glass shards and ultimately synthetic images that simulate image acquisition of the physical system. In total, we generated $5{,}118$ individual glass shards images with full annotation: $1{,}598$ *brown*, $2{,}516$ *green*, and $1{,}004$ *clear* shards.

3.3 Metric

Since the datasets are considerably imbalanced, we use Matthews Correlation Coefficient (MCC, [9]) to evaluate performance, as MCC is symmetric and therefore less sensitive to class imbalance than other metrics. MCC can be interpreted as the correlation between classifier prediction and ground truth. A value near zero indicates that classification is not better than random guessing, while a value of one means perfect classification.



(a) Emission spectrum of the fluorescent light source.

(b) Raw sensor response of the *ELiiXA UC4/UC8* line scan camera by *e2v*.

Figure 5: Spectra of light source and sensor response of the *HR Fine* system.

MCC is a binary metric, but we investigate a three-class setting. We therefore report the MCC for each class individually, and we consider the target class as positive class and all other classes as negative class.

3.4 Implementation details

In all experiments, we used a vocabulary of $K = 10$ words for the bag of words descriptors. Images used to learn the vocabulary were not used to train the classifier. The hyperparameters of the SVM classifiers were chosen in a randomized search. In the first set of experiments, stratified fivefold cross validation was used to estimate classification performance. In the cross-source experiments, we used bootstrapping to estimate confidence intervals of classification performance.

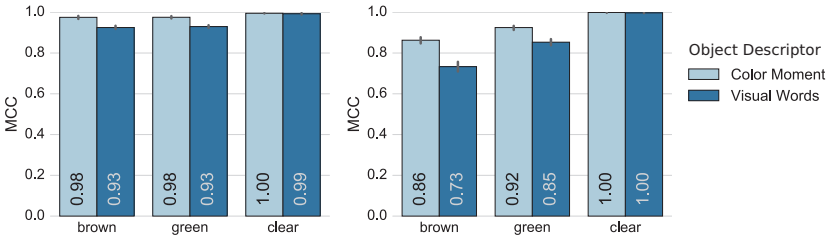
3.5 Intra-source experiments

In the first sets of experiments, the classification pipelines were trained and evaluated with data from the same source, either using only synthetic images, or using only recorded images. These experiments will provide insights which pipeline is more suited for the application and what classification performance can be expected for the particular data.

Figure 6 shows the MCC for each class and classification pipeline. For the synthetic dataset, classification is a bit more reliable with color moments, but very accurate with either descriptor. With recorded images,

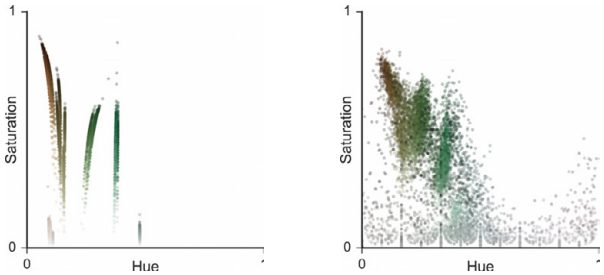
brown and green shards are more difficult to classify, which confirms our expectation from section 3.2. Furthermore, the baseline color moment descriptor significantly outperforms the visual word approach.

Figure 7 indicates a reason: The color in the synthetic images is distributed in discrete bands, while the recorded images show large variation in the hues. The variation is most likely due to image noise as well as a mismatch between the shutter speed of the camera and the speed of the shards on the slide. Furthermore, each virtual shard exhibits the spectral characteristics of one of the 23 measured shards, while real shards may show more varied spectra due to varying glass mixtures. The visual vocabulary is unable to adequately represent this variation.



(a) Synthetic images (b) Recorded images

Figure 6: Classification performance in the intra-source experiments.



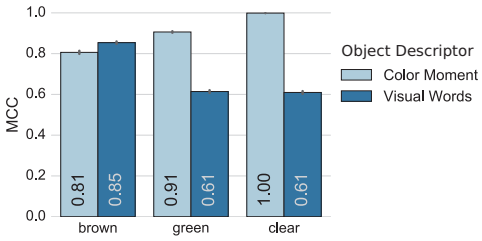
(a) Synthetic images (b) Recorded images

Figure 7: Color distribution of foreground pixels in the datasets in the hue-saturation space.

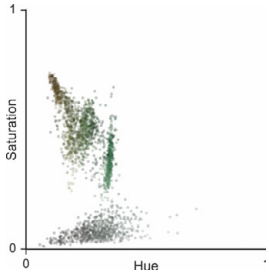
3.6 Cross-source experiments

In the transfer experiment, the classifier was trained using synthetic images, but evaluated with captured images. Figure 8(a) summarizes the results. The color moment descriptor enables significantly more reliable classification of green and clear glass shards than the visual word descriptor. A reason can again be found in Figure 7: The visual vocabulary is adapted to the discrete color bands in the synthetic images, since it was learned from that data. The variation in the captured images cannot be represented by the vocabulary, which leads to noisy image descriptors. The color moment, on the other hand, removes much of the noise (cf. Fig. 8(b)), and allows the classifier to find decision regions that generalize well.

Despite the noise in the recorded images, classification performance in the cross-source setting comes very close to classification performance when the recorded images are used for both training and evaluation (cf. Figs. 6(b) and 8(a)). Going into more detail, the confusion matrices in Figure 9 reveal that in cross-source classification, brown shards are more often classified as green shards, while green shards are less likely to be confused with brown shards. In other words: the classifier shows a small bias towards the green class, but otherwise is well suited to replace the classifier trained on real images.

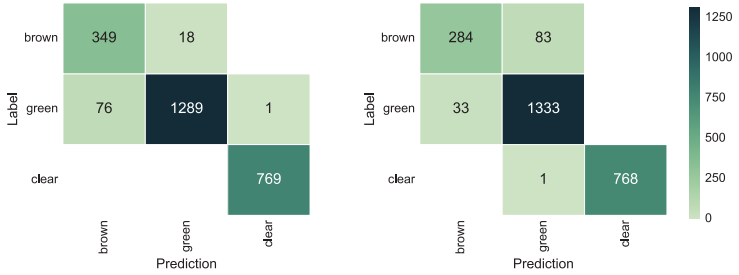


(a) Classification performance for each class with respect to the object descriptor.



(b) Distribution of object descriptors in the hue-saturation color space.

Figure 8: Results of the cross-source experiments.



(a) Training and testing with recorded images.

(b) Training with synthetic, testing with recorded images.

Figure 9: Confusion matrices for two experiments (color moment descriptors).

4 Conclusion

Altogether, we investigated the combination of synthetic image acquisition (SIA) and machine learning to accelerate design and deployment of automated optical inspection (AOI) systems. In the context of glass sorting, we have demonstrated that a classifier can be trained using synthetic images. The SIA system was parametrized to recreate the measuring conditions and hardware parameters of a real-world system. We have shown that on physically acquired images, this classifier performs on par with a classifier that was also trained using these images, even though the synthetic training images currently show less variation than physically acquired images.

The difference in variation also results in a large performance gap between classification exclusively on synthetic versus exclusively on recorded data. In order to avoid overfitting when designing an AOI system, classification performance on synthetic data should not be the only optimization goal. This issue could be addressed by statistically varying the simulated absorption spectra, which we plan for future work, and also by simulating sensor noise, chromatic distortions, and other disturbing factors.

Still, the deployment of new or modified AOI systems can be accelerated considerably by replacing the training data collection step by SIA. A classifier could even be trained remotely, while the sorting system is being integrated or modified.

References

1. A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.
2. K. A. Tarabanis, P. K. Allen, and R. Y. Tsai, "A Survey of Sensor Planning in Computer Vision," *IEEE transactions on Robotics and Automation*, vol. 11, no. 1, pp. 86–104, 1995.
3. J. Nilsson, M. Ericsson, and F. Danielsson, "Virtual machine vision in computer aided robotics," in *2009 IEEE Conference on Emerging Technologies & Factory Automation*. IEEE, 2009, pp. 1–8.
4. J. Reiner, "Rendering for machine vision prototyping," in *Optical Systems Design*. International Society for Optics and Photonics, 2008, pp. 710 009–710 009.
5. S. Irgenfried, F. Dittrich, and H. Wörn, "Realization and evaluation of image processing tasks based on synthetic sensor data: 2 use cases," in *Forum Bildverarbeitung 2014*, vol. 82. KIT Scientific Publishing, 2014, p. 35.
6. M.-G. Retzlaff, J. Hanika, J. Beyerer, and C. Dachsbacher, "Potential and challenges of using computer graphics for the simulation of optical measurement systems," *18. GMA/ITG Fachtagung: Sensoren und Messysteme 2016*, pp. 322–329, May 2016.
7. M. Richter, T. Längle, and J. Beyerer, "Visual words for automated visual inspection of bulk materials," *14th IAPR International Conference on Machine Vision Applications (MVA)*, pp. 210–213, Mar. 2015.
8. M.-G. Retzlaff, J. Stabenow, J. Beyerer, and C. Dachsbacher, "Synthesizing images using parameterized models for automated optical inspection (AOI)," *tm – Technisches Messen*, vol. 82, no. 5, pp. 251–261, Apr. 2016.
9. B. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

Realistic simulation of camera images of micro-scale defects for automated defect inspection

Haiyue Yang[1,2], Tobias Haist[1], Marc Gronle[1] and Wolfgang Osten[1]

[1] Universität Stuttgart, Institut für Technische Optik,
Pfaffenwaldring 9, 70569 Stuttgart
[2] Universität Stuttgart, GSaME,
Allmandring 35, 70569 Stuttgart

Abstract Imaging-based automatic inspection technology has developed rapidly in the past decades and is widely applied in industry. Nowadays, the image processing method in this technology, applied posterior to the measurement, can provide high recognition rate of the defect detection. But the success of image processing with respect to defect detection still requires a huge set of training data. In real production processes it is necessary to obtain and generate enough representative samples of every single part. This process results in as well as a wasting of parts and time. Here, a virtual surface defect simulation is applied to substitute the image acquisitions in imaging-based automatic inspection systems. In this paper, we focus on simulating the micro-scale defects on metal surfaces. We compare the results between ray tracing and scalar diffraction approximation methods.

Keywords Micro-scale defect, ray tracing, TEA, LPIA.

1 Introduction

Industrial automation has developed rapidly in the past decades. Customized fabrications and short production time require flexible and high speed inspection systems. Based on these requirements, imaging-based automatic inspection technology for detecting surface defects becomes more and more popular. Since the classification of defects depends on the images captured by the camera, in order to obtain high

accurate classification rates for detecting defects, it is necessary to store sufficient realistic training data (images) in the detection system beforehand [1]. However, in a modern and flexible production environment, not enough of those training images are available at the moment of the design of the inspection systems, especially for some rare or expensive samples. To circumvent this problem, modelling and rendering the training sample images for the multi-sensor surface inspection systems [2] instead of realisitic images becomes an alternative efficient choice.

The research on the defect simulation by means of ray tracing methods in the field of the computer graphics has been started more than twenty years ago [3–5]. Normally, the 3D objects in the rendering scene are modelled by many small patches, e. g., triangles. Since the diameter of these triangles is normally larger than 10 mm, in order to render triangles based on parameters, such as the roughness and bright spots, different bidirectional reflectance distribution function (BRDF) models are used [6]. The BRDF can vary the reflection probability according to the incident angle. After the optimization of BRDF, it is even possible to obtain rendered images, that look very similar to real acquired images, by combining BRDF models with the properties of lenses, light sources and cameras.

However, when the detection scale of the objects becomes smaller, especially in the micro-scale range, this BRDF-based approach is not suitable any more, since the BRDF only describes statistical effects of entire areas of the object's surface and does not consider the contribution of the individual microfacets. Furthermore, diffraction effects can also not be rendered. Hence, it is important to find the limiting scale of the ray tracing rendering method for micro-scale structures.

In this paper, we show several methods to simulate realistic images of microstructures, whose surface topography have been measured using a confocal microscope. We compare a GPU-based ray tracing method and wave optical methods for 2D imaging of coarse metal surfaces. In the ray tracing method, each microfacet is treated as a perfect mirror, and the reflection intensity is calculated according to the Fresnel equation. In the wave optical methods, several scalar diffraction approximation methods are tested, such as thin element method (TEA) and local plane interface approximation (LPIA).

2 Experimental setup

In order to evaluate the results of simulated images, an inspection system for the micro scale has been designed and built. Since the image of the sample varies by different incident angles, the rotated stages for the sample and light source are mounted in the system, see Fig. 1. During the simulation process, the profiles of the samples are obtained by the measured data, and this measurement is implemented by a confocal microscopy. The whole rendering process is done by the GPU based ray tracing software Macrosim [7].

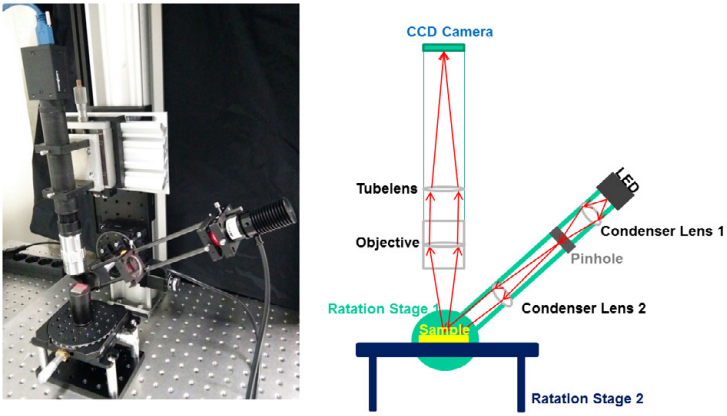


Figure 1: Left: inspection setup for the microstructure. Right: Schematic of system.

A LED light source with a wavelength of 625 nm (LCS-0625-38-11, Mightex) is collimated by two condenser lenses. The illumination system is mounted on a motorized rotation stage (8MR191-30-28, Standa) to achieve different incident angles from 20° to 65°. The real images of the samples are captured by a 5.0 MP camera (GS3-U3-50S5M-C, Point Grey) together with a magnification objective (10x Mitutoyo Plan Apo Infinity-Corrected Long WD Objective, NA $= 0.28$). By inserting a 160 mm tube lens, the system can acquire images with a magnification of 8.22 (1.36 µm lateral resolution). The profiles of samples were obtained by a spinning disk confocal microscope with a 1392×1024 pixel

CCD camera (PCO Pixelfly QE, 10 bit dynamic range), and a 20x objective (Olympus LMPlanFI, NA = 0.4).

3 Rendering methods

3.1 Ray tracing method

In order to obtain accurate rendering images for the inspection system, it is necessary to take all the parameters, which influence the image quality, into account. These parameters in this microscopic inspection system are the light source, the detected sample and the camera system. Since the collimated light source and the camera system are easy to handle in the rendering system, the difficult task here becomes dealing with the intersection between light beam and the surface of sample. According to different applications, there are lots of available methods that can be used. The physical optical methods are Kirchhoff approximation [8], small perturbation method [9], tangent plane approximation [10] and surface integral equations method [11]. The geometrical optical method is ray tracing [12]. Although the physical optical methods, treating light as an electromagnetic field, describe the scattering precisely in the micrometer region, the simulation time is quite long. Therefore, the ray tracing method is firstly applied here to test its feasibility.

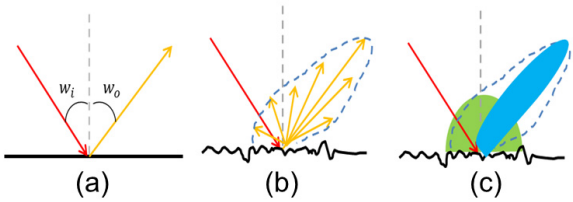


Figure 2: (a): Reflection on mirrorlike surface. (b): Scattering on rough surface. (c): Fitted analytical BRDF Model for specific scattering on rough surface.

Figure 2 (a) shows a total reflection situation, when the material is mirror-like. However, in most cases, the surface is rough and the outgoing light will be scattered into different directions (Fig. 2 (b)) and the reflected energy distribution can be measured by a BRDF setup. To sim-

plify the computation of the BRDF, analytical BRDF models are fit into the measured energy distribution of the scattered light (blue dashed line in Fig. 2 (c)).

By the implementation of BRDF on the large objects, the rendered image looks always real. However, when the BRDF models are applied for rendering the microstructure of metal, the parameters of BRDF are hard to be optimized in order to obtain best rendering images [13]. The reason is, that in the micro-scale range each microfacet should be treated as a flat reflective surface. The addiation of the each microfacet's contribution defines the BRDF in this area. Therefore, the different arrangement of the microfacets of the material provides the variable BRDF and forms different scattering effect [14]. Based on the above considerations, we compared the simulation results with the measured data for the calibration standards made from chrome steel 100 Cr6. Three samples with different roughnesses Ra $= 1\,\mu$m, Ra $= 4\,\mu$m and Ra $= 10\,\mu$m have been tested, see Fig. 3. During the simulation, each microfacet has been considered as a flat surface, the reflectance was calculated by Fresnel's equations.

Figure 3: Calibration standards with the different roughness Ra $= 1\,\mu$m, Ra $= 4\,\mu$m and Ra $= 10\,\mu$m (HALLE Praez.-Kalibriernormale GmbH).

Since most metal surfaces have an anisotropic BRDF, it is also important to simulate the sample at different viewing angles. Hence, the images of three samples were simulated by rotating the sample surface around the z axis, which is perpendicular to the whole stage, for three different angles. Figures 4, 5 and 6 show the realistic images and the corresponding simulation results. The entire highlight effect of the rendered images has the same alterations with the various viewing angles compared to the realistic images. For instance, the tested material has

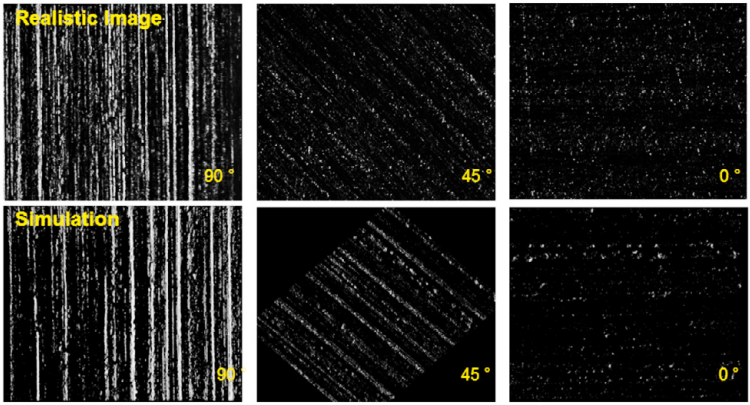


Figure 4: Realistic image captured by camera and simulation images of the calibration standards of Ra $= 1$ μm in Fig. 3 by rotating sample with three different angles along the axis, which is perpendicular to the xy-plane (plane of sample).

.

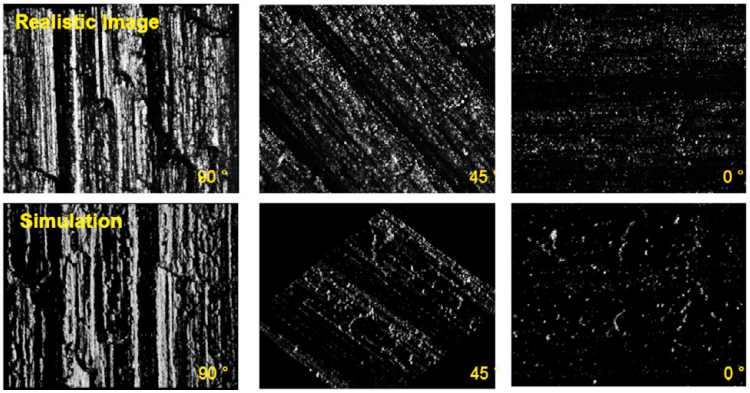


Figure 5: Realistic image captured by camera and simulation images of the calibration standards of Ra $= 4$ μm in Fig. 3 by rotating sample with three different angles along the axis, which is perpendicular to the xy-plane (plane of sample).

.

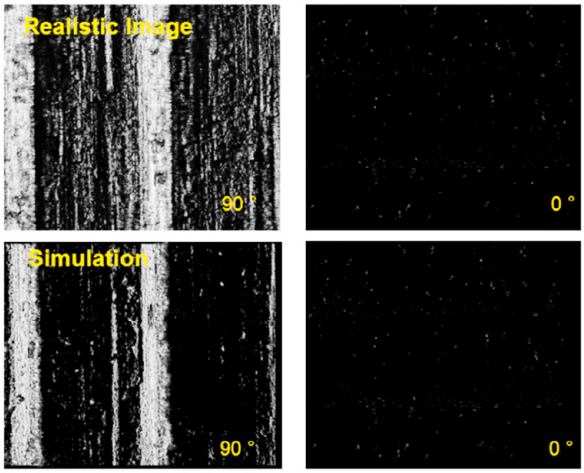


Figure 6: Realistic image captured by camera and simulation images of the calibration standards of Ra $= 10$ μm in Fig. 3 by rotating sample with two different angles along the axis, which is perpendicular to the xy-plane (plane of sample).

.

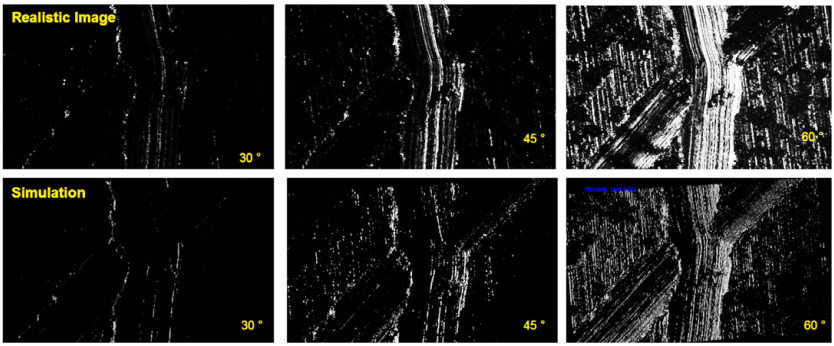


Figure 7: Realistic images captured by camera and simulation images for a scratch on stainless steel with different incident angles of the light source.

the high light when the rotation angle is 0°, whereas it has the lowest reflectance at 90°. Nevertheless, it is important to compare the same position of the realistic and simulated images on the object. Since the microstructure of the sample with Ra $= 1\,\mu$m is too dense, it is hard to find out the same position of the realistic image and the simulated image. In addition, one cross scratch on the stainless steel at different incident angles was also tested. The same conclusion can be indicated from Fig. 7.

3.2 Scalar diffraction approximation theory

Geometrical ray tracing can describe the light propagation from the light source, through the reflection or refraction of the media to the camera in a clear way. The last chapter shows, that the metal surface in the micro-scale range can be rendered acceptably by the ray tracing method. However, when the light interacts with a grating, the ray cannot be regarded as a simple vector any more. To solve these problems, physical-optics-based simulation is needed. Nowadays, there are plenty of numerical methods to simulate these phenomena, such as FDTD, FEM and RCWA. But the computational cost of these methods is very high.

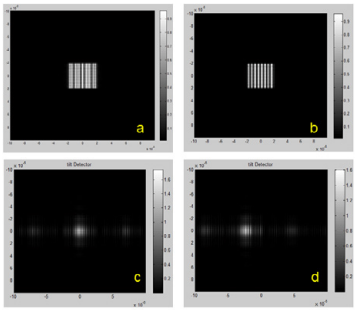


Figure 8: (a): Simulation image of grating with incident angle 0°. (b): Incident angle 45°. (c): Simulation image at 20 mm away from the grating with incident angle 0°. (d): Incident angle 45°.

To address the computational problem and keep the accuracy of the simulation of wave optical phenomena at the same time, the scalar

diffraction approximation method [15] becomes a fairly good choice. This method combines the geometric and diffractive field tracing. In the homogeneous media region, the wave optics method is applied. In the inhomogeneous region, the beam propagates like a ray but with an additional phase shift along the optical path. In our study, we have applied two scalar diffraction approximation methods, thin element approximation (TEA) and local plane-interface approximation (LPIA).

Figure 8 shows a rendered image of a grating by means of TEA [16]. The simulated image of the grating at the camera position with two different incident angles are shown in Fig. 8 (a) and (b). Here we can see, TEA can render the diffraction light after the diffraction element. Figure 8 (c) and (d) show the diffracted light after the grating by using TEA. However, due to TEA's strong approximation, its application is strictly limited. For example, it cannot show the darkness of the blocked area. Figure 9 shows the highlight spot in the scratch area with TEA method, but the same position in the real image the area is dark. The reason is, that TEA only calculates the first reflection on the object, and assumes that all these reflected rays will leave the object. When the refleced rays are blocked again by the objects (shadow effect), the result of TEA shows some error.

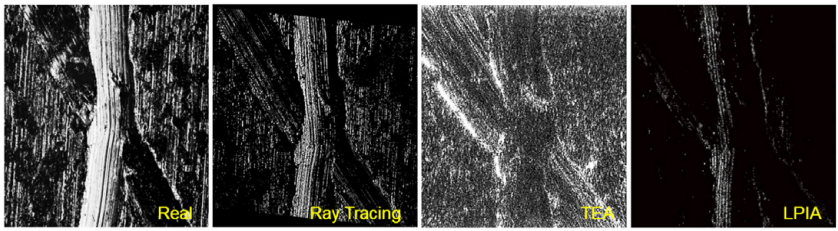


Figure 9: Comparsion among realistic image, rendered image by ray tracing method, rendered image by TEA method and rendered image by LPIA method.

In order to render the shadow effect, LPIA [17] was also used here. Unlike TEA, LPIA takes the impact of the slope of the sample's surface together with the second intersection of the light with the sample into account. LPIA in Fig. 9 shows the shadow effect in the area of the scratch, whereas the same area with TEA method is light. However,

LPIA loses a lot of details compared to the real image as well. This is because the rays in the scratch area reflect normally couple of times. The phase difference between adjacent rays would due to the multiple reflections become very large. This huge phase difference leads to a huge fluctuation of the wavefront. However, in the real case, this fluctuation does not exist. Therefore, LPIA works well, only if the adjacent rays do not intersect with each other during the propagation.

4 Conclusions

We have shown several virtual surface defect rendering methods in the micro scale. The ray tracing method can provide a fairly realistic rendered image for typical metal surfaces. In case of the diffractive element, the scalar diffraction approximation methods, such as TEA and LPIA, are used. These methods work well if the sample is only dominated by the diffractive effect. If the profile of the sample is too irregular, such as the grating incorportated with a deep hole, these scalar approximation methods work not any more. In this case, more rigorous simulation methods are needed.

References

1. Y. Ai and K. Xu, "Feature extraction based on contourlet transform and its application to surface inspection of metals," *Optical Engineering*, vol. 51, no. 11, p. 113605, 2012.
2. W. Lyda, A. Burla, T. Haist, and W. Osten, "Implementation and analysis of an automated multiscale measurement strategy for wafer scale inspection of micro electromechanical systems," *International Journal of Precision Engineering and Manufacturing*, vol. 13, no. 4, pp. 483–489, 2012.
3. J.-H. Kim, J. Im, C.-H. Kim, and J. Lee, "Subtle features of ice with cloudy effects and scratches from collision damage," *Computer Animation and Virtual Worlds*, vol. 27, no. 3-4, 2016.
4. M.-P. Cani and M. Slater, "A physically-based model for rendering realistic scratches," *EUROGRAPHICS 2004*, vol. 23, no. 3, 2004.
5. M. Schwärzler and M. Wimmer, "Rendering imperfections: Dust, scratches, aging,..." Institute of Computer Graphics and Algorithms, Vienna University of Technology, Tech. Rep. TR-186-2-07-09, 2007.

6. K. Nagano, G. Fyffe, O. Alexander, J. Barbic, H. Li, A. Ghosh, and P. Debevec, "Skin microstructure deformation with displacement map convolution," *ACM Trans.Graph.*, vol. 34, no. 4, p. 109:1–109:10, 2015.
7. F. Mauch, D. Fleischle, W. Lyda, W. Osten, T. Krug, and R. Häring, "Combining rigorous diffraction calculation and GPU accelerated nonsequential raytracing for high precision simulation of a linear grating spectrometer," in *Proc. SPIE, 8083, 80830F-2*, 2011.
8. L. Tsang, J. A. Kong, and K.-H. Ding, *Scattering of Electromagnetic Waves, Theories and Applications*. John Wiley and Sons, 2002.
9. S. Afifi, R. Dusseaux, and A. Berrouk, "Electromagnetic scattering from 3D layered structures with randomly rough interfaces: Analysis with the small perturbation method and the small slope approximation," *IEEE Transactions on Antennas and Propagation*, vol. 62, no. 10, 2014.
10. A. G. Voronovich, "Tangent plane approximation and some of its generalizations," *Acoustical Physics*, vol. 53, no. 3, pp. 298–304, 2007.
11. L. Fu, K. Frenner, and W. Osten, "Rigorous speckle simulation using surface integral equations and higher order boundary element method," *Optics Letters*, vol. 39, no. 14, p. 4104, 2014. [Online]. Available: http://dx.doi.org/10.1364/OL.39.004104
12. A. Ngan, F. Durand, and W. Matusik, "Experimental analysis of BRDF models," *Eurographics Symposium on Rendering 2005*, 2005.
13. H. Yang, T. Haist, M. Gronle, and W. Osten, "Realistic simulation of camera images of local surface defects in the context of multi-sensor inspection systems," in *Proc. SPIE 9525*, Munich, Germany, 2015.
14. Z. Dong, B. Walter, S. Marschner, and D. Greenberg, "Predicting appearance from measured microgeometry of metal surfaces," *Proc. SPIE 9525*, vol. 35, no. 1, pp. 9:1 – 9:13, 2015.
15. F. Wyrowsk, H. Zhong, S. Zhang, and C. Hellmann, "Approximate solution of Maxwell's equations by geometrical optics," in *Proc. SPIE 9630*, Jena, Germany, 2015.
16. H. Zhong, S. Zhang, and F. Wyrowski, "Parabasal thin-element approximation approach for the analysis of microstructured interfaces and freeform surfaces," *Journal of the Optical Society of America A*, vol. 32, no. 1, pp. 124–129, 2015.
17. A. Pfeil, F. Wyrowski, A. Drauschke, and H. Aagedal, "Analysis of optical elements with the local plane-interface approximation," *Appl Opt*, vol. 39, no. 19, pp. 3304–13, 2000.

Erfassung und Verarbeitung von Lichttransportmatrizen zur automatischen Sichtprüfung transparenter Objekte

Acquisition and processing of light transport matrices for automated transparent object inspection

Johannes Meyer[1], Thomas Längle[2] und Jürgen Beyerer[2]

[1] Karlsruher Institut für Technologie,
Lehrstuhl für Interaktive Echtzeitsysteme,
Adenauerring 4, 76131 Karlsruhe
[2] Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung IOSB,
Abteilung Sichtprüfsysteme,
Fraunhoferstr. 1, 76131 Karlsruhe

Zusammenfassung Transparente Materialien kommen in diversen Produkten zum Einsatz und müssen zur korrekten Erfüllung ihres Zwecks oft hohen Qualitätsansprüchen genügen. Dazu müssen diese Materialien insbesondere frei sein von sogenannten streuenden Defekten wie beispielsweise eingeschlossenen Luftblasen. Prüfsysteme auf Basis von Dunkelfeldanordnungen sind prinzipiell in der Lage, diese Defekte abzubilden, haben jedoch einen hohen manuellen Einrichtungsaufwand. In diesem Artikel wird dargelegt, wie Lichttransportmatrizen für ein optisches System bestehend aus einer programmierbaren Flächenlichtquelle und einer telezentrischen Kamera berechnet werden können. Es werden zwei Merkmale vorgestellt, die aus diesen Matrizen extrahiert werden können und die Abbildung streuender Defekte in transparenten Objekten ermöglichen, ohne dass das System speziell an den konkret vorliegenden Prüfling angepasst werden muss. Synthetische Experimente mit Hilfe eines physikalisch basierten und entsprechend erweiterten Renderingframeworks erlaubten eine positive Evaluation des Ansatzes und bestätigten eine gewisse Überlegenheit gegenüber klassischen Inspektionssystemen.

Schlagwörter Sichtprüfung transparenter Objekte, Bildverarbeitung, Lichttransportmatrizen.

Abstract Transparent materials are employed for creating different kinds of products and have to meet high quality requirements. First of all, transparent materials have to be free from so-called scattering defects, e. g., enclosed air bubbles. Visual inspection systems based on dark field setups are principally capable of imaging these kinds of defects, however, it usually requires much effort to adapt them to the test object on hand. This article shows how light transport matrices can be calculated for an optical system consisting of a programmable area light source and a telecentric camera. Two features are proposed that can be extracted out of these matrices and that allow to image scattering defects present in a transparent object without the need of adapting the system to the actual test object. The results of synthetic experiments obtained using a physically based and adequately extended rendering framework approved the proposed approach and showed that it even outperforms classical inspection systems in some situations.

Keywords Transparent object inspection, image processing, light transport matrices.

1 Einleitung

Transparente Materialien spielen in vielen Industriezweigen eine entscheidende Rolle. Sie kommen beispielsweise als Windschutzscheiben von Automobilen und Flugzeugen zum Einsatz, werden in medizinischen Augenoperationen zur Führung von Laserstrahlen an die gewünschte Position verwendet und sie sind Teil hochpräziser optischer Bauteile. Die genannten Anwendungsgebiete machen deutlich, dass diese Materialien hohen Qualitätsansprüchen genügen müssen.

Typische Fehler, die bei der Herstellung transparenter Objekte auftreten können, sind eingeschlossene absorbierende Partikel, eingeschlossene Luftblasen, Oberflächenkratzer, Fehler bezüglich der 3D-Form oder Abweichungen des Brechungsindexes. Je nach Anwendungsgebiet können diese Fehler gravierende Folgen haben, weswegen eine Qualitätskontrolle in Form einer Sichtprüfung unabdingbar ist. Für Menschen ist diese Prüfaufgabe ermüdend und fehleranfällig, weshalb automatisierte Lösungen gefunden werden müssen.

Für einige der genannten Defekttypen existieren bereits ausgereifte automatisierte Inspektionsverfahren [1–5]. Die Detektion von streuenden Defekten (eingeschlossene Luftblasen, Schlieren auf der Oberfläche) ist ein schwieriges Problem. Diese Defekte äußern sich nicht in einer Intensitätsänderung des transmittierten Lichts, sondern in der Verteilung der Richtung ausgehender Lichtstrahlen. Je nach Größe des Defekts führen unterschiedliche physikalische Effekte wie Mie-Streuung oder Rayleigh-Streuung dazu, dass inzidente Lichtstrahlen aufgefächert oder abgelenkt werden [6,7]. Ein gängiger Ansatz zur Abbildung solcher Defekte sind sogenannte Dunkelfeldanordnungen [8,9]. Dabei wird das Prüfobjekt so beleuchtet, dass unter Abwesenheit streuender Defekte möglichst wenig Licht die Kamera erreicht. Ist hingegen ein streuender Defekt vorhanden, so streut er das Licht der Beleuchtung in die Kamera und der Defekt wird sichtbar. Ein großer Nachteil dieser Verfahren ist, dass die relative Anordnung von Kamera, Beleuchtung und Prüfobjekt genau abgestimmt sein muss und häufig mit einem hohen empirischen Aufwand einhergeht. Daher sind diese Ansätze beispielsweise für Sichtprüfungsanwendungen mit einer hohen und variablen Teilevielfalt nicht oder nur beschränkt einsetzbar.

Dieser Artikel führt eine neue Methode ein, mit der anhand der Richtungsinformation des Lichts streuende Defekte in transparenten Objekten abgebildet werden können, ohne dass die Beleuchtungs- und Bildaufnahmekonstellation an die Prüflingsgeometrie angepasst werden muss. Der Ansatz basiert auf der Theorie sogenannter Lichttransportmatrizen (LTMs) [10–15]. In Abschnitt 2 wird der optische Aufbau der Methode vorgestellt und die LTM definiert. Zur späteren Evaluation des Verfahrens wurde ein physikalisch basierter Renderer verwendet. Das entsprechende Framework wurde wie in Abschnitt 3 gezeigt erweitert, sodass sich damit komfortabel und effizient LTMs für die simulierten Szenen berechnen lassen. In Abschnitt 4 werden zwei Merkmale vorgestellt, die aus LTMs berechnet werden können und für die angestrebte Defektdetektion geeignet sind. Abschnitt 5 ist den durchgeführten Experimenten gewidmet und Abschnitt 6 schließt den Artikel mit einer Zusammenfassung und einem Ausblick.

2 Optischer Aufbau

Dieser Beitrag schlägt einen optischen Aufbau basierend auf einer telezentrischen Kamera und einer örtlich programmierbaren Flächenlichtquelle vor. Abbildung 1 zeigt den schematischen Aufbau des optischen Systems. Die programmierbare Lichtquelle könnte beispielsweise ein Matrixdisplay sein, bei dem sich einzelne Pixel gezielt ein- und ausschalten lassen. Im gezeigten Beispiel wird eine Doppelkonvexlinse inspiziert. Die telezentrische Kamera registriert nur solche

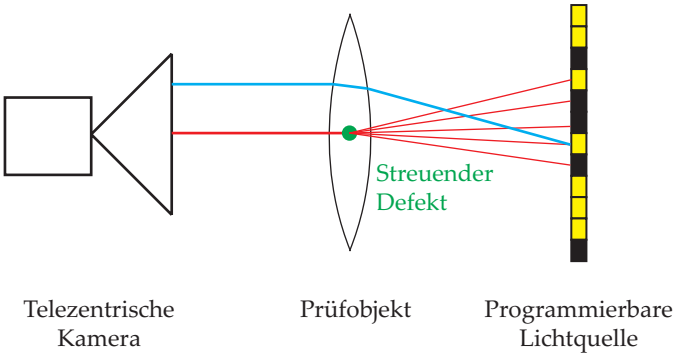


Abbildung 1: Optischer Aufbau des vorgeschlagenen Ansatzes: Das Prüfobjekt befindet sich im Fokus einer telezentrischen Kamera. Die Beleuchtung erfolgt durch eine örtlich programmierbare Flächenlichtquelle. Betrachtet ein Sichtstrahl der Kamera einen defektfreien Bereich des Prüflings (blauer Strahlengang), so trifft er auf nur wenige Pixel der Lichtquelle. Trifft der Sichtstrahl hingegen auf einen streuenden Defekt (roter Strahlengang), so erreicht er mehrere, örtlich verteilte Lichtquellenpixel.

Lichtstrahlen, die annähernd parallel zur optischen Achse verlaufen [9]. Für Kamerapixel, die einen defektfreien Bereich des Prüflings abbilden, verlaufen die Sichtstrahlen eng gebündelt durch das optische System und treffen auf nur wenige Pixel der Lichtquelle. Liegt hingegen ein streuender Defekt vor, so werden viele Lichtstrahlen von unterschiedlichen Pixel der Lichtquelle gestreut, sodass auch immer ein Anteil par-

allel zur optischen Achse verläuft und von der telezentrischen Kamera eingefangen wird. Es ist also zu erwarten, dass an dem Signal, das Kamerapixel registrieren, die einen streuenden Defekt abbilden, mehrere Pixel der Lichtquelle beteiligt sind. Im weiteren Verlauf des Artikels wird gezeigt, wie sich dieser Zusammenhang zur Detektion streuender Defekte nutzen lässt.

Lichttransportmatrix Im vorgestellten optischen Aufbau kann prinzipiell jedes Kamerapixel $\mathbf{k} = (k_m, k_n)^{\mathrm{T}}$ von jedem Lichtquellenpixel $\mathbf{l} = (l_p, l_q)^{\mathrm{T}}$ einen Beitrag zum empfangenen Signal erhalten. Somit lässt sich für eine Kamera mit $M \times N$ Pixel und eine programmierbare Lichtquelle mit $P \times Q$ Pixel der komplette Lichttransport der Szene mittels der sogenannten Lichttransportmatrix $\mathbf{T}$ darstellen:

$$\mathbf{T} = \begin{pmatrix} \mathbf{c}(1,1)^{\mathrm{T}} \\ \mathbf{c}(1,2)^{\mathrm{T}} \\ \vdots \\ \mathbf{c}(M,N-1)^{\mathrm{T}} \\ \mathbf{c}(M,N)^{\mathrm{T}} \end{pmatrix}, \tag{1}$$

wobei die Zeilenvektoren $\mathbf{c}(m,n)^{\mathrm{T}}$ sogenannte Korrespondenzvektoren darstellen [11]. Ein Korrespondenzvektor $\mathbf{c}(m,n) = (c_1(m,n), c_2(m,n), \ldots, c_{PQ-1}(m,n), c_{PQ}(m,n))^{\mathrm{T}}$ beinhaltet die Signalanteile der einzelnen PQ Lichtquellenpixel zum Kamerapixel $(k_m, k_n)^{\mathrm{T}}$. Jede Zeile von $\mathbf{T}$ kodiert also das Zustandekommen des Signals am entsprechenden Kamerapixel.

Durch Kenntnis von $\mathbf{T}$ einer Szene lasst sich das Kamerabild $\mathbf{y}$ fur ein beliebiges Beleuchtungsmuster $\mathbf{x}$ der Lichtquelle synthetisch durch eine einfache Matrix-Vektor-Multiplikation berechnen:

$$\mathbf{y} = \mathbf{T}\mathbf{x}\,. \tag{2}$$

Der folgende Abschnitt zeigt, wie sich LTMs für simulierte Szenen approximieren lassen.

3 Erweiterung des Renderingframeworks zur Approximation von Lichttransportmatrizen

Um den vorgeschlagenen Ansatz mit möglichst geringem Aufwand evaluieren zu können, wird im Rahmen dieses Artikels das physikalisch basierte Renderingframework Mitsuba verwendet und entsprechend erweitert [16]. Abbildung 2 zeigt den schematischen Aufbau des Frameworks. Die in Schwarz gezeichneten Komponenten sind die Basisbestandteile von Mitsuba. Um das Bild eines Sensors (z. B. einer Kamera) für eine modellierte Szene zu berechnen, die aus verschiedenen Objekten und Lichtquellen bestehen kann, laufen folgende Schritte ab: Die Hauptkomponente des Frameworks, der Renderer, wählt nach einer bestimmten Strategie ein Sensorsample (Pixel der Kamera) und ein Aperturesample aus. Die Sensor-Komponente berechnet aus diesen Samples einen Sichtstrahl, dessen Verlauf der Renderer durch die Szene soweit verfolgt, bis er nicht mehr weiterreflektiert wird oder auf eine Lichtquelle trifft. Strahlt die Lichtquelle Licht in die Richtung des Sichtstrahls ab, so wird die entsprechende Strahldichte entlang des optischen Pfads des Sichtstrahls zurückpropagiert, mit den Reflektanzspektren eventuell beteiligter Objektoberflächen verrechnet und schließlich zusammen

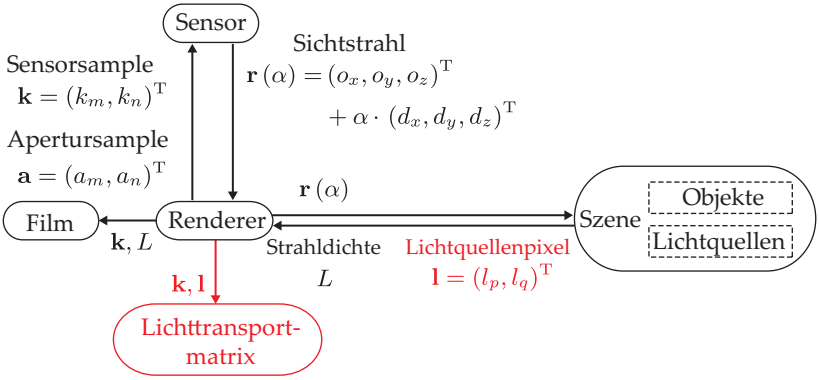


Abbildung 2: Schematischer Aufbau des verwendeten Renderingframeworks. Die in Schwarz gezeichneten Grundbestandteile des Mitsuba-Renderers wurden um die rot markierten Komponenten im Rahmen dieses Artikels ergänzt.

mit dem Sensorsample der Filmkomponente übergeben. Diese Komponente aggregiert sukzessive alle Paare aus Sensorsamples und Strahldichten, um letztendlich das simulierte Sensorbild zu erzeugen.

Um nun für eine gegebene Szene die Lichttransportmatrix berechnen zu können, wurde das Mitsuba Renderingframework um die in Abb. 2 rot gekennzeichneten Komponenten erweitert. Trifft ein Lichtstrahl auf den Pixel $(l_p, l_q)^{\mathrm{T}}$ einer programmierbaren Lichtquelle, so transportiert er diese Information zusätzlich zu der Strahldichte L. Der Renderer kann nun anhand der Paare aus Sensorsamples $\mathbf{k}$ und Lichtquellenpixel $\mathbf{l}$ fortlaufend die Korrespondenzvektoren $\mathbf{c}$ befüllen und schließlich die Lichttransportmatrix $\mathbf{T}$ wie in Absch. 2 gezeigt zusammensetzen.

4 Extraktion von Merkmalen aus Lichttransportmatrizen

Aus Lichttransportmatrizen, die für Prüfszenen wie in Abb. 1 berechnet wurden, können geeignete Merkmale extrahiert werden, die eine Abbildung von streuenden Materialfehlern in transparenten Objekten ermöglichen. Im Folgenden werden zwei solcher Merkmale beschrieben.

4.1 Merkmal ScatterCount

Wie in Absch. 2 beschrieben, führen streuende Defekte dazu, dass die Sichtstrahlen der telezentrischen Kamera aufgefächert werden und mehrere Lichtquellenpixel erreichen, bzw., dass die Lichtstrahlen vieler Lichtquellenpixel so gebrochen werden, dass sie parallel zur optischen Achse verlaufen und von der telezentrischen Kamera erfasst werden. Für ein Kamerapixel $(k_m, k_n)^{\mathrm{T}}$, das einen streuenden Defekt abbildet, bedeutet das, dass viele Komponenten des entsprechenden Korrespondenzvektors $\mathbf{c}(m, n)$ größer Null sind.

Das Merkmal *ScatterCount*

$$sc(m, n) := |\{i \in [1, \ldots, PQ] : \mathbf{c}_i(m, n) \geq 0\}| \tag{3}$$

enthält für jedes Kamerapixel $(k_m, k_n)^{\mathrm{T}}$ die Anzahl an Komponenten von $\mathbf{c}(m, n)$, die größer sind als Null.

4.2 Merkmal ScatterWidth

Neben der vom vorher beschriebenen Merkmal erfassten Anzahl an unterschiedlichen Lichtquellenpixel, die zum Signal eines Kamerapixel beitragen, spielt auch die maximale Entfernung, die sog. *ScatterWidth* zwischen den beteiligten Pixel auf der Lichtquelle, eine Rolle. Stärker streuende Materialdefekt führen beispielsweise dazu, dass eine größere Region von Lichtquellenpixel zu dem entsprechenden Kamerasignal beiträgt. Das Merkmal *ScatterWidth* ist definiert als:

$$sw(m,n) := \begin{cases} 0, & \text{falls } \max \mathbf{c}(m,n) = 0, \\ \Omega - A & \text{sonst}, \end{cases} \tag{4}$$

mit

$$A, \Omega \in [1, \ldots, PQ]\,, \tag{5}$$

$$\forall i \in [1, \ldots, A-1] : \mathbf{c}_i(m,n) = 0, \tag{6}$$

$$\forall i \in [\Omega+1, \ldots, PQ] : \mathbf{c}_i(m,n) = 0, \tag{7}$$

$$\mathbf{c}_A(m,n) > 0, \mathbf{c}_\Omega(m,n) > 0\,, \tag{8}$$

A ist also der Index des ersten und Ω der Index des letzten Eintrags in $\mathbf{c}(m,n)$, der größer Null ist.

5 Experimente

Zur Evaluation des vorgeschlagenen Verfahrens kamen die in Abb. 1 gezeigte Prüfszene und das wie in Absch. 3 beschriebene, erweiterte Renderingframework Mitsuba zum Einsatz. Der Prüfling, eine Doppelkonvexlinse, stellt für herkömmliche Prüfansätze eine besondere Herausforderung dar, da er selbst ein optisches Element mit abbildenden Eigenschaften ist, die beim Design des Prüfsystems beachtet werden müssen. Es wurden vier Instanzen der Prüfszene angelegt. In der ersten Prüfszene ist der Prüfling defektfrei. In den weiteren drei Prüfszenen wurde das Prüfobjekt jeweils mit drei eingeschlossene Luftblasen versehen. Die Größe der Defekte wurde für jede Prüfszene variiert, dabei hat die kleinste Defektinstanz die Größe der Projektion eines halben Kamerapixel. Die Positionen der Defekte im Prüfling wurden für

alle Prüfszenen gleich gewählt: Ein Defekt befindet sich im Zentrum der Doppelkonvexlinse und die weiteren zwei Defekte wurden mit äquidistanten Schritten zum linken Rand des Prüflings verschoben.

Um eine Vergleichbarkeit der Ergebnisse zu erreichen, wurde zusätzlich ein konventionelles Prüfsystem bestehend aus einer telezentrischen Kamera und einer uniformen Flächenbeleuchtung simuliert.

Abbildung 3 zeigt eine Übersicht der Ergebnisse des Experiments. Das konventionelle Prüfsystem ist in der Lage, die Defekte aller simulierten Größen abzubilden, wobei die Abbildungen der kleinsten Defekte einen geringen Kontrast aufweisen und kaum noch wahrnehmbar sind. Im Gegensatz dazu bilden beide Merkmale des vorgestellten Verfahrens selbst die kleinsten Defekte mit hohem Kontrast ab, was für eine nachgeschaltete Defektdetektion von Vorteil wäre. Insbesondere in dem für das Merkmal *ScatterWidth* berechneten Inspektionsbild sind die Defekte deutlich zu erkennen. Die vereinzelt auftretenden weißen Punkte, die in den Ergebnisbildern des Merkmals *ScatterWidth* zu sehen sind, sind auf das Samplingrauschen des Renderingframeworks zurückzuführen.

6 Zusammenfassung

Dieser Artikel legt dar, wie ein optisches System bestehend aus einer telezentrischen Kamera und einer örtlich programmierbaren Flächenlichtquelle mit Methoden basierend auf der Theorie der Lichttransportmatrizen kombiniert werden kann, um ein Prüfsystem zur Inspektion transparenter Objekte in Hinblick auf streuende Materialdefekte zu erhalten. Des Weiteren wurde gezeigt, dass sich streuende Defekte in transparenten Materialien hauptsächlich in einer Änderung der Richtungsverteilung inzidenter Lichtstrahlen und nicht in einer Änderung der Lichtintensität äußern. Die Lichttransportmatrix einer Szene enthält die Information, welche Lichtquelle zu welchem Anteil an den von den Sensorpixeln erfassten Signalen beteiligt sind. In dieser Information ist für das vorgestellte optische System insbesondere auch die Richtung der erfassten Lichtstrahlen enthalten.

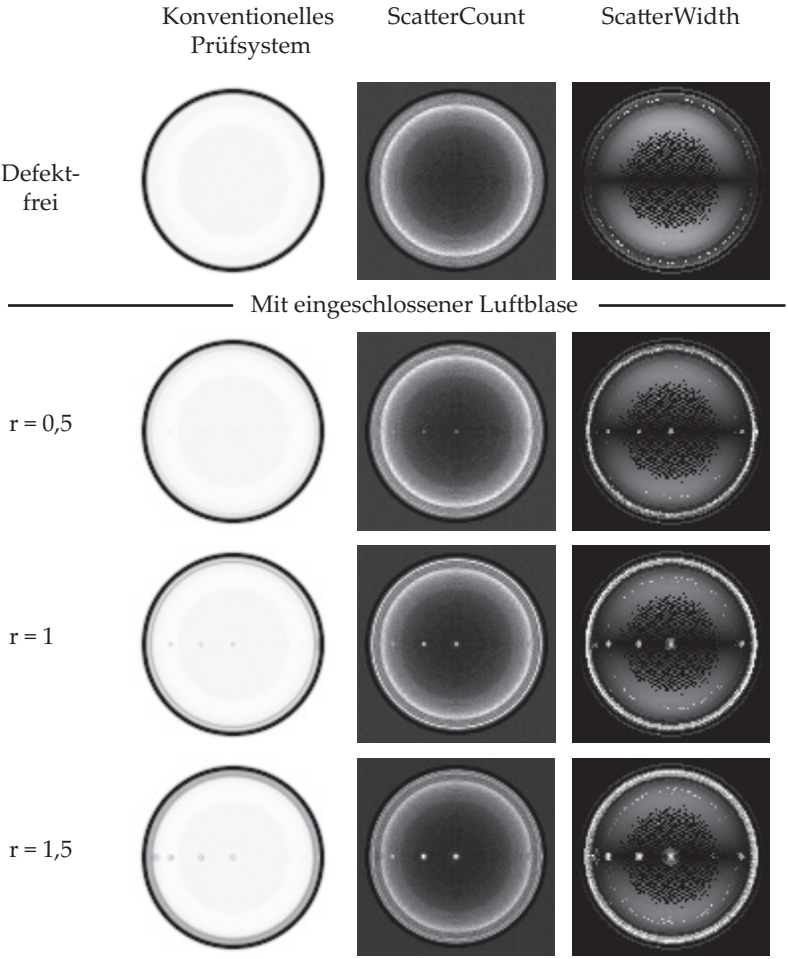


Abbildung 3: Simulierte Inspektionsbilder des konventionellen Prüfsystems und Falschfarbendarstellungen der auf den approximierten Lichttransportmatrizen basierenden Merkmalen *ScatterCount* und *ScatterWidth*. Die Größe r der eingebrachten Defekte ist als Faktor bezüglich der Fläche eines in die Szene abgebildeten Kamerapixel zu verstehen.

Im Rahmen dieses Artikels wurden zwei Merkmale vorgestellt, die auf Basis der Lichttransportmatrix berechnet werden und zur Detektion von streuenden Materialdefekten in transparenten Objekten verwendet werden können. Ein physikalisch basiertes Renderingframework wurde so angepasst, dass sich damit effizient Lichttransportmatrizen für synthetische Prüfszenen approximieren ließen. Anhand dieser Simulationen konnte gezeigt werden, dass der Ansatz zur Prüfung transparenter Objekte auf streuende Defekte geeignet ist und sogar konventionellen Prüfsystemen überlegen sein könnte.

Als nächsten Schritt planen die Autoren eine praktische Umsetzung des Verfahrens und die Durchführung realer Experimente. Hierbei muss insbesondere untersucht werden, wie sich die relevanten Teile der Lichttransportmatrix effizient optisch approximieren lassen. Des Weiteren könnten weitere Merkmale in Bezug auf die Lichttransportmatrizen definiert werden. Beispielsweise berücksichtigt das vorgestellte Merkmal *ScatterWidth* bisher nicht die zweidimensionale Struktur der programmierbaren Lichtquelle. Durch Auswerten dieser zusätzlichen Information könnte die Nützlichkeit des Merkmals erhöht werden und es wäre denkbar, dass sich damit auch verschiedene Defektausprägungen unterscheiden ließen.

Literatur

1. M. Hartrumpf und R. Heintz, „Device and method for the classification of transparent components in a material flow“, Patent WO 2009/049594 A1, 2009.
2. M. Hartrumpf, K.-U. Vieth, T. Längle und G. Struck, „Neues Verfahren zur Sichtprüfung transparenter Materialien“, in *Sensorgestützte Sortierung*, 2008, S. 57–58.
3. J. Meyer, „Visual inspection of transparent objects – physical basics, existing methods and novel ideas“, Karlsruhe Institute of Technology, Tech. Rep. IES-2014-04, 2014.
4. ——, „Overview on machine vision methods for finding defects in transparent objects“, Karlsruhe Institute of Technology, Tech. Rep. IES-2015-08, 2015.

5. S. Chatterjee, „Determination of refractive index in-homogeneity of transparent, isotropic optical materials“, in *Advances in Optical Science and Engineering*. Springer, 2015, S. 61–66.
6. C. F. Bohren und D. R. Huffman, *Absorption and Scattering of Light by Small Particles*. Wiley-VCH Verlag GmbH, 2007.
7. H. van de Hulst, *Light Scattering by Small Particles*, Ser. Dover Books on Physics. Dover Publications, 1957.
8. J. Beyerer, F. Puente León und C. Frese, *Machine Vision: Automated Visual Inspection: Theory, Practice and Applications*. Springer Berlin Heidelberg, 2015.
9. J. Beyerer, F. P. León und C. Frese, *Automatische Sichtprüfung: Grundlagen, Methoden und Praxis der Bildgewinnung und Bildauswertung*, 2. Aufl. Springer Vieweg, 2016.
10. J. Bai, M. Chandraker, T.-T. Ng und R. Ramamoorthi, „A dual theory of inverse and forward light transport“, in *Computer Vision–ECCV 2010*. Springer, 2010, S. 294–307.
11. M. O'Toole und K. N. Kutulakos, „Optical computing for fast light transport analysis.“ *ACM Trans. Graph.*, Vol. 29, Nr. 6, S. 164, 2010.
12. J. Wang, Y. Dong, X. Tong, Z. Lin und B. Guo, „Kernel Nyström method for light transport“, in *ACM Transactions on Graphics (TOG)*, Vol. 28. ACM, 2009, S. 29.
13. M. Chandraker, J. Bai, T.-T. Ng und R. Ramamoorthi, „On the duality of forward and inverse light transport“, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 33, Nr. 10, S. 2122–2128, 2011.
14. S. M. Seitz, Y. Matsushita und K. N. Kutulakos, „A theory of inverse light transport“, in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, Vol. 2. IEEE, 2005, S. 1440–1447.
15. M. O'Toole, J. Mather und K. Kutulakos, „3d shape and indirect appearance by structured light transport“, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, S. 3246–3253.
16. W. Jakob, „Mitsuba renderer“, 2010, http://www.mitsuba-renderer.org.

Extraction of regular textures from real images

Pilar Hernández Mesa and Fernando Puente León

Karlsruhe Institute of Technology,
Institute of Industrial Information Technology,
Hertzstraße 16, Bldg. 06.35, 76187 Karlsruhe

Abstract A texel repeated at equal distances over space creates a regular texture. Such textures appear everywhere, in nature (e. g., the trunk of a palm tree) as well as out of it (e. g., tiles of floors). However, regular textures are rarely found in real images, due to, i. a., lighting variations and the projection from the three dimensions onto two. The detection and processing of such near-regular textures is a challenging task with many application fields. In this paper different methods are considered and compared to detect near-regular textures from real images and to extract their regular textures.

Keywords Near-regular textures, detection, texel extraction.

1 Introduction

Textures can be roughly classified into regular and irregular, if they are of random nature [1]. Regular textures are seldom found in images, whereas near-regular textures [2] are more common. The detection and the extraction of their regular texture is a challenging task with many application fields like geotagging [3], content-based image retrieval [4], and texture synthesis [2, 5]. In this work the extraction of regular textures from surfaces in real images is considered. For this purpose, the repeated texel must be extracted from the image. The approach that we follow can be divided into three main blocks: the extraction of characteristic points and the sorting of repeating characteristic points into groups, the extraction of the median texel per group, and additional post-processing. This modus operandi is inspired by [5]. For the resolution of the goal of each block, several different conditions and methods are considered here that may even be fused. The results show that

an appropriate selection and fusion of the proposed methods extract 96.49 % of the regular textures in the tested database correctly, without the extraction of false positives.

This paper is organized as follows. In Sec. 2 an overview of existing methods is given. The extraction and sorting of repeating characteristic points into groups (Sec. 3), the extraction of the median texel per group (Sec. 4), and additional post-processing (Sec. 5) explain the considered methods and information to extract the regular textures. The results of the methods are shown in Sec. 6, and conclusions are drawn in Sec. 7.

2 State of the art

Liu et al. analyze in [2] near-regular textures to obtain their synthesis information and to replace them to create other textures. They represent a near-regular texture as a composition of geometric deformations, lighting variations, and color changes. The interaction with a user is necessary in their work. To automatically extract the regular texture information, interest point detectors are used. Hays et al. [6] propose the use of MSER [7] and a normalized cross-correlation method in case that the points of interest are not enough to detect the regular texture. Schindler et al. [3] extract SIFT features first, and group them according to the similarity of their descriptors into N groups. As the features are similar, clusters shall contain repeating points that span texels. Park et al. [8] use a corner detector to extract interest points and group them automatically into clusters of similar points using mean-shift. The benefit is that the number of clusters does not have to be predefined. But in their work, to extract the texel information of the texture, a fixed window size is set at each interest point not considering different scales and orientations of the texels. Hilsmann et al. [5] use SIFT features to obtain the synthesis information of a near-regular texture, which are also grouped by mean-shift. Their texture is searched from three characteristic points that are neighbors and span a L-shaped pair of vectors, which are expanded to a lattice.

3 Sorted characteristic texel points

To extract a regular texture from a near-regular one, the texel of the texture is searched. For this purpose, interest points are obtained from the grey-level image using SURF features [9, 10], which are robust to affine transforms. As the texel appears several times in the texture, we expect many similar features at repeating points. Interest points with different feature vectors may describe other texel configurations. The extracted feature points must therefore be grouped according to their similarity first (Sec. 3.1). Next, the clusters may be processed to separate textures from spatially separated objects (Sec. 3.2), and to eliminate clusters consisting merely of points located at contours (Sec. 3.3).

3.1 Proposed clustering method

Obtaining different clusters from the extracted feature points is an unsupervised learning problem with an unknown number of clusters. To solve this issue, we represent each point in the image as a node of a graph. Edges in the graph are inserted between nodes of interest points if the Euclidean distance of their feature vectors is smaller than a predefined threshold. The connected graphs are the clusters of similar interest points. This method can be seen as a simplified version of the DBSCAN algorithm [11]. In contrast to [11] we do not consider the number of edges per node.

3.2 Compact point clouds

An image containing several objects with a similar grey-level texture will have interest points from different objects at the same cluster. To extract regular textures as close as possible to the textures of each object's surface, the obtained clusters from the previous section are further considered and, if necessary, divided into new clusters with only one dense concentration of points (spatially). For this purpose, the characteristic points in each cluster are considered as nodes and connected by edges to their spatially nearest d_{neigh} neighbors. The connected graphs are the final clusters containing characteristic points with similar features that are spatially agglomerated at one place.

3.3 Eliminate merely contour points

A cluster of characteristic points may only contain points along the contour of an object, not describing the texture of a surface. The former suppression and therefore recognition of such clusters avoids extra processing and time consumption. From all the points in a cluster ($\mathbf{p}_j, 1 \leq j \leq J$), the points that describe the most compact boundary ($\mathbf{p}_r^{\text{bound}}, 1 \leq r \leq R, R < J$) and envelop all the points are searched. The minimum distance from a point in the cluster to the centre of all points is compared to the mean distance of the points at the boundary to the centre:

$$q = \frac{\min_j \sqrt{\mathbf{p}_j - \frac{\sum_j \mathbf{p}_j}{J}}}{\frac{\sum_r \sqrt{\mathbf{p}_r^{\text{bound}} - \frac{\sum_j \mathbf{p}_j}{J}}}{R}}. \tag{1}$$

If q is higher than a predefined threshold, then the cluster is suppresed.

4 Texel extraction

The texels of the regular textures are searched from the clusters containing the repeated characteristic points. Each point in a cluster gets maximally four points assigned as neighbors (Sec. 4.1), which will be used to find possible texels (Sec. 4.2). The expected texel of a regular texture $t^{\text{Exp}}(\mathbf{x})$ is obtained filtering with a median filter a maximal predefined number of texels that are randomly selected. This restriction ensures a faster extraction of the regular texture. Different methods are proposed to suppress incorrectly extracted possible texels (Sec. 4.3).

4.1 Valid neighbors from characteristic points

Once the characteristic points are grouped, each point gets a maximum of four neighbors from their same cluster assigned. The spatially closest point in the cluster with a similar grey-level value becomes the first neighbor of the considered point. Up to four neighbors with similar grey-level values are assigned if the following criteria are fulfilled:

1. The vectors spanned between the considered point and its neighbors lie each in a quadrant of a coordinate system with origin at

the considered point.

2. The spatially closest points to the considered point that fulfill condition 1. are selected.

In contrast to [5] we force neighbor points to have a similar grey-level value to improve a correct assignment of repeated points, as we assume the existence of local regular textures at near-regular textures.

4.2 Possible texels

To obtain the texel of the near-regular texture, a start lattice describing a local regular texture is determined per cluster in contrast to [5], where the lattice is expanded from one characteristic point and a pair of vectors. Next, each start lattice is expanded. As near-regular textures are not perfectly regular, bigger variations of the texel and the displacement vectors are allowed. Once the lattice is determined, the cells of the lattice that involve possible texels are projected onto quadrangles of equal shape to become robust against scale and rotation variations.

The start lattice is composed of spatially connected cells that describe possible texels at a local and almost regular texture. The higher the number of cells in a lattice, the higher the probability of extracting real texels from the texture, as repeating elements with almost similar vectors are found. Every characteristic point in a cluster with four neighbors is further considered to determine the start lattice. At local and almost regular textures these vectors can be sorted in two groups of almost anti-parallel vectors with similar vector lengths. A graph is considered containing the characteristic points in the cluster fulfilling these criteria as nodes. Furthermore, edges are set in the graph between nodes containing neighboring characteristic points. The start lattice is the biggest connected graph.

The expansion of the lattice found at a local and almost regular texture is considered now. In the following $v_n^{\mathrm{latt}}, 1 \leq n \leq N$, are the characteristic points in the lattice (nodes in the graph), $\left(v_i^{\mathrm{latt}}, v_j^{\mathrm{latt}}\right)$ an edge between the points v_i^{latt} and v_j^{latt}, and $\mathcal{V}_n^{\mathrm{latt}} = \left\{ v_k^{\mathrm{latt}} \middle| \exists \left(v_n^{\mathrm{latt}}, v_k^{\mathrm{latt}}\right) \right\}$ the set of all characteristic points connected by one edge to v_n^{latt} directly. For the lattice expansion each point in the lattice v_n^{latt} with less than four neighbors $\left(\left|\mathcal{V}_n^{\mathrm{latt}}\right| < 4\right)$ is considered until no new points are added to

the lattice. New neighbors are searched from the characteristic points in the cluster. As the texture is now expanded over a non-regular texture, neighbors searched now must not fulfill the criteria of Sec. 4.1 anymore. The only condition is that a new neighbor from v_n^{latt} must lie within an environment of the expected position that the neighbor point would have in case that the texture was regular. The expected position of the point is given from the vectors spanned by the neighbors of v_n^{latt} and their respective neighbors. Due to the fact that the texture is not regular, the vectors will be different depending on the considered neighbor point. Because of this, when the cells are searched, the neighbor points of a characteristic point will always depend on the point used to calculate them. Finally, texels $t_z(\mathbf{x})\,,\ 1 \leq z \leq Z$, are found at the closed cells of the lattice. The texels will appear as many times as they are extracted as closed cell from the characteristic points to give cells that are essential for the texture higher weights.

A projection onto quadrangles of equal shapes from the detected texels at the original texture and an adjustment of the brightness are necessary, as the detected texels may vary in grey-level values, size, and shear. To extract the texel of the regular texture from the near-regular one, all of the detected texels are projected onto squared texels $t_z^{\mathrm{Proj}}(\mathbf{x})\,,\ z = 1, \ldots, Z$, of predefined length l^{Proj}. We assume that all texels in the near-regular texture can be described by an affine transformation [12] of the texels in the regular texture. Each texel in the near-regular texture is described by its four points decribing the corners of the cell. The six degrees of freedom from the affine transformation are determined by least-squares [13] using the correspondences of the four corner points. The possible texels independent of luminance are obtained by the following relation:

$$t_z^{\mathrm{Poss}}(\mathbf{x}) = \frac{t_z^{\mathrm{Proj}}(\mathbf{x}) - \min_{\mathbf{x}}\left(t_z^{\mathrm{Proj}}(\mathbf{x})\right)}{\max_{\mathbf{x}}\left(t_z^{\mathrm{Proj}}(\mathbf{x})\right)}. \tag{2}$$

4.3 Noise suppression

In this section we propose three different methods to overcome errors at the extraction of the texel. All of the methods can be combined.

The occlusion of faulty texels that appear seldom is considered first. Eigentexels $e_z^{\mathrm{tex}}(\mathbf{x})$ with their eigenvalues λ_z are extracted from the pos-

sible texels $t_z^{\mathrm{Poss}}(\mathbf{x})$ via principal component analysis [14] together with the mean texel over all possible texels. Approximated possible texels $\hat{t}_z^{\mathrm{Poss}}(\mathbf{x})$ are constructed next from the eigentexels with the M' strongest eigenvalues and the mean texel. All of the texels $t_z^{\mathrm{Poss}}(\mathbf{x})$ that vary significantly from their reconstructed $\hat{t}_z^{\mathrm{Poss}}(\mathbf{x})$ are discarded.

A similarity comparison between the extracted texels is considered next. In real images, characteristic points may be grouped in the same cluster even if they belong to different texels. Furthermore, errors may occur under the assumption of an affine transformation of the texels. Because of this, we propose to group the proposed texels by similarity first and extract as many texels as similar groups found. The two dimensional correlation coefficient [14]

$$r\left(t_i^{\mathrm{Poss}}(\mathbf{x}), t_j^{\mathrm{Poss}}(\mathbf{x})\right) = \frac{\sum_{\mathbf{x}} \left(h_i(\mathbf{x})\, h_j(\mathbf{x})\right)}{\sqrt{\left(\sum_{\mathbf{x}} \left(h_i(\mathbf{x})\right)^2\right)\left(\sum_{\mathbf{x}} \left(h_j(\mathbf{x})\right)^2\right)}}, \tag{3}$$

is calculated between all possible texels, where $h_g(\mathbf{x}) = t_g^{\mathrm{Poss}}(\mathbf{x}) - \overline{t_g^{\mathrm{Poss}}(\mathbf{x})}$, $\overline{t_g^{\mathrm{Poss}}(\mathbf{x})}$ is the average component of $t_g^{\mathrm{Poss}}(\mathbf{x})$, and $g \in \{i, j\}$. As in Sec. 3.1, different clusters of texels are obtained from the connected graphs between the possible texels. The texels are connected with each other if their correlation coefficient is high. Only clusters containing a minimum number of texels are further considered.

The repeatability of the texels at the contours can also be selected to discard proposed texels as in [5]. Here we perform a comparison via the correlation coefficient of the left and right contour, respectively, upper and lower contour. Texels with almost homogenous grey-level contours are always allowed as possible texels.

5 Post-processing of the extracted regular textures

The extracted texels $t_s^{\mathrm{Exp}}(\mathbf{x})$, $1 \leq s \leq S$, may repeat texels from a regular texture or be faulty. One possibility to overcome faulty extracted texels is to proof the repeatability of the texels at the contours if this method was not used to suppress possible texels. Repeated extracted texels can be merged as in Sec. 5.1 and overdetermined extracted texels can be recognized following the method from Sec. 5.2.

5.1 Fusion of similar texels

There is not a unique texel in a regular texture, as cyclic shifts will deliver new texels. However, all of the texels that become equal after an adequate cyclic shift represent the same texture. We group all of these texels together to avoid redundant information. For this purpose, pairwise comparisons $c(t_i^{\mathrm{Exp}}(\mathbf{x}), t_j^{\mathrm{Exp}}(\mathbf{x}))$ are done between all extracted texels. The correlation coefficients are calculated between $t_i^{\mathrm{Exp}}(\mathbf{x})$ and $t_j^{\mathrm{Exp}}(\mathbf{x})$ cyclically shifted ($(i,j) \in S$):

$$c\left(t_i^{\mathrm{Exp}}(\mathbf{x}), t_j^{\mathrm{Exp}}(\mathbf{x})\right) = \max_{\forall \mathbf{y} \in \mathbb{R}^2} \left(r\left(t_i^{\mathrm{Exp}}(\mathbf{x}), t_j^{\mathrm{Exp}}(\mathbf{x}+\mathbf{y})\right)\right). \tag{4}$$

Texels having a high comparison coefficient are grouped together.

Objects with similar textures of different colors will be grouped as the textures will be similar at the grey-level texels. To overcome this problem, the color is taken into account. The colors of the textures are represented with a signature and compared. The texels are only grouped together if their colors and their colors' proportions are similar.

5.2 Suppresion of overdetermined texels

To suppress texels that contain the real texel several times (overdetermined texels), an image $b_s(\mathbf{x})$ per extracted texel is considered:

$$b_s(\mathbf{x}) = \sum_{k=1}^{l^{\mathrm{Rep}}} \sum_{d=1}^{l^{\mathrm{Rep}}} t_s^{\mathrm{Exp}}(\mathbf{x}) **\delta\left(\mathbf{x} - k \cdot l^{\mathrm{Proj}} \begin{pmatrix} 1 \\ 0 \end{pmatrix} - d \cdot l^{\mathrm{Proj}} \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right). \tag{5}$$

Next, the two-dimensional Fourier transform is calculated [15]:

$$B_s(\mathbf{f}) \propto T_s^{\mathrm{Exp}}(f) \sum_{k=1}^{l^{\mathrm{Proj}}} \sum_{d=1}^{l^{\mathrm{Proj}}} \delta(\mathbf{f} - k\mathbf{f}_1 - d\mathbf{f}_2), \tag{6}$$

where $\begin{pmatrix} \mathbf{f}_1^{\mathrm{T}} \\ \mathbf{f}_2^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} l^{\mathrm{Proj}} & 0 \\ 0 & l^{\mathrm{Proj}} \end{pmatrix}^{-1}$. Every extracted texel $t_s^{\mathrm{Exp}}(\mathrm{x})$ whose first peak does not appear at the expected positions is discarded, as well as texels whose magnitude of the first peaks at the expected positions varies considerably.

Table 1: Analyzed extraction methods.

		T1	T2	T3	T4	T5	T6	T7
Sorted characteristic points Sec. 3	Mean-shift					•		
	Proposed clustering (Sec. 3.1)	•	•	•	•		•	•
	Compact point clouds (Sec. 3.2)	•	•			•		•
	Eliminate contour points (Sec. 3.3)	•	•			•		•
Texel extraction Sec. 4	Neighbors with similar grey-level (Sec. 4.1)	•	•			•		•
	Neighbors independent of the grey-level			•	•		•	
	Proposed possible texels (Sec. 4.2)	•	•	•		•		•
	Start from one characteristic point				•		•	
	Occlusion of faulty texels (Sec. 4.3)	•	•			•		•
	Similarity of the texels (Sec. 4.3)	•	•			•		•
	Repeatability at the contours (Sec. 4.3)		•	•	•		•	
Regular texture post-processing Sec. 5	Fusion of similar texels (Sec. 5.1)	•	•	•	•	•		
	Color fusion of similar texels (Sec. 5.1)							•
	Overdetermined texels (Sec. 5.2)	•	•	•	•	•		•
	Repeatability at the contours (Sec. 4.3)	•		•	•	•		•

6 Results

Different methods and combinations are proposed in the previous sections. Seven different combinations are tested (T1,...,T7) that are shown in Table 1. T3, T4, and T6 do not require neighbor points at the start lattice to have a similar grey-level value (similar to [5]). Furthermore, in T4 and T6 the start lattice is composed from one characteristic point and its neighbors and not as many cells as possible (similar to [5]). On the other hand T1, T2, T5, and T7 require neighbor points at the start lattice to have a similar grey-level value. T1, T2, and T7 use the method proposed here to sort the characteristic points whereas T5 uses a mean-shift approach [16]. The consequences of suppressing texels depending on the repeatability of their contours as a post-processing step or at the texel extraction can be deduced by comparing T1 and T2. T7 considers the color information. The combinations are tested at a database containing 50 images of textured objects, clothes, and surfaces. Some of the objects and clothes are set in front of a patterned background.

Table 2: Evaluation Results.

	Detect	True	Repeat	Almost True	False
T1	52	50	3	2	0
T2	52	48	4	2	0
T3	52	44	2	2	2
T4	52	36	0	2	3
T5	52	44	3	1	2
T6	52	37	19	3	13
T7	57	55	3	2	0

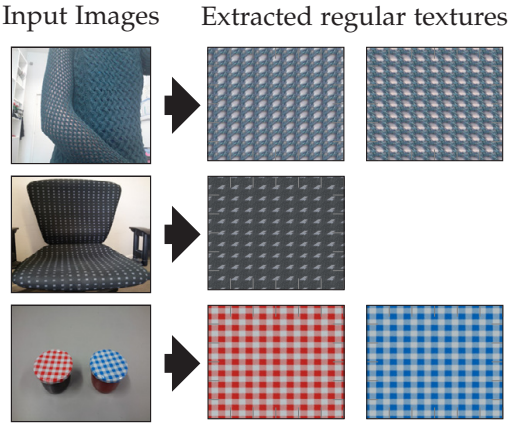


Figure 1: Example of a repeated detected texture (top), an almost right extracted texture (middle), and correctly extracted regular textures.

The results are shown in Table 2. *Detect* is the total number of textures that should be extracted as regular from the database, 52 at grey-level images and 57 at color images. *True* represents the number of textures that are correctly extracted (e. g., Fig. 1 bottom), *Repeat* the number of repeated textures (e. g., Fig. 1 top), *Almost True* the number of textures that are not perfectly extracted but close to the real ones (e. g., Fig. 1 middle), and *False* the number of wrongly extracted textures.

If our proposed method is used to cluster the characteristic points and the start lattice is obtained from neighbors with similar grey-level

values (T1, T2, T7), no wrong textures are extracted (in contrast to T3, T4, T5, T6). Furthermore, T1, T2, and T7 extract overall the highest number of correct textures (96.15 %, 92.31 %, 96.49 %) in contrast to T3, T4, T5, and T6 (84.62 %, 69.23 %, 84.62 %, 71.15 %). The fusion of the methods proposed to post-process the extracted regular textures (Sec. 5) decrease the number of repeated (0 from T4 vs. 19 from T6) and wrong (3 from T4 vs. 13 from T6) extracted textures considerably. Overall T7 delivers the best results for color images and T1 for grey-level images.

7 Conclusions

In this work the extraction of regular textures from surfaces in real images is analyzed. An approach that can be divided into three stages is used. For each stage different methods are proposed and implemented. Different combinations of the proposed methods per stage are tested at a database containing 50 images. Our proposed method to group repeated characteristic points together with the search of a start lattice at a local regular texture from the near-regular one containing as many texels as possible and the fusion of the methods proposed at the post-processing of the extracted texels show the best results with no false positives and up to 96.49 % correctly extracted regular textures.

References

1. E. R. Davies, *Machine vision: theory, algorithms, practicalities*, 3rd ed. Amsterdam [i. a.]: Morgan Kaufmann, 2005.
2. Y. Liu, W.-C. Lin, and J. Hays, "Near-regular texture analysis and manipulation," in *ACM Transactions on Graphics (TOG) – Proceedings of ACM SIGGRAPH 2004*, vol. 23, no. 3, 2004, pp. 368–376.
3. G. Schindler, P. Krishnamurthy, R. Lublinerman, Y. Liu, and F. Dellaert, "Detecting and matching repeated patterns for automatic geo-tagging in urban environments," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.
4. P. Hernández Mesa, J. Anastasiadis, and F. Puente León, "Identification and sorting of regular textures according to their similarity," in *Automated Visual Inspection and Machine Vision*, ser. Proceedings of SPIE, J. Beyerer and F. Puente León, Eds., vol. 9530. Bellingham, WA: SPIE, 2015.

5. A. Hilsmann, D. C. Schneider, and P. Eisert, "Warp-based near-regular texture analysis for image-based texture overlay." in *Vision, Modeling and Visualization Workshop*. The Eurographics Association, 2011.
6. J. Hays, M. Leordeanu, A. A. Efros, and Y. Liu, "Discovering texture regularity as a higher-order correspondence problem," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria*. Berlin Heidelberg: Springer, 2006, pp. 522–535.
7. J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.
8. M. Park, K. Brocklehurst, R. T. Collins, and Y. Liu, "Deformed lattice detection in real-world images using mean-shift belief propagation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1804–1816, 2009.
9. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
10. H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria*. Berlin Heidelberg: Springer, 2006, pp. 404–417.
11. M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996, pp. 226–231.
12. W. E. Snyder and H. Qi, *Machine vision*. Cambridge: Cambridge Univ. Press, 2004.
13. M. S. Grewal and A. P. Andrews, *Kalman filtering : theory and practice using MATLAB*, 2nd ed. New York [i. a.]: Wiley, 2001.
14. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. New York [i. a.]: Wiley, 2001.
15. J. Beyerer, F. Puente León, and C. Frese, *Machine Vision – Automated Visual Inspection: Theory, Practice and Applications*. Berlin Heidelberg: Springer, 2016.
16. D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

High-throughput sensor-based sorting via approximate computing

Georg Maier[1], Michael Bromberger[2], Thomas Längle[1] and Wolfgang Karl[2]

[1] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB),
Fraunhoferstr. 1, 76131 Karlsruhe, Germany
[2] Karlsruhe Institute of Technology (KIT), Institute of Computer Science & Engineering (ITEC),
Kaiserstr. 12, 76131 Karlsruhe, Germany

Abstract Sensor-based sorting provides solutions for separating cohesive, granular materials. In order to reliably locate the position of material objects with deviating velocity, perception and separation shall be close together. This in turn poses challenges on the data analysis systems, since available processing time depends on this distance and the velocity of the object. Whenever the sorting decision for an object cannot be derived in time, no information about this object is taken into account, potentially leading to a sorting error. In this paper, we present an analysis of the impact of this distance and an approach which allows utilizing information about an object before the final classification result is available. Therefore, we apply the concept of anytime algorithms to a decision tree-based classifier. First results suggest that the approach can indeed increase the sorting quality for complex objects for which the deadline would else not be met.

Keywords Sensor-based sorting, optical sorting, approximate computing.

1 Introduction

Sensor-based sorting is an established technology for sorting various products, for instance according to quality aspects [1]. Among others,

fields of application include food processing [2], recycling [3], and sorting of industrial minerals [4]. Systems for sorting cohesive, granular materials, so-called bulk goods, typically use scanning sensors, such as line-scan cameras. Hence, the material has to move. This provides several advantages in system design, for instance regarding necessary illumination. Nevertheless, corresponding systems typically suffer from a delay between perception of the material by one or more sensors and the physical separation thereof [5], see Figure 1. In order to achieve high sorting quality, this delay is to be minimized. This implies an as small as possible physical distance between perception and separation. However, the minimum delay and hence the distance depends on the performance of the included signal processing systems. In many cases, information retrieved from applied sensors can be represented as an image, turning the signal processing task into an image evaluation problem. Furthermore, compressed air jets are commonly used for separation [6,7].

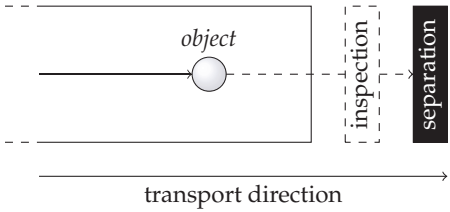


Figure 1: Illustration of the delay between perception and separation.

In order to keep the jet activation time window in case of rejecting an object as small as possible and hence minimize the amount of falsely co-deflected objects located nearby [8], the physical distance between sensors and separation mechanism has to be small. This poses challenges upon the image evaluation system because the system needs to respect a real-time criterion. However, the time required to process a perceived object may strongly vary dependent on its properties. The number of concurrent objects also varies. In times of high system load, caused by high material density or encounter of several computational-wise complicated objects, sorting quality may drop. This is due to missing the required deadline for certain objects and therefore not having

derived a sorting decision in time. When designing a system, the minimum manageable distance may be determined empirically. However, the resulting deadline may be too strict for outliers in terms of computational burden which might occur during final operation. Implementing data processing techniques in hardware has been proposed in order to achieve high throughput during evaluation [9–11]. Compared with software solutions, corresponding implementations typically lack flexibility towards the product to be sorted. This is disadvantageous whenever a system shall realize several sorting tasks and may also result in higher development cost. Hence, existing solutions focus on a certain application to improve performance by optimizations in software, exploiting hardware acceleration or setting the distance according to the average classification time. These approaches are time-consuming to apply and still offer optimization potential. Therefore, we target towards a system that itself adapts the used algorithms according to the deadline and the required execution time for an object. Since the data analysis works with noisy sensor data, which introduce uncertainties, approximation techniques are applicable with less impact on the accuracy of image processing tasks. This enables to apply methods from the recent research topic *approximate computing* [12] in the field of sensor-based sorting. Specifically, we use these methods to comply with real-time constraints instead of reducing the energy consumption which is typical in that domain. Anytime algorithms are a specific kind of approximated algorithms, which improve the accuracy of results over time.

In this paper, we present a study on the correlation between quality of physical separation, the distance between the perception and separation line and its influence on available processing time to derive a sorting decision. By that we show that minimizing the distance is crucial, especially when strong deviations in the individual objects' velocities exist. To tackle the problem of reduced processing time when shortening this distance, we further present an approach turning a decision tree-based classifier into an *anytime algorithm*. This method allows deriving an approximate sorting decision which can be calculated faster than conventional classification. More precisely, we enable to accelerate object classification from image data in high-load situations. This way, under heavy system load, sorting decisions based on incomplete knowledge about an object can be executed instead of missing the deadline and consequently not using any knowledge about an object at all.

Results show that utilizing uncertain classification results leads to better sorting performance when computation time is strongly limited. Considering two data-sets, we show that $\sim 50 - 70\%$ of the processing time of a conventional approach suffice in order to achieve equal classification performance.

2 Role of the distance between perception and physical separation

In the following, we discuss the impact of the distance between the perception and the separation line in sensor-based sorting. Firstly, it is shown in Section 2.1 that minimizing this distance is crucial for sorting quality, especially in case of imperfect flow control. In this case, individual objects may not come to rest until reaching the perception stage which typically is due to properties of the product, such as shape, as well as the transport mechanism. Secondly, Section 2.2 discusses the challenges caused by above optimization posed upon data analysis systems. Finally, we present the dilemma of requiring a close distance to ensure good physical separation but thereby tightening the firm real-time limits by means of this distance in Section 2.3.

2.1 Necessity for a close distance between perception and separation

Typically, for the purpose of separation, an array of air jets is aligned perpendicular to transport direction at a given distance. For rejecting an object, one or several of these jets are activated for a certain time window in order to deflect the object. Conventional systems assume a fixed average velocity for each object. Knowledge about transport velocity determines this velocity, for instance in case of a belt sorter the belt speed. The known distance between perception and separation as well as the average object velocity predicts the point in time when the object reaches the separation line. Besides the point in time, activating the correct air jet(s) also requires selection of the right jets in the array.

In case of imperfect flow control, the velocity of individual objects is not equal to the average velocity. Consequently, the system predicts the actual point in time as well as jets to be activated falsely. The quantity of

this error depends on the deviation between average and individual velocity as well as the distance between perception and separation. Whenever this deviation grows too big, the corresponding object is missed, potentially leading to a sorting error.

Calculating the jet activation window based on the area of the object, for instance the surrounding bounding box, presumably increases the deflection accuracy. Also, whenever purity of the *accept* product is the primary goal, a larger deflection window further increases chances to deflect an object. Yet it is important to note that enlarging the activation window also increases the risk of falsely co-deflecting other objects which are located nearby.

In summary, the following factors are relevant when determining whether an object is going to be successfully deflected:

1. The deviation between the objects velocity and the expected velocity.

2. The distance between inspection and separation line.

3. The resolution of the array of air jets.

Hence, it becomes obvious that minimizing the distance between perception and separation increases sorting quality. However, this only holds true as long as the time available for data analysis, which directly depends on this distance, suffices.

2.2 Firm real-time requirements depending on the distance between perception and separation

Whenever the sorting decision is derived after the object already passed the separation stage, the result of data analysis is worthless, since the system cannot execute the sorting decision on the right time. Consequently, the utility function of deriving the sorting decision corresponds to firm real-time requirements, see Figure 2. At any time before the deadline passed, the sorting decision can be executed and hence the utility is at its maximum. After this point in time, it drops to zero.

Additionally, the data analysis typically has to analyze numerous objects simultaneously. Hence, several objects are competing over processing time. The time required to derive a sorting decision for an object

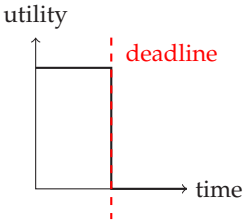


Figure 2: Utility function for deriving of the sorting decision. The deadline is not met whenever the object to be potentially deflected has physically passed the separation line.

typically varies in dependence to its properties. For instance, the calculation of features may require more time for bigger objects than for smaller ones in terms of the area. To further support this statement, the time elapsed between perception, represented by an image time stamp, and having calculated the sorting decision and being ready to activate the corresponding air jets was determined experimentally for numerous objects. For this purpose, we recorded image sequences from a sorting system containing mainly wheat seeds. This data was fed into a corresponding evaluation system using a camera simulator and hence under stable, repeatable conditions. Results are depicted in Figure 3. As can be seen, while most objects require $4 - 15$ ms, outliers exist which require significantly more time from perception to the sorting decision.

2.3 Correlation of distance between perception and separation and real-time requirements

Figure 4 illustrates the dependency between the physical distance, here denoted in meters, from perception to separation and available processing time (shown as colors from red to blue). For this simulation, an expected velocity of 3 m/s was assumed. This simplified model further assumes that an object is represented by a single point. For an arbitrary object, deviations from the expected velocity ranging from -1 to 1 m/s were respected for both direction components (Δv_x, Δv_y) on the transportation plane. With regard to the range of an air jet, we assume 10 mm and no expansion of the activation window. Based on the deviation of

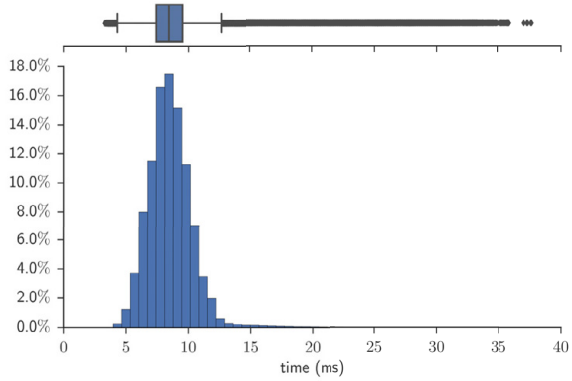


Figure 3: Time passing between perception of an object and deriving the sorting decision. Data was collected for $\sim$ 15 million objects.

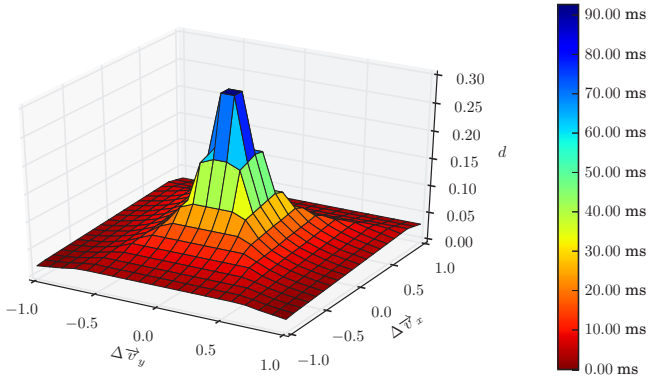


Figure 4: Illustration of the dependency between flow control, deflection success rate and time available for deriving the sorting decision.

the object's velocity to the expected velocity and the distance between perception and separation, we can determine whether the object is going to be successfully deflected or not.

As can be seen, the successful deflection rate, which is illustrated in terms of the area covered by the surface, dramatically decreases for

higher deviations in velocity already at a distance of 5 to 10 cm. If more than 40 ms of processing time is required, almost no deviation in velocity can be accounted for because the expected point in time and location of the object at the separation stage differs too much from the actual.

3 Approximate sorting decisions to increase efficiency in sensor-based sorting

Data analysis in sensor-based sorting typically contains several processing steps. For instance, pre-processing of the data in terms of filtering might be necessary due to noisy input data. Also, regions containing objects need to be extracted from the input image. For these regions, the system performs classifications based on calculated features. Classification results provide the basis for the sorting decisions.

While certain processing steps require the same execution time independent of the input data, for other steps, e. g., feature calculation and classification, the required time heavily depends on the data representing an individual object. To tackle this challenge, we present an interruptible classifier based on a decision tree which allows deriving a sorting decision before the final classification decision. A noteworthy strength of this approach is that the system only applies approximation when the situation requires it, e. g., in times of exceptional high system load.

3.1 Approximation of classification with decision trees

Decision trees can perform classifications in the field of sensor-based sorting. Each node of such a tree evaluates a feature against a certain threshold. Utilizing lazy evaluation, features are calculated on demand in order to minimize required execution time. A leaf of the tree represents a certain class.

While sorting can often be regarded as a two class problem, i. e., accept or reject, systems often differentiate between more classes. This allows gathering important information, for instance with respect to the nature of the material feed. Assuming a contaminated material with known foreign objects, this gives insights regarding the distribution of individual, foreign materials in the bulk. Pragmatically, this allows fa-

cility operators to derive conclusions regarding their suppliers, i. e., the quality of the material.

The decision tree for a sorting task can be learned using a labeled training set, for instance by the well-known ID3 algorithm [13]. Here, starting at the root, the learning algorithm selects a split point optimal to some criteria, for instance information gain, for each node. In our case, this split point includes a feature as well as a threshold. The algorithm then splits the training data accordingly and continues the procedure for the left and right sub-tree in a recursive manner.

3.2 Additional information from the training phase

During learning of the tree, we save additional information regarding the distribution of the training data in each node. More precisely, each node is annotated by the class ratios that were used during learning to determine the split feature and value in this node. As will be discussed later, this is crucial information when stopping the classification procedure before it terminates.

Furthermore, we introduce a second training phase for which we use a separate training set in order to simulate classification using the already learned decision tree. For every object contained in the set, we check in each node it passes during this simulated classification whether the class yielding the highest ratio of the node corresponds to the true class of the object. This way, a *node hit ratio* is calculated. This key figure quantifies how often the correct class of an object was already determined in the node.

3.3 Extending classification by order planning

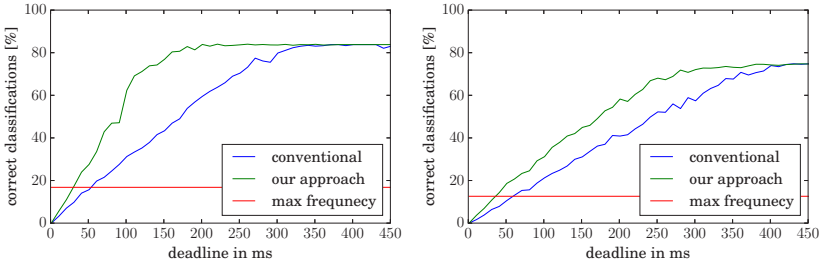
During runtime, we divide the classification procedure into single steps, whereas each step corresponds to an object passing through a node of the decision tree. We extend classification of several objects by an order planning phase which is executed after each step. The purpose of the planning mechanism is to determine which object is to perform the next step. The planning is realized by means of a priority queue. The priority is formulated in terms of minimizing the node hit ratio whereas each object is assigned the node hit ratio of the node it is currently located in. The goal of this approach is to grant additional computing

steps to those object, for which it is likely that their correct class is not yet known, while terminating classification early for those objects, for which chances are good that it already is.

3.4 Test methodology and experimental results

We evaluate our approach on two different sorting problems. The first data-set consists of six different classes of glass shards which mainly differ in their color. The second data-set consists of eight classes of lego bricks which mainly differ in geometric properties.

For the evaluation, we randomly select 50 objects from a testing set. As a reference, we perform conventional, non-interruptible classification with the same decision tree. For all objects not classified within a deadline, we consider the classification result to be not correct. We do not consider the additional time required for order planning. Results for different deadlines are illustrated in Figure 5. All experiments run on the same hardware setup.



(a) Glas shards (b) Lego bricks

Figure 5: Correct classification ratios of an interruptible classifier for different deadlines.

As can be seen from Figure 5(a), the proposed approach achieves significantly more correct classifications compared to the conventional classification for all simulated deadlines on the glass shards data-set. It also can be seen that when $\sim$ 350 ms are available the classification of each of the 50 objects can be completed by the conventional approach and hence the maximum correct classification rate possible with the

learned decision tree is reached. The approach presented in this paper reaches this rate already at around ~ 200 ms.

It also holds that the proposed approach performs superior to the conventional classification for the Lego data-set, see Figure 5(b). Here, ~ 400 ms are necessary to achieve the maximum correct classification rate reachable by the learned decision tree. However, our approach requires around the same time. Yet, for closer deadlines, better correct classification rates can be achieved.

4 Conclusion

In this paper, we illustrated the significance of the distance between the inspection and physical separation line in sensor-based sorting. It was shown that minimizing this distance is crucial in order to achieve reliable separation whenever flow control is imperfect. However, it has also been demonstrated that minimizing the distance poses high challenges upon data analysis. To tackle this problem, an approach was presented which allows to derive uncertain classification results which serve as the basis of the sorting decision. The presented results demonstrated that realizing a decision tree-based classifier as an anytime algorithm enables significantly speeding up the classification process while keeping classification errors low. In the context of sensor-based sorting, this also means that incomplete information about an object can be utilized to derive the sorting decision instead of missing the deadline and hence not respecting any of the object's properties.

With respect to future work, we intent to extend the approach in order to also work with diverse deadlines. While experiments presented in this paper assumed a group of objects all with the same deadline, the planning can be improved by incorporating the individual deadlines.

References

1. H. Wotruba, "Stand der Technik der sensorgestützten Sortierung," *BHM Berg-und Hüttenmännische Monatshefte*, vol. 153, no. 6, pp. 221–224, 2008.
2. V. Narendra and K. Hareesha, "Prospects of computer vision automated grading and sorting systems in agricultural and food products for quality

evaluation," *International Journal of Computer Applications*, vol. 1, no. 4, pp. 1–9, 2010.

3. M. Bigum, L. Brogaard, and T. H. Christensen, "Metal recovery from high-grade WEEE: a life cycle assessment," *Journal of hazardous materials*, vol. 207, pp. 8–14, 2012.
4. J. Lessard, J. de Bakker, and L. McHugh, "Development of ore sorting and its impact on mineral processing economics," *Minerals Engineering*, vol. 65, pp. 88–97, 2014.
5. F. Pfaff, C. Pieper, G. Maier, B. Noack, H. Kruggel-Emden, R. Gruna, U. D. Hanebeck, S. Wirtz, V. Scherer, T. Längle *et al.*, "Improving optical sorting of bulk materials using sophisticated motion models," *tm–Technisches Messen*, vol. 83, no. 2, pp. 77–84, 2016.
6. J. Huang, T. Pretz, and Z. Bian, "Intelligent solid waste processing using optical sensor based sorting technology," in *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol. 4, Oct 2010, pp. 1657–1661.
7. T. Pearson, "High-speed sorting of grains by color and surface texture," *Applied engineering in agriculture*, vol. 26, no. 3, pp. 499–505, 2010.
8. R. Pascoe, O. Udoudo, and H. Glass, "Efficiency of automated sorter performance based on particle proximity information," *Minerals Engineering*, vol. 23, no. 10, pp. 806–812, 2010.
9. Z. Ruoyu, K. Za, and J. Yinlan, "Design of Tomato Color Sorting Multichannel Real-Time Data Acquisition and Processing System Based on FPGA," in *Information Science and Engineering (ISISE), 2010 International Symposium on*. IEEE, 2010, pp. 207–212.
10. B. W. House, D. W. Capson, and D. C. Schuurman, "Towards real-time sorting of recyclable goods using support vector machines," in *Sustainable Systems and Technology (ISSST), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 1–6.
11. M. A. Nuño-Maganda, Y. Hernandez-Mier, C. Torres-Huitzil, and J. Jimenez-Arteaga, "FPGA-based real-time citrus classification system," in *Circuits and Systems (LASCAS), 2014 IEEE 5th Latin American Symposium on*. IEEE, 2014, pp. 1–4.
12. C. Plessl, M. Platzner, and P. J. Schreier, "Approximate Computing," *Informatik-Spektrum*, vol. 38, no. 5, pp. 396–399, 2015.
13. J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

Multimodal convolutional neural networks for road detection

Stefan Held[2], Hassene Khelil[1,2] and Matthias Killat[2]

[1] Technische Universität München,
Arcisstraße 21, 80333 München
[2] ITK-Engineering AG,
Lochhamer Str. 15, 82152 Planegg

Abstract In this work we propose convolutional networks for the task of road detection. This task is handled as a semantic segmentation, i. e., we predict a label at each input image pixel. First we consider a fully convolutional architecture that relies on proven classification networks, which we adapt for semantic segmentation. We train the network end-to-end based on road scene images from the KITTI benchmark. Motivated by the fact that scale invariant features such as road boundaries are important for road detection, we investigate a multiscale convolutional network that we apply on a Haar-wavelet image representation. Considering the different characteristics of our proposed approaches, we fuse features learned by single and multiscale architecture into a multimodal network. This network achieves 87.78 % F1-score on the KITTI benchmark at an inference time of 0.3 seconds per image (QuadroK620 3.3 GHz, Xeon E3-1226 v3).

Keywords Pattern recognition, mathematical models, environmental perception for vehicles.

1 Introduction

Analysing and understanding visual data is currently one of the major topics in automotive engineering. Especially driver assistance systems and autonomous driving applications require a clear understanding of the vehicle environment including the road area and road users. Visual information about this environment is then necessary for building

intelligent systems (e. g., lane keeping, collision avoidance and road following). Depending on the application, a road scene can be segmented into many types of objects like road, pedestrians and cars. In this work we focus on road detection, since it represents the basic step for a robust lane detection especially in the case of unmarked roads.

1.1 Related work

Road detection has received considerable attention since the mid 1980s. According to a recent survey on the progress in road and lane detection [1], road detection systems are mainly based on vision, other modalities (LIDAR [2], RADAR [3]), or a fusion of many sensors. Road detection also gained attention in the area of computer vision, especially due to the presence of visual cues and landmarks in the road scene. In our work, we focus on using only visual data from a monocular camera, which enables the application of our algorithm on a wide spectrum of vehicles. The dataset applied is the KITTI Benchmark [4], which is a widely used dataset. We approach the task as a binary semantic segmentation of a road scene, where each pixel is labelled "road" or "background".

Semantic segmentation: The research in this area is very active and has witnessed a rapid progress, especially due to the huge success of deep learning in object classification problems. Based on this success, researchers started investigating deep convolutional neural networks for more structured predictions such as the semantic segmentation problem. One of the most recent methods is the SegNet network [5], which presents the state of the art in various datasets such as Pascal VOC12 [6]. The authors presented in [5] the most successful approaches for solving the semantic segmentation problem.

Road detection: In this section we highlight the state of the art of vision-based road detection using convolutional neural networks. A comprehensive review of all existing road detection methods is out of the scope of this paper. Brust et al. [7] proposed a convolutional neural network, which learns image patches for labeling one pixel centred at each patch. To enforce the contextual informa-

tion for each pixel, they incorporated prior information from pixel positions and achieved F1-score of $86.5\,\%$ on the KITTI dataset. Due to the patch-based labeling, this method yields a slow inference time of $30\,s$. Fetaya et al. [8] reduced the segmentation task to a column-wise regression, which they solve with convolutional neural networks by finding the road limit in each column. They achieved $89\,\%$ F1-score on the KITTI Benchmark at an inference time of $1\,s$. Mohan et al. [9] combined convolutional with deconvolutional neural networks. They broke down the segmentation problem by using image patches instead of full images. Looping their network over each patch, they predicted only the pixel labels of that specific patch. The approach has the advantage of being completely automated without the need of additional post-processing. It yields the state of the art performance of the KITTI dataset with $93.65\,\%$ F1-score at a runtime of $2\,s$.

1.2 Convolutional Neural Networks for semantic segmentation

Convolutional Neural Networks (CNNs) [10] belong to the family of methods referred to as deep learning and can be applied to various problems of computer vision and digital image processing. These networks were re-interpreted in 1998 by (LeCun et al.) [11] for recognizing digits in a document recognition application. The main idea of CNNs is to transform the input into a higher dimension by generating layers of feature maps, which can be used to solve a certain task (e. g., semantic segmentation).

We consider $x^{(i)}$ an input color (or grayscale) image from a defined set of images $X = \{x^{(1)}, ..., x^{(i)}, ..., x^{(n)}\}$. By assuming the existence of a function $\mathcal{G}$ mapping an input image to its ground truth labels, we can solve our segmentation problem by training a CNN model $\mathcal{M}_w$ with parameters w so that the error between the network predictions $\mathcal{M}_w(x^{(i)})$ and the ground truth $\mathcal{G}(x^{(i)})$ is minimal:

$$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} \mathcal{L}(\mathcal{M}_w(x^{(i)}), \mathcal{G}(x^{(i)})) \tag{1}$$

$\hat{w}$ represents the optimal network parameters for the dataset X and $\mathcal{L}$ is the error function between $\mathcal{M}_w$ and $\mathcal{G}$.

2 Single scale network

2.1 Architecture

Our first approach is a single scale network similar to the model proposed by Long et al. [12] (see Figure 1). The network uses images in full resolution instead of operating on extracted image patches [9], [7]. Each convolutional layer applies learnable filters on its input and generates multiple feature maps. An activation function is then applied on these maps (e. g., Rectified Linear Units [10], which forwards only positive activations). This non-linearity is then followed by a max-pooling layer that selects the highest activation in a 2×2 window at stride 2 resulting in downsampled feature maps. This is important to increase translation and distortion invariance of the feature maps. By processing the input through a hierarchy of these layers, we get a set of low resolution feature maps, which are learned in a second stage with two convolutional layers that preserve their spatial sizes. The final output is then upsampled back to the original input size, so that each pixel has a prediction score for each class. The kernel size of the convolutional filters controls the field of view of the network in the input image. This field of view determines the contextual window centred at each pixel and therefore influences its labelling. By experimenting with different fields of view by varying the kernel sizes, we could achieve a strong yet compact architecture (see Table 1).

Type of Layer	kernel size	stride	pad	Output size
conv1	3×3	1	1	$192 \times 640 \times 12$
relu1	-	-	-	$192 \times 640 \times 12$
pool1	2×2	2	1	$96 \times 320 \times 12$
conv2	3×3	1	1	$96 \times 320 \times 24$
relu2	-	-	-	$96 \times 320 \times 24$
pool2	2×2	2	1	$48 \times 160 \times 24$
conv3	3×3	1	1	$48 \times 160 \times 48$
relu3	-	-	-	$48 \times 160 \times 48$
pool3	2×2	2	1	$48 \times 160 \times 48$
conv4	3×3	1	1	$24 \times 80 \times 48$
relu4	-	-	-	$24 \times 80 \times 48$
pool4	2×2	2	1	$24 \times 80 \times 48$
conv5	7×7	1	3	$12 \times 40 \times 48$
relu5	-	-	-	$12 \times 40 \times 48$
conv6	3×3	1	1	$12 \times 40 \times 48$
relu6	-	-	-	$12 \times 40 \times 48$
conv7	1×1	0	0	$12 \times 40 \times 2$
upsample	–	–	–	$192 \times 640 \times 2$
softmax	–	–	–	$192 \times 640 \times 2$

Table 1: Single scale network architecture.

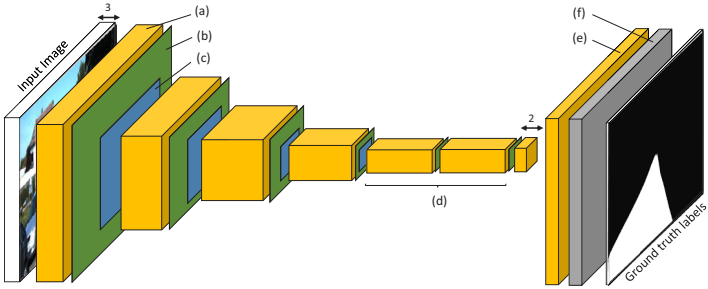


Figure 1: Single scale architecture: (a) convolution feature maps; (b) non-linearity; (c) max pooling; (d) fully convolutional layers; (e) upsampled feature maps; (f) softmax layer.

2.2 Training procedure

For training, we apply the logarithmic loss function that calculates the training error in each iteration. This error is then back-propagated through the network layers, where the network parameters are correspondingly updated. By considering $x \in \mathbb{R}^{H \times W}$, $y \in \mathbb{R}^{H \times W}$ and $\tilde{y} \in \mathbb{R}^{H \times W}$ an input image, its prediction labels and its ground truth labels respectively, our logarithmic loss function is defined as:

$$\mathcal{L}(y, \tilde{y}) = -\sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{k=1}^{2} \mathbb{1}_{\{k=\tilde{y}_{ij}\}} log(c_{ijk}) \tag{2}$$

with c_{ijk} the score value for the label y_{ij} corresponding to the class k. We train our network with 200 annotated images available in the KITTI dataset. Due to this quite small dataset we perform data augmentation by flipping the training images along the y-axis. As preprocessing, we apply a data normalization and zero-centering for each color channel separately. The training of the single scale network as well as the following networks in this paper is done using stochastic gradient descent with momentum.

2.3 Results

We test our network on $n_{test} = 60$ annotated images of the KITTI dataset. As an evaluation metric we use the error of the maximum F1-score [4]:

$$maxF = \frac{1}{n_{test}} \cdot \sum_{i=1}^{n_{test}} \textit{F1-score}(y^{(i)}(\hat{\alpha}), \tilde{y}^{(i)}) \tag{3}$$

with $\hat{\alpha}$ the optimal threshold in $[0, 1]$. Our single scale architecture yields a maximum F1-score of $92.7\,\%$ with 400 RGB training images. As shown in Figure 2 the single scale network qualitative results reveal certain limitations. The first major problem is the coarse prediction caused by the low resolution feature maps. This restricts the ability to detect fine details especially far ahead in the road image. Another issue is caused by regions of similar color and structure, which are misinterpreted by the network as a road region. Some road boundaries are also not perfectly detected by the network and lead to an over-segmentation in the boundary region. This is mainly due to the shadows that mask road boundaries. Shadows also limit the detection performance especially by drawing high intensity edges in the road area.

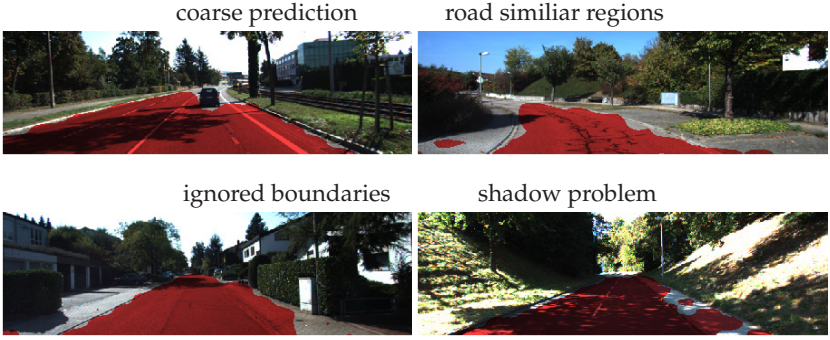


Figure 2: Qualitative results of the single scale network.

3 Multiscale network

3.1 Architecture

The analysis of the single scale network demonstrates that road boundaries are important visual cues for road detection, especially since most missclassifications occur in the boundary regions. Another great challenge for the single scale structure is the problem of coarse predictions. We address these two main problems by implementing a multiscale network (see Figure 3). This architecture exploits different resolutions of road images, which enables the learning of scale invariant features and enhances therefore the road boundaries. Based on the work of Farabet et al. [13], our multiscale network extracts multiscale features in a first stage, then feed-forwards these features through fully convolutional layers similar to the single scale network. By using this new architecture we need $30\,\%$ less parameters and achieve a finer prediction due to the smaller number of pooling layers. Given the kernel sizes in Table 2, the network has a field of view of 46×46 at each scale. This covers $1.7\,\%$, $6.8\,\%$ and $27\,\%$ of the first, second and third scale respectively, which represents a sufficient contextual window for each pixel.

Type of Layer	kernel size	stride	pad	Output size
conv1	7×7	1	3	$192 \times 640 \times 16$
tanh	$--$			$192 \times 640 \times 16$
pool1	2×2	2		$96 \times 320 \times 16$
conv2	7×7	1	3	$96 \times 320 \times 16$
tanh	$--$			$96 \times 320 \times 16$
pool2	2×2	2		$48 \times 160 \times 16$
concat	$--$			$48 \times 160 \times 48$
conv3	7×7	1	3	$12 \times 40 \times 48$
relu	$--$			$12 \times 40 \times 48$
conv4	3×3	1	1	$12 \times 40 \times 48$
relu	$--$			$12 \times 40 \times 48$
conv5	1×1			$12 \times 40 \times 2$
upsample	$--$			$192 \times 640 \times 2$
softmax	$--$			$192 \times 640 \times 2$

Table 2: Multiscale network architecture.

3.2 Training procedure

We define f_s the feature extractor network of the first stage with parameters w_s. Applied on images x_s at N scales, f_s generates the feature maps y_s:

$$y_s = f_s(w_s; x_s), \quad \forall s \in \{1, ..., N\} \tag{4}$$

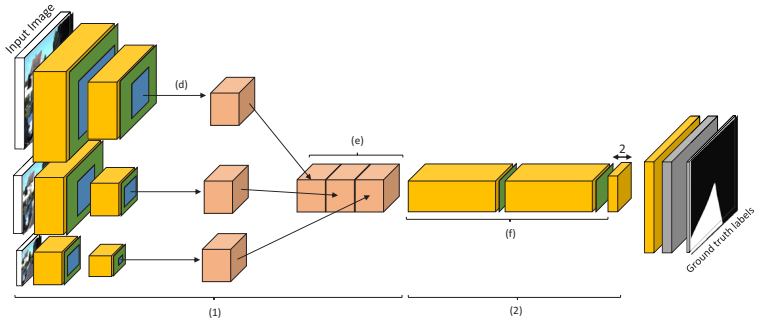


Figure 3: multiscale architecture: (1) feature extraction (2) classification. (a) upsampling the feature maps; (e) concatenated feature maps; (f) fully convolutional layers.

For learning scale invariant features, the same network parameters w_s are shared across the scales $w_s := w$, $\forall s \in \{1, ..., N\}$. Our experiments have shown that these perform better than using different parameters for each scale [13]. The raw image input is transformed in 3 scales x_1, x_2 and x_3. The coarser-scale feature maps y_2 and y_3 are upsampled to match the finest scale maps y_1. All maps are then concatenated and forward-propagated through the classification network of the second stage f_c, which has a fully convolutional structure similar to the single scale network.

3.3 Multiscale image representation

Since we aim to enhance the edge information, which helps in detecting the road boundaries, we use Haar-wavelets as multiscale decomposition of the image. Input images are decomposed into 4 sub-images (a horizontal, a vertical and a diagonal detail and an approximation image). This process is then repeated for many scales resulting in a multiscale wavelet representation. We applied this decomposition on each RGB channel separately and trained the network only with the multiscale details, which contain the useful edge information.

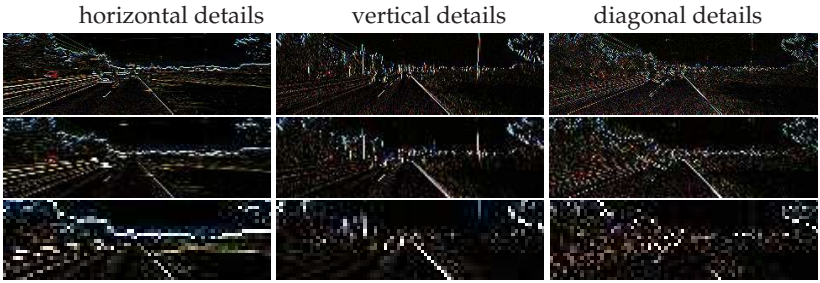


Figure 4: 2D-Haar wavelets details at 3 resolutions.

3.4 Results

From the qualitative results of the multiscale network we draw two interpretations: The first interpretation is that the scale invariant features help with finding finer details and detecting road regions far ahead in the image. This is better demonstrated in the bird eye view in Figure 5. The second interpretation is the sensitivity to high intensity edges like road markings and hard shadows (see Figure 6).

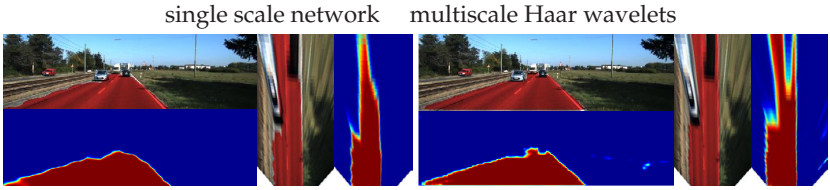


Figure 5: Comparison of the results between the single scale and multiscale networks in perspective view and bird eye view.

4 Multimodal network

4.1 Architecture

The results of the previous architectures show the advantages and drawbacks of each approach. The single scale network detects most of the road region successfully with coarser prediction and less sensi-

tivity to edges, while the multiscale network detects finer structures but comes with a high edge sensitivity. To combine the strengths of both approaches we introduce a multimodal network. As shown in Figure 7 the network is composed of two stages (i. e., a two-stream feature extractor and a fusion stream).

Figure 6: Qualitative results of multiscale network with Haar-wavelets.

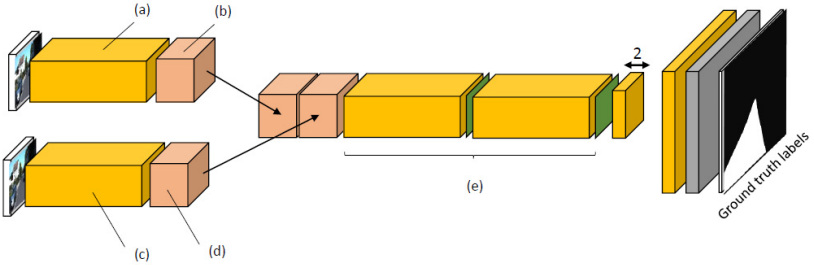


Figure 7: Multimodal network architecture: (a)-(b) single scale network and its feature maps; (c)-(d) multiscale network and its feature maps; (e) fusion network.

4.2 Training procedure

Given an image x, we generate the inputs x_1 and x_2 for the streams f_1 and f_2 by applying corresponding preprocessing and multiscale transformation. We fix the weights w_1 and w_2 for each stream and randomly initialize the weights w_3 of the second stage fusion stream. The feature maps y_1 and y_2 are then generated with the trained streams of the first stage, upsampled to the finest maps scale and concatenated. The result-

ing maps are then forward-propagated through the second stage where the weights w_3 are learned.

4.3 Results

The results in Figure 8 show a performance increase of the $maxF$ from $93.13\,\%$ with the multiscale network with input Haar-wavelets to $93.74\,\%$ with the multimodal network. The qualitative results in Figure 9 also demonstrate how the misclassifications due to the high edge sensitivity are reduced, by keeping the multiscale network fine prediction.

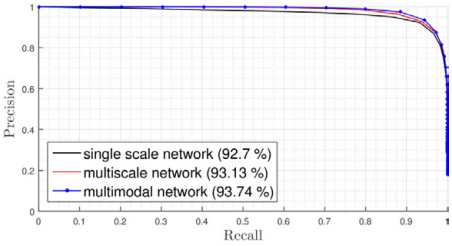


Figure 8: Results comparison of the single scale, multiscale and multimodal network.

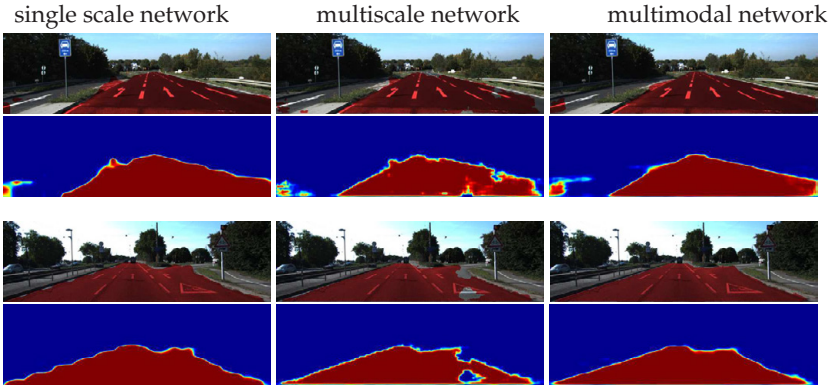


Figure 9: Visualization of the prediciton results of the multimodal network in comparison to the single scale and multiscale network.

5 Conclusions

This paper demonstrates how well-adapted architectures of convolutional neural networks trained end-to-end can achieve competitive de-

tection results for the task of road detection. It also presents guidelines on convolutional network architectural choices, which rely on filter sizes, number of convolutional layers and an adequate field of view. Two major insights were gained in this paper. Firstly, the extensive analysis of qualitative results of our single scale network shows the importance of road boundaries as visual cues for road detection. This is the motivation to use a multiscale architecture fed with enhanced boundary information, e. g., Haar-wavelets. By doing this, we are able to address the coarseness of the prediction of the single scale network and detect the road far ahead in the image. The latter is especially apparent when transforming our results into the bird-eye-view. Secondly, we demonstrate the advantages of combining different network features in a second learning stage. The multimodal network that enables this combination leads to performance improvement in comparison to the single and multiscale networks trained separately. We quantified the multimodal network on 3 road categories of the KITTI benchmark [4] and achieved $87.78\,\%$ F1-score at an inference time of 0.3 seconds (QuadroK620 3.3 GHz, Xeon E3-1226 v3) . This approach suggests the interpretation that the finest scale of raw-images represents a useful aid to Haar-wavelet features that rely mainly on boundary information. This is due to the fact that the raw images can be interpreted as the approximation of the Haar-wavelet decomposition, which can not be used directly in the training of the multiscale network.

References

1. A. B. Hillel, R. Lerner, D. Levi, and G. Raz, "Recent progress in road and lane detection: a survey," *Machine Vision and Applications*, vol. 25, no. 3, pp. 727–745, 2014.
2. R. Fernandes, C. Premebida, P. Peixoto, D. Wolf, and U. Nunes, "Road detection using high resolution lidar," in *Vehicle Power and Propulsion Conference (VPPC), 2014 IEEE*. IEEE, 2014, pp. 1–6.
3. C. Adam, R. Schubert, N. Mattern, and G. Wanielik, "Probabilistic road estimation and lane association using radar detections," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, July 2011, pp. 1–8.

4. J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.
5. V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.
6. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.
7. C.-A. Brust, S. Sickert, M. Simon, E. Rodner, and J. Denzler, "Convolutional patch networks with spatial prior for road detection and urban scene understanding," *arXiv preprint arXiv:1502.06344*, 2015.
8. D. Levi, N. Garnett, and E. Fetaya, "Stixelnet: A deep convolutional network for obstacle detection and road segmentation." in *26TH British Machine Vision Conference (BMVC)*, 2015.
9. R. Mohan, "Deep deconvolutional networks for scene parsing," *arXiv preprint arXiv:1411.4101*, 2014.
10. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
11. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
12. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CVPR (to appear)*, Nov. 2015.
13. C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, 2013.

Projektive Invarianten höherdimensionaler Punktkonfigurationen

Projective invariants of higher dimensional point configurations

Bastian Erdnüß

Karlsruher Institut für Technologie,
Institut für Photogrammetrie und Fernerkundung,
Englerstraße 7, 76131 Karlsruhe
und
Fraunhofer Institut für Optronik, Systemtechnik und Bildauswertung,
Fraunhoferstraße 1, 76131 Karlsruhe

Zusammenfassung Invarianten in der Geometrie helfen, unwesentliche Aspekte eines Problems beiseite zu lassen, um so das Wesentliche in den Vordergrund zu rücken. Die fundamentale Invariante der projektiven Geometrie ist das Doppelverhältnis, eine Invariante von vier kollinearen Punkten. Sie ist für das Verständnis und die anschauliche Vorstellung der projektiven Ebene von zentraler Bedeutung. Alle anderen Invarianten der projektiven Geometrie lassen sich auf das Doppelverhältnis zurückführen, allerdings wertet dies andere Invarianten nicht ab, denn auch sie tragen zum Verständnis der projektiven Geometrie bei. In diesem Artikel wird eine naheliegende Verallgemeinerung des Doppelverhältnisses auf Punktkonfigurationen in höherdimensionalen projektiven Räumen erarbeitet. Die wesentliche Einsicht wird dabei sein, dass das Verhältnis baryzentrischer Koordinaten projektiv invariant ist.

Schlagwörter Projektive Geometrie, Geometrische Invarianten, Doppelverhältnis, Baryzentrische Koordinaten, Projektive Invarianz, Baryzentrisches Verhältnis, Tripelverhältnis.

Abstract Geometric invariants help to focus on the main aspects of a geometric problem by hiding irrelevant information. The fundamental invariant of projective geometry is the cross-ratio, an invariant of four collinear points. It is from particular importance for the understanding and conception of the projective plane. All other projective invariants can be expressed in terms of cross-ratios but this does not depricate other invariants as they are still beneficial for the understanding of the projective geometry. This article provides a natural generalisation of the cross-ratio to point configurations in higher dimensional projective spaces. The central insight will be that the ratio of barycentric coordinates is projectively invariant.

Keywords Projective geometry, geometric invariance, cross-ratio, barycentric coordinates, projective invariance.

1 Einleitung

Invarianten sind Größen, die sich unter einer bestimmten Klasse von Transformationen nicht ändern. Z. B. ist der Abstand zweier Punkte invariant unter euklidischen Transformationen, oder der Winkel zweier sich schneidender Geraden ist invariant unter Ähnlichkeitstransformationen. Eine Schwierigkeit in der Invariantentheorie[1] entsteht dadurch, dass mit den Invarianten $a, b, c, \dots$ auch jede Funktion f von den Invarianten wieder eine neue Invariante $f(a, b, c, \dots)$ ergibt, die jedoch abhängig von den zugrundeliegenden Invarianten $a, b, c, \dots$ ist. Die Invariantentheorie befasst sich daher vor allem mit der Frage, wie viele wesentlich verschiedene Invarianten es zu einer gegebenen Situation gibt und ob ein vorgegebenes System von Invarianten unabhängig ist. Jedoch kann sie ein Invariantensystem nicht gegenüber einem äquivalenten System als *angepasster*, *treffender*, oder *zweckmäßiger* auszeichnen. In diesem Text wird ein System von Invarianten für den projektiven Raum angeben. Das System ist dabei in dem Sinne ausgezeichnet, das es dazu eine einfache und anschauliche Interpretation gibt und es so das Verständnis und weitere Nachdenken über die entsprechenden Punktkonfigurationen fördern kann.

[1] vgl. z. B. Mumford [1]

1.1 Einordnung

Die hier beschriebenen Invarianten sind zwar nicht gänzlich unbekannt, jedoch nach Wissen des Autors zumindest in dieser Darstellung unveröffentlicht. In Mundy [2] tauchen zwei Invarianten I_1 und I_2 für die projektive Ebene auf, die zu den hier erarbeiteten Invarianten äquivalent sind und sich einfach ineinander umrechnen lassen. Pamfilos [3] beschreibt die meisten hier genutzten Werkzeuge, jedoch ohne sie in dieser Form zusammenzusetzen. Quan [4] verwendet die hier erarbeiteten Invarianten bereits in derselben Form, ohne sie allerdings als ausgezeichnet zu fassen und gibt die notwendigen Formeln nur in algorithmischer Form an. Dieser Artikel ist hauptsächlich durch das Paper von Quan [4] motiviert und baut auf die in Richter-Gebert [5] verwendete Methodik und Notation auf.

Eine enge Analogie existiert zum affinen Raum. Dort ist die grundliegende Invariante das Teilverhältnis auf Geraden. Das Teilverhältnis verallgemeinert sich auf räumliche Strukturen zu baryzentrischen Koordinaten, die ebenfalls affin invariant sind. Sie bekommen im Zusammenspiel eine sinnvolle Bedeutung und können so als ein einzelnes eigenständiges Objekt angesehen werden, eine *Rauminvariante*. Gleiches gibt es im projektiven Raum. Das Doppelverhältnis ist als Verhältnis zweier Teilverhältnisse projektiv invariant. Das Verhältnis der baryzentrischen Koordinaten zweier Punkte ist ebenfalls projektiv invariant und liefert die hier betrachtete projektive Rauminvariante.

1.2 Vorgehen

Zu $n+3$ Punkten im n-dimensionalen projektiven Raum existiert eine Rauminvariante. Der Leitgedanke des Artikels ist, aus $n+2$ der Punkte ein Koordinatensystem festzulegen und die Koordinaten des letzten Punktes darin anzugeben. Diese sind dann nach Konstruktion projektiv invariant.

Um aus den ersten $n+2$ Punkten ein Koordinatensystem zu berechnen und den letzten Punkt in diesem darzustellen, müssen lineare Gleichungssysteme gelöst werden. Dies erfolgt mit der Cramer'schen Regel. In dem hier vorliegenden Fall hat die Cramer'sche Regel einen engen Bezug zu baryzentrischen Koordinaten, der auch die schlussendliche Interpretation der gefundenen Invarianten liefert.

2 Grundlagen

Im diesem Kapitel werden zunächst die notwendigen Werkzeuge aus der projektiven Geometrie und der linearen Algebra eingeführt, wie der projektive Raum, projektive Transformationen, baryzentrische Koordinaten und die Cramer'sche Regel. Die Einführung richtet sich an Leser, denen die Mittel zumindest prinzipiell bekannt sind, und dient in erster Linie der Einführung der hier verwendeten Notation.

2.1 Der projektive Raum

Der n-dimensionale projektive Raum $\mathbb{P}^n$ ist eine Erweiterung des reellen Raums $\mathbb{R}^n$ um *Punkte im Unendlichen*. Modelliert wird er als Menge der Ursprungsgeraden im $\mathbb{R}^{n+1}$. Jeder von 0 verschiedene Punkt $x \in \mathbb{R}^{n+1}$ wird mit der Ursprungsgerade $\langle x \rangle := \{\lambda x : \lambda \in \mathbb{R}\} \in \mathbb{P}^n$ identifiziert und repräsentiert so einen Punkt im $\mathbb{P}^n$. x wird auch als *homogene Koordinaten* des projektiven Punktes $\langle x \rangle$ bezeichnet. Zwei homogene Koordinaten $x, y \in \mathbb{R}^{n+1}$ mit $x = \lambda y$ für ein $\lambda \in \mathbb{R}$ repräsentieren denselben projektiven Punkt. Dafür wird $x \sim y$ geschrieben.

Häufig wird in der Notation nicht zwischen den homogenen Koordinaten $x \in \mathbb{R}^{n+1}$ und dem projektiven Punkt $\langle x \rangle \in \mathbb{P}^n$ unterschieden und beide werden mit x bezeichnet.

2.2 Einbettung des reellen Raums

Der reelle Raum $\mathbb{R}^n$ wird in eine affine Hyperebene H des $\mathbb{R}^{n+1}$ eingebettet, die nicht durch den Ursprung geht. H wird als *Parameterraum* des $\mathbb{P}^n$ bezeichnet. Jeder Punkte $x \in H$ wird mit dem Punkt $\langle x \rangle \in \mathbb{P}^n$ identifiziert. Die zu H parallel liegenden Geraden $\langle y \rangle \in \mathbb{P}^n$ bilden die *Punkte im Unendlichen*. Der Parameterraum und die Einbettung

$$H_0 := \{(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n+1} : x_{n+1} = 1\} \tag{1}$$

$$h_0 : \mathbb{R}^n \to H_0, (x_1, \dots, x_n) \mapsto (x_1, \dots, x_n, 1) \tag{2}$$

$$(x_1, \dots, x_n, x_{n+1}) \cong \left(\frac{x_1}{x_{n+1}}, \dots, \frac{x_n}{x_{n+1}} \right) \qquad \text{für } x_{n+1} \neq 0 \tag{3}$$

sind in der Computer Vision verbreitet und finden z. B. in Hartley [6] Anwendung. (3) identifiziert einen Teil des $\mathbb{P}^n$ mit dem $\mathbb{R}^n$. Punkte im Unendlichen haben keine Entsprechung als Punkte des $\mathbb{R}^n$.

2.3 Allgemeine Lage

Höchstens $n+1$ projektive Punkte des $\mathbb{P}^n$ sind *in allgemeiner Lage*, wenn sie für irgendwelche homogenen Koordinaten linear unabhängig im $\mathbb{R}^{n+1}$ sind. Mehr als $n+1$ projektive Punkte des $\mathbb{P}^n$ sind in allgemeiner Lage, wenn jeweils beliebige $n+1$ davon in allgemeiner Lage sind. Zu $n+1$ homogenen Koordinaten im $\mathbb{R}^{n+1}$ in allgemeiner Lage gibt es genau eine Parameterebene H, die sie enthält.

2.4 Projektive Transformationen

Jede invertierbare lineare Abbildung A auf dem $\mathbb{R}^{n+1}$ induziert eine projektive Transformation auf dem $\mathbb{P}^n$ via $A\langle x\rangle := \langle Ax\rangle$. Aus der Linearität folgt die Wohldefiniertheit der Abbildung. Zwei projektive Transformationen A und B sind genau dann gleich, wenn A als Matrix ein Vielfaches von B ist. Das wird $A \sim B$ notiert.

Projektive Transformationen transformieren Punkte in allgemeiner Lage auf Punkte in allgemeiner Lage. Eine projektive Transformation ist durch die Angabe von $n+2$ Punktepaaren in allgemeiner Lage eindeutig festgelegt. Präziser, sind $x_1, \ldots, x_{n+2} \in \mathbb{P}^n$ und $y_1, \ldots, y_{n+2} \in \mathbb{P}^n$ jeweils in allgemeiner Lage, so gibt es genau eine projektive Transformation A mit $Ax_k \sim y_k$ für $k = 1, \ldots, n+2$.

2.5 Projektive Koordinatensysteme

Ein projektives Koordinatensystem im $\mathbb{P}^n$ besteht aus $n+2$ Punkten des $\mathbb{P}^n$ in allgemeiner Lage. $n+1$ der $n+2$ Punkte fungieren als Koordinatenachsen im $\mathbb{R}^{n+1}$ und der letzte Punkt setzt die Maßstäbe auf den $n+1$ Achsen fest. Pamfilos [3] spricht hier vom *Kalibrieren* des Koordinatensystems. Zu zwei Koordinatensystemen gibt es genau eine projektive Transformation, die das eine in das andere überführt.

Das *Standardkoordinatensystem* besteht aus den $n+1$ Einheitsvektoren e_k des $\mathbb{R}^{n+1}$ und der Standardkalibrierung $e = (1, \ldots, 1)$. Es wird als $(e_1, \ldots, e_{n+1}; e)$ notiert. Unter der Einbettung h_0 aus Abschnitt 2.2 entsprechen e_1 bis e_n den Punkten im Unendlichen, die zu den n Koordinatenachsen des R^n gehören, und e_{n+1} und e entspricht dem Koordinatenursprung und dem im Einheitshyperwürfel davon diagonal gegenüber liegenden Punkt.

2.6 Matrizen, Vektoren und Determinanten

Matrizen $A = (a_1, \dots, a_n)$ werden als Auflistung ihrer Spaltenvektoren geschrieben. Determinanten werden wie in Richter-Gebert [5] mit eckigen Klammern notiert: $[a_1, \dots, a_n] := \det(A)$. Alle Vektoren in diesem Text sind als Spaltenvektoren zu verstehen, auch wenn sie im Fließtext oder aus Platzgründen liegend notiert sind. Zeilenvektoren im Sinn von $1 \times n$-Matrizen kommen in diese Text nicht vor.

2.7 Baryzentrische Koordinaten

Baryzentrische Koordinaten werden in diesem Text nur für Parameterebenen definiert und gebraucht. Ist H eine Parameterebene des $\mathbb{P}^n$ und sind $x_1, \dots, x_{n+1} \in H$ in allgemeiner Lage, so existieren zu jedem $y \in H$ eindeutige Koeffizienten $\lambda = (\lambda_1, \dots, \lambda_{n+1}) \in \mathbb{R}^{n+1}$, so dass mit $X = (x_1, \dots, x_{n+1})$

$$y = \lambda_1 x_1 + \dots + \lambda_{n+1} x_{n+1} = X\lambda \tag{4}$$

gilt. $\lambda_1, \dots, \lambda_{n+1}$ sind die *baryzentrischen Koordinaten von y bzgl. $x_1, \dots, x_{n+1}$*. Sie sind die eindeutige Lösung λ des linearen Gleichungssystems $X\lambda = y$. Die baryzentrischen Koordinaten sind nicht projektiv invariant, d. h. sie hängen von der konkreten Wahl der Repräsentanten von $x_1, \dots, x_{n+1}$ ab.

2.8 Cramer'sche Regel

Ist $X = (x_1, \dots, x_{n+1})$ invertierbar und $\lambda, y \in \mathbb{R}^{n+1}$, so hat das lineare Gleichungssystem $X\lambda = y$ die eindeutige Lösung

$$\lambda = \frac{1}{\det(X)} \begin{pmatrix} [y, x_2, \dots, x_{n+1}] \\ [x_1, y, \dots, x_{n+1}] \\ \vdots \\ [x_1, x_2, \dots, y] \end{pmatrix} \sim \begin{pmatrix} [y, x_2, \dots, x_{n+1}] \\ [x_1, y, \dots, x_{n+1}] \\ \vdots \\ [x_1, x_2, \dots, y] \end{pmatrix}. \tag{5}$$

y ersetzt dabei auf der rechten Seite der Reihe nach alle Spalten von X einmal in den Determinanten.

Liegt y in derselben Parameterebene wie $x_1, \dots, x_{n+1}$, so liefert die Cramer'sche Regel insbesondere eine Formel zur Berechnung der baryzentrischen Koordinaten von y bzgl. $x_1, \dots, x_{n+1}$.

3 Hauptteil

Mit dem letzten Kapitel sind die Grundlagen für die Betrachtung des eigentlichen Problem gegeben. Zunächst wird die projektive Invariante definiert und benannt, um die es in diesem Text geht. Im Anschluss daran wird die Herleitung durchgeführt und begründet, dass es sich bei der definierten Größe tatsächlich um eine projektive Invariante handelt. Schließlich wird der Zusammenhang mit bekannten Invarianten aufgezeigt. Im Anschlusskapitel folgt dann noch eine kleine Anwendung.

3.1 Rauminvariante der projektiven Geometrie

Zu $n+3$ projektiven Punkten $x_1, \ldots, x_{n+1}, s, y \in \mathbb{P}^n$ in allgemeiner Lage wird die folgende projektive Invariante $z \in \mathbb{P}^n$ definiert:

$$z = (x_1, \ldots, x_{n+1}; s, y) := \begin{pmatrix} \frac{[y, x_2, \ldots, x_{n+1}]}{[s, x_2, \ldots, x_{n+1}]} \\ \frac{[x_1, y, \ldots, x_{n+1}]}{[x_1, s, \ldots, x_{n+1}]} \\ \vdots \\ \frac{[x_1, x_2, \ldots, y]}{[x_1, x_2, \ldots, s]} \end{pmatrix} \tag{6}$$

Dabei stehen $x_1, \ldots, x_{n+1}, s, y$ und z sowohl für Punkte im $\mathbb{P}^n$ als auch gleichzeitig für beliebige Repräsentanten im $\mathbb{R}^{n+1}$. Die Formel liefert einen Repräsentanten für z im $\mathbb{R}^{n+1}$. Das Ergebnis der Formel ist wohldefiniert, d. h. der projektive Punkt $\langle z \rangle$ hängt nicht von der Wahl der Repräsentanten von $x_1, \ldots, x_{n+1}, s$ und y ab.

Liegen s und y in derselben Parameterebene wie $x_1, \ldots, x_{n+1}$, stehen abgesehen von einem gemeinsamen Skalierungsfaktor in den Zählern von z die baryzentrischen Koordinaten von y und in den Nennern von z die baryzentrischen Koordinaten von s, jeweils bzgl. $x_1, \ldots, x_{n+1}$. Bei z handelt es sich also um das Verhältnis der baryzentrischen Koordinaten von y zu s. Daher scheint *baryzentrisches Verhältnis* ein angemessener Name für (6) zu sein.

Die Schreibweise $(\cdots ; \cdots)$ für das baryzentrische Verhältnis, bei der das Semikolon die Basis der baryzentrischen Koordinaten von den beiden Punkten absetzt, deren Koordinaten ins Verhältnis zueinander gesetzt werden, ist der häufigen Schreibweise $(a, b; c, d)$ für das Doppelverhältnis nachempfunden, die dieselbe Interpretation zulässt.

3.2 Spezialfall Flächeninvariante

Besonders wichtig ist das baryzentrische Verhältnis als Flächeninvariante der projektiven Ebene. Da sie dort das gemeinsame Verhältnis von drei Verhältnissen ist, scheint *Tripelverhältnis* ein angemessener Name dafür zu sein.

$$(x_1, x_2, x_3; s, y) = \left(\frac{[y, x_2, x_3]}{[s, x_2, x_3]}, \frac{[x_1, y, x_3]}{[x_1, s, x_3]}, \frac{[x_1, x_2, y]}{[x_1, x_2, s]} \right) \tag{7}$$

Die Formel (7) zeigt bereits deutlich das Bildungsmuster, das auch für den allgemeinen Fall (6) anzuwenden ist. Der geneigte Leser kann sich diese als Essenz des Artikels im Kopf behalten. Die Schreibweise mit den Determinanten zeigt deutlich die Linearität jeder einzelnen Komponente im Zähler und im Nenner, die in theoretischen Überlegungen und für Umformungen verwendet werden kann.

3.3 Herleitung

Zur Herleitung wird die eindeutige projektive Transformation T gesucht, die das Koordinatensystem $(x_1, \dots, x_{n+1}; s)$ in das Standardkoordinatensystem $(e_1, \dots, e_{n+1}; e)$ aus Abschnitt 2.5 überführt, und $z = Ty$ berechnet. z enthält somit die Koordinaten von y bzgl. des Koordinatensystems $(x_1, \dots, x_{n+1}; s)$ (die Koordinaten von z verstehen sich dabei bzgl. des Standardkoordinatensystems).

Die Herleitung geschieht in zwei Schritten. Zunächst wird die umgekehrte Transformationsmatrix A berechnet, die das Standardkoordinatensystem in $(x_1, \dots, x_{n+1}; s)$ überführt. Dieser Schritt wird als *Vorwärtstransformation* bezeichnet, da damit Standardkoordinaten in das Koordinatensystem $(x_1, \dots, x_{n+1}; s)$ übergeführt werden können. Im nächsten Schritt werden diejenigen Standardkoordinaten z gesucht, die unter A auf y abgebildet werden. Dazu wird das Gleichungssystem $Az \sim y$ gelöst. Der zweite Schritt wird als *Rückwärtstransformation* bezeichnet, da er den Punkt y zurück in das Standardkoordinatensystem transformiert.

Damit $(x_1, \dots, x_{n+1}; s)$ ein projektives Koordinatensystem bildet, müssen sich $x_1, \dots, x_{n+1}$ und s in allgemeiner Lage befinden. Ist das nicht der Fall, würde in der Formel (6) in manchen Nennern durch 0 geteilt werden. Auf diese Einschränkung kann daher nicht verzichtet werden. Allerdings könnte y prinzipiell frei gewählt werden.

Vorwärtstransformation

Zunächst wird die Transformation A von der Standardbasis $(e_1, \ldots, e_{n+1}; e)$ auf eine neue Basis $(x_1, \ldots, x_{n+1}; s)$ betrachtet[2]. Da A den k. Einheitsvektor e_k auf ein Vielfaches von x_k abbildet, hat A die Form

$$A \sim (\lambda_1 x_1, \ \ldots, \ \lambda_{n+1} x_{n+1}) \tag{8}$$

mit bis auf Skalierung zu bestimmendem $\lambda = (\lambda_1, \ldots, \lambda_{n+1}) \in \mathbb{R}^{n+1}$. Da $e = (1, \ldots, 1)$ auf ein Vielfaches von s abgebildet werden soll, gilt $Ae \sim \lambda_1 x_1 + \cdots + \lambda_{n+1} x_{n+1} = X\lambda \sim s$ mit $X = (x_1, \ldots, x_{n+1})$. Nach der Cramer'schen Regel (5) ist

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{n+1} \end{pmatrix} \sim \begin{pmatrix} [s, x_2, \ldots, x_{n+1}] \\ [x_1, s, \ldots, x_{n+1}] \\ \vdots \\ [x_1, x_2, \ldots, s] \end{pmatrix} \tag{9}$$

und damit folgt eingesetzt in (8) für die Vorwärtstransformation

$$A \sim \left([s, x_2, \ldots, x_{n+1}] x_1, \ldots, [x_1, x_2, \ldots, s] x_{n+1} \right) \ . \tag{10}$$

Rückwärtstransformation

A, λ und X sind wie gehabt. Für die Koordinaten z von y bzgl. des Koordinatensystems $(x_1, \ldots, x_{n+1}; s)$ gilt $Az \sim y$. Mit Formel (8) und $w = (w_1, \ldots, w_{n+1}) \in \mathbb{R}^{n+1}$ mit $w_k = z_k \lambda_k$ für $k = 1, \ldots, n+1$ ist

$$Az \sim \sum_{k=1}^{n+1} z_k \lambda_k x_k = Xw \sim y \ . \tag{11}$$

Das Gleichungssystem $Xw \sim y$ wird mit der Cramer'schen Regel (5) nach w gelöst und z mit $z_k = w_k / \lambda_k$ berechnet. Es ergibt sich

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_{n+1} \end{pmatrix} \sim \begin{pmatrix} [y, x_2, \ldots, x_{n+1}] \\ [x_1, y, \ldots, x_{n+1}] \\ \vdots \\ [x_1, x_2, \ldots, y] \end{pmatrix}, \quad \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_{n+1} \end{pmatrix} \sim \begin{pmatrix} \frac{[y, x_2, \ldots, x_{n+1}]}{[s, x_2, \ldots, x_{n+1}]} \\ \frac{[x_1, y, \ldots, x_{n+1}]}{[x_1, s, \ldots, x_{n+1}]} \\ \vdots \\ \frac{[x_1, x_2, \ldots, y]}{[x_1, x_2, \ldots, s]} \end{pmatrix} . \tag{12}$$

[2] vgl. zum Ansatz auch Richter-Gebert [5, S. 23]

3.4 Projektive Invarianz

Das baryzentrische Verhältnis (6) ist projektiv invariant, wenn $z \sim w$ für jede projektive Transformation A mit $z = (x_1, \dots, x_{n+1}; s, y)$ und $w = (Ax_1, \dots, Ax_{n+1}; As, Ay)$ gilt. Nach Konstruktion sind $z = Ty$ und $w = SAy$ für die eindeutigen projektiven Transformationen T und S, die $(x_1, \dots, x_{n+1}; s)$ bzw. $(Ax_1, \dots, Ax_{n+1}; As)$ in das Standardkoordinatensystem $(e_1, \dots, e_{n+1}; e)$ überführen. T und SA stimmen auf den $n+2$ Punkten $x_1, \dots, x_{n+1}, s$ in allgemeiner Lage überein. Nach Abschnitt 2.4 sind sie daher bereits gleich und es gilt $z = Ty \sim SAy = w$.

3.5 Zusammenhang mit bekannten Invarianten

Im Folgenden wird der reelle Raum $\mathbb{R}^n$ mittels h_0 (siehe Abschnitt 2.2, Gleichung (2)) in den $\mathbb{P}^n$ eingebettet und die Identifikation (3) genutzt.

Die projektive Gerade $\mathbb{P}^1$

Das baryzentrische Verhältnis (6) der vier reellen Zahlen $a, b, c, d \in \mathbb{R}$

$$(a, b; c, d) \cong \left(\frac{\begin{bmatrix} d & b \\ 1 & 1 \end{bmatrix}}{\begin{bmatrix} c & b \\ 1 & 1 \end{bmatrix}}, \frac{\begin{bmatrix} a & d \\ 1 & 1 \end{bmatrix}}{\begin{bmatrix} a & c \\ 1 & 1 \end{bmatrix}} \right) \cong \frac{(d-b)(a-c)}{(c-b)(a-d)} \tag{13}$$

ist das übliche Doppelverhältnis und gibt die Koordinaten des Punktes d im projektiven Koordinatensystem $(a, b; c)$ an.

Die projektive Ebene $\mathbb{P}^2$

Das Tripelverhältnis (7) von fünf Punkten $p_1, \dots, p_5 \in H_0$ ist

$$\begin{aligned} (p_1, p_2, p_3; p_4, p_5) &= \left(\frac{[p_5, p_2, p_3]}{[p_4, p_2, p_3]}, \frac{[p_1, p_5, p_3]}{[p_1, p_4, p_3]}, \frac{[p_1, p_2, p_5]}{[p_1, p_2, p_4]} \right) \\ &= \left(\frac{|m_{523}|}{|m_{423}|}, \frac{|m_{153}|}{|m_{143}|}, \frac{|m_{125}|}{|m_{124}|} \right) \cong \left(I_2, \frac{1}{I_1} \right) \end{aligned} \tag{14}$$

wobei $|m_{ijk}|$ und I_1, I_2 die Notation aus Mundy [2, S. 16f] ist und I_1 und I_2 die dort angegeben Invarianten für die Punktkonfiguration $p_1, \dots, p_5$ sind.

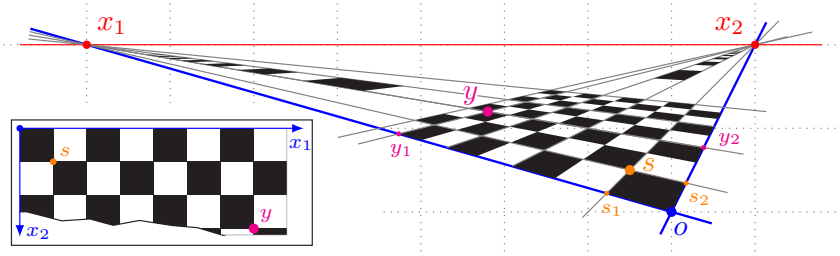


Abbildung 1: Projektiv verzerrtes Schachbrett mit Ursprung o und Fluchtpunkten x_1 und x_2. Maßstab ist durch s gegeben, y ein beliebiger Punkt auf dem Schachbrett. s_1, s_2 und y_1, y_2 sind die Projektionen von s und y auf die Koordinatenachse ox_1 und ox_2. Links unten wird das entzerrte Schachbrett gezeigt. Die entzerrten Koordinaten von y sind $(7, 3)$.

4 Anwendung

Eine Anwendung des Doppelverhältnisses ist das Messen in projektiv verzerrten Bildern. $(x, o; s, y)$ ist der metrische Abstand zwischen o und y auf der Geraden mit Fluchtpunkt x und Normierung 1 zwischen o und s. So lassen sich die metrischen Abstände[3] $(x_1, o; s_1, y_1) = 7$ für oy_1 und $(x_2, o; s_2, y_2) = 3$ für oy_2 in Abbildung 1 ermitteln.

In der Ebene können die metrischen Koordinaten eines Punktes mit dem Tripelverhältnis (7) bestimmt werden: $(x_1, x_2, o; s, y)$ sind unter der Einbettung h_0 die metrischen Koordinaten des Punktes y, von o aus gemessen entlang der beiden Achsen mit Fluchtpunkten x_1 und x_2, wobei s die o gegenüberliegende Ecke des Einheitsquadrats ist. In Abbildung 1 führt das auf die metrischen Koordinaten[4] $(x_1, x_2, o; s, y) \cong (7, 3)$ von y bzgl. des Koordinatensystems $(x_1, x_2, o; s)$.

Über die Projektionen auf die Koordinatenachsen kann das Tripelverhältnis auf das Doppelverhältnis zurückgeführt werden. Das Tripelverhältnis kann allerdings direkt mit y arbeiten, ohne umprojizieren zu müssen.

---

[3] In der Abbildung 1 sind die im angedeuteten Raster gemessenen Abstände im Bild $ox_1 = \sqrt{53}, os_1 = \frac{1}{9}\sqrt{53}, oy_1 = \frac{7}{15}\sqrt{53}$ und $ox_2 = \sqrt{5}, os_2 = \frac{5}{29}\sqrt{5}, oy_2 = \frac{5}{13}\sqrt{5}$.

[4] In Abbildung 1 sind die Koordinaten der Punkte im Bild im angedeuteten Raster $x_1(0, 0), x_2(8, 0), o(7, 2), s(6.5, 1.5)$ und $y(4.8, 0.8)$.

5 Zusammenfassung

In diesem Artikel wurde das baryzentrische Verhältnis (6) als projektive Invariante von $n+3$ Punkte aus dem $\mathbb{P}^n$ identifiziert. Sie verallgemeinert das Doppelverhältnis zu einer Rauminvariante in derselben Weise, wie in der affinen Geometrie baryzentrische Koordinaten das Teilverhältnis zu einer Rauminvariante verallgemeinern.

Da die Invariante so nahe liegt, ist es unwahrscheinlich, dass sie komplett unbekannt ist. Jedoch konnte der Autor sie nirgendwo explizit beschrieben finden. Darüber hinaus deuten die Texte, in denen sie implizit oder in anderer Form verwendet werden, darauf hin, das es methodisch wertvoll ist, sie als explizites Objekt zu fassen. Aus diesem Grunde wurde ihr eine knappe Schreibweise $(x_1, \dots, x_{n+1}; s, y)$ und ein Name (*baryzentrisches Verhältnis*, bzw. *Tripelverhältnis*) gegeben.

Auf weitere Anwendungen konnte in diesem Artikel aus Platzgründen leider nicht mehr eingegangen werden, dazu sei auf Mundy [7] und Quan [4] verwiesen. Einige davon können mit dem baryzentrischen Verhältnis klarer dargestellt werden.

Literatur

1. D. Mumford, J. Fogarty und F. Kirwan, *Geometric Invariant Theory*, 3. Aufl. Berlin Heidelberg: Springer-Verlag, 1994.
2. J. L. Mundy und A. Zisserman, „Introduction – towards a new framework for vision“, in *Geometric Invariance in Computer Vision*, J. L. Mundy und A. Zisserman, Hrsg. Cambridge, MA: MIT Press, 1992, S. 1–39.
3. P. Pamfilos, „Geometrikon.“ [Online]. Available: http://users.math.uoc.gr/~pamfilos/eGallery/Gallery.html
4. L. Quan, „Invariants of six points and projective reconstruction from three uncalibrated images“, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, Nr. 1, S. 34–46, 1995.
5. J. Richter-Gebert und T. Orendt, *Geometriekalküle*. Berlin Heidelberg: Springer-Verlag, 2009.
6. R. I. Hartley und A. Zisserman, *Multiple View Geometry in Computer Vision*, 2. Aufl. Cambridge: Cambridge University Press, 2004.
7. J. L. Mundy und A. Zisserman, Hrsg., *Geometric Invariance in Computer Vision*. Cambridge, MA: MIT Press, 1992.

Automatisierte Qualitätsbeurteilung von (S)VHS-Digitalisaten

Automated quality assessment of digitized (S)VHS-tapes

Stefanie Müller, Stefan Kahl und Maximilian Eibl

Technische Universität Chemnitz,
Professur Medieninformatik,
Straße der Nationen 62, 09111 Chemnitz

Zusammenfassung In den frühen 1990er Jahren waren es vor allem (S)VHS-Kassetten, die kleinen lokalen Fernsehstationen zur Archivierung von Roh- und Sendematerial gedient haben. In den letzten Jahren haben immer mehr dieser TV-Sender begonnen, ihre Videodaten zu digitalisieren. In diesem Beitrag stellen wir ein Verfahren zur automatisierten Bewertung der Bildqualität für digitalisierte (S)VHS-Videoaufnahmen vor. Wir stützen uns dabei auf ein Convolutional Neural Network auf Basis einer modifizierten Variante der bewährten VGG-16 Architektur. Zum Training nutzen wir ca. 10.000 nutzergenerierte Bewertungen von extrahierten Keyframes.

Schlagwörter VHS, Digitalisierung, CNN, Deep Learning.

Abstract In the early 1990s German local TV stations used (S)VHS tapes to store raw material as well as broadcasted productions. In recent years, some TV stations began to successively digitize these tapes. We present an approach of an automated quality assessment in digitized (S)VHS footage. We propose a convolutional neural network based on a modified version of the well-known VGG-16 architecture which is trained on user-generated ratings of approx. 10.000 keyframes.

Keywords VHS, digitizing, CNN, deep learning.

1 Einleitung

Nachdem sich nach dem Zweiten Weltkrieg zunächst der öffentlich-rechtliche Rundfunk in Westdeutschland entwickelte, etablierte sich ab 1984 auch der private Rundfunk. Nach der Wiedervereinigung im Jahr 1990 entstanden vor allem im Osten Deutschlands zahlreiche lokale Fernsehsender. Aus heutiger Sicht dokumentierte deren Arbeit auf unterschiedliche Weise die vielfältigen Herausforderungen, welche die Menschen aus Ostdeutschland während des Umbruchs von einem sozialistischen zu einem marktwirtschaftlichen System meistern mussten. Die meist sehr kleinen Fernsehsender nutzten für die Produktion und Lagerung aus wirtschaftlichen Gründen hauptsächlich VHS- und SVHS-Magnetbänder. Diese wurden seither ohne Archivstandards gelagert: Oftmals war es den lokal ansässigen Fernsehsendern nicht möglich, ein professionelles Archiv mit langlebiger Technik, optimalen klimatischen Bedingungen und entsprechend ausreichender Dokumentation oder Indizierung zu betreiben. So wurden Bänder nicht katalogisiert und ohne Inhaltsangabe in einfachen Kartons gelagert. Videoarchivdaten nach der Digitalisierung für die Einbindung in heutige Produktionen manuell zu durchsuchen und entsprechend aufzubereiten, ist ein zeitaufwändiger Prozess, den lokale TV-Sender nicht bewältigen können.

Den Ausgangspunkt fur unsere Untersuchungen stellt ein Pilotprojekt zur hochwertigen Digitalisierung sächsischer Lokalfernseharchive dar [1]. Die Professur Medieninformatik hat es sich zur Aufgabe gemacht, Teile dieser Archivbestände (aktuell 183 (S)VHS-Kassetten und 106 Beta-Kassetten) zu digitalisieren, zu annotieren und für eine Weiterverarbeitung aufzubereiten. Dazu zählen u. a. bisherige Arbeiten zur automatisierten Gesichts-, Schnittgrenzen- und Spracherkennung [2], laufende Forschungen zur Detektion von Bildstörungen sowie die qualitative Einordnung und Verbesserung des Archivmaterials [3]. Die in diesem Beitrag vorgestellte Methodik liefert eine Aussage über die Qualität des vorliegenden Materials, verringert durch eine automatisierte Untersuchung des Archivbestandes den Arbeitsaufwand des Produzenten und vereinfacht somit den Weiterverarbeitungsprozess erheblich.

2 Stand der Technik

In den vergangenen Jahren wurden zahlreiche Studien mit unterschiedlichem Fokus zur qualitativen Beurteilung von Videoinhalten durchgeführt. Dabei wurden Erkenntnisse zur objektiven Qualitätseinschätzung von Videos vorgestellt [4] und subjektive Methodiken untereinander verglichen [5].

Optimalerweise sollte bei Qualitätsstudien zu jedem Bild eine Reihe menschlicher Bewertungen vorliegen. Dies ist jedoch mit einem erhöhten Zeit- und Kostenaufwand verbunden. Daher wird nach Algorithmen bzw. Indizes geforscht, die die ablaufenden Prozesse des menschlichen visuellen Systems (Human Visual System, HMS) zu simulieren versuchen, um menschennahe Bewertungen abzugeben. Sheikh et al. analysierten und evaluierten diese computionalen Systeme [6], Wang & Bovik entwickelten mathematisch basierte Indizes und forschten an strukturbasierten Algorithmen [7]. Zhang et al. fokussierten sich auf Low Level-Features wie Kanten und Nulldurchgänge [8].

Neuronale Netze sind inzwischen fester Bestandteil des State-of-the-Art der Bildverarbeitung. Auch im Bereich der Bildrestauration kommen spezialisierte neuronale Netze zum Einsatz [9]. Für die Klassifikation von Bilddaten eignen sich besonders tiefe Convolutional Neural Networks (CNN) [10]. In zahlreichen Anwendungsfällen konnte die Erfahrung gewonnen werden, dass künstliche neuronale Netze, vor allem durch ihre starke Fähigkeit der Generalisierung, klassische Deskriptoren und Detektoren in Anwendungsdomänen mit sehr heterogenem Bildmaterial in puncto Genauigkeit deutlich übertreffen [11].

Voraussetzung dafu̇r sind allerdings ausreichend große Trainingsdatensets, die oft aufwendig manuell erstellt werden müssen und leicht mehrere tausend Instanzen pro Klasse beinhalten können. Hier fällt vor allem der Umstand unterschiedlich stark ausgeprägter Features und Klassen ins Gewicht. Kann eine (ungefähre) Gleichverteilung der Trainingsdaten nicht sichergestellt werden, sind entsprechende Transformationen und Augmentationen der Trainingsdaten erforderlich. Ansätze dafür sind weit verbreitet, müssen aber individuell auf den Anwendungsfall angepasst werden.

3 Methodik

3.1 System zur Qualitätsbewertung

Die Festlegung der Anzahl möglicher Qualitätsstufen erfolgte in unserem Ansatz durch eine Orientierung am Mean-Opinion-Score-Verfahren, welches sich auf fünf Abstufungen (mangelhaft, mäßig, ordentlich, gut, ausgezeichnet) stützt. Um nun den Probanden ein leichteres Verständnis für die Qualitätsbeurteilung zu ermöglichen, lehnt sich die Methode in diesem Beitrag an die geläufige Bewertungsmetapher des Sternen-Ratings an. Ein Stern steht für unbrauchbar oder stark beschädigt, zwei Sterne für mittel bis stark beschädigt, drei Sterne für leicht beschädigt oder weiß nicht, vier Sterne für gut und fünf Sterne schließlich für sehr gut. Die Bilddaten wurden jedem Probanden einzeln und in einer randomisierten Reihenfolge dargeboten, um zu verhindern, dass die unvermeidbar stattfindenden Vergleiche nicht zu systematischen Fehlerquellen führen können. Um identische Versuchsbedingungen zu schaffen, orientierte sich die Durchführung der Studie an der Norm ISO 20462.

3.2 Auswahl der Studiendaten

Der Professur Medieninformatik liegen aus einem Pilotprojekt insgesamt 473.478 extrahierte Keyframes aus 237 digitalisierten Videos mit einer Spieldauer von ca. 200 Stunden vor. Aus diesen wurden zufällig 10.000 repräsentative Keyframes ausgewählt. Die Auswahl der durch eine Shot-Detection erzeugten Auszüge konzentrierte sich dabei auf die sogenannten Least-Motion-Keyframes, da davon auszugehen ist, dass diese durch die geringste Bewegung die höchste Bildqualität bieten und bspw. nicht durch Bewegungsunschärfe verfälscht sind [12]. Parallel erfolgte eine Bereinigung der Daten, sodass alle nicht eindeutig erkennbaren oder verfälschten Frames entfernt wurden. Dazu zählen ein- oder mehrfarbige Flächen ohne Text, reines Bandrauschen ohne Hintergrundbild, Blendeneffekte sowie Test- und Doppelbilder. Die verbliebenen 9.855 Frames bilden die Basis unserer Studiendaten (Abbildung 1). Diese wurden durch zehn Multimedia-affine Experten (fünf männliche sowie fünf weibliche Personen im Alter von 23 bis 33 Jahren) im Rahmen einer Studie zur Ermittlung der subjektiv wahrgenommen

Bildqualität bewertet. In einer nachträglichen Befragung gaben die Probanden an, welche Kriterien für sie während ihrer individuellen subjektiven Bildwahrnehmung und -einschätzung ausschlaggebend waren. Als störende Faktoren wurden Falschfarben, Unschärfe, Rauschen, Über- und Unterbelichtung oder das Vorhandensein von Artefakten wie Dropouts, Chroma Bleeding oder Moiré angegeben. Die Nachbefragung macht deutlich, dass eine automatisierte Beurteilung der Bildqualität durch eine Verkettung klassischer Detektoren der Bildverarbeitung nur sehr schwer und oft unvollständig zu erreichen wäre.

Abbildung 1: Beispiele für die Qualitätsstufen stark geschädigt/unbrauchbar (links – Störzone), mittlere Beschädigungen (Mitte – Falschfarben, geringes Rauschen) und sehr gute Bildqualität (rechts – klare Konturen, kräftige Farben)

3.3 Auswahl der Trainingsdaten

Die Annotation von Massendaten ist ein zeitaufwändiger und fehleranfälliger Prozess. Besonderes Augenmerk muss daher auf der Qualitätssicherung der Trainingsdaten liegen. Gerade bei Klassifikationsproblemen können Beispieldaten mit fehlerhaften Labels zu unerwarteten Effekten und verringerter Effizienz der Verfahren führen. Es ist in unserem Fall daher sinnvoll, die Zahl der Trainingsdaten zu reduzieren, um ein homogenes Datenset zu erhalten. Dieser Schritt geht aber auch mit dem Verlust wertvoller Samples einher und muss entsprechend sensibel gestaltet sein.

Die Annotation der Keyframes durch mehrere Personen erlaubt auch die Vergabe unterschiedlicher Labels. Der Mean-Opinion-Score allein liefert keine Aussage zur Eignung einer Annotation zum Training. Wir haben daher nur genau die Frames zum Training verwendet, deren Be-

wertung durch die Nutzer sehr ähnlich waren. Also solche, die bei allen Annotatoren ein ähnlich subjektives Qualitätsempfinden ausgelöst haben. Schwellwert war dabei eine Standardabweichung von nicht mehr als 0.5 über die Einzelbewertungen eines Frames hinweg. Die Menge der Keyframes mit dieser Eigenschaft belief sich auf 5.560 von 9.855 Frames und ist damit bereits deutlich reduziert.

Die ursprüngliche Ungleichverteilung der Bewertungen bleibt nach dieser Reduktion jedoch erhalten. Mehr als die Hälfte der annotierten Frames wurde von den Probanden mit einem Mean-Opinion-Score zwischen zwei und drei bewertet. Hier musste das Trainingsset weiter reduziert werden, um eine übermäßige Verfälschung der Klassifikation zu verhindern. Durch die Begrenzung der Zahl der Keyframes mit einem Mean-Opinion-Score zwischen 2.0 und 3.0 auf 25 Prozent der ursprünglichen Anzahl konnte der Einfluss dieser Ungleichmäßigkeit ausgeglichen werden. Allerdings bestand das finale Trainingsset somit aus nur noch 2.245 von ursprünglich 9.855 annotierten Keyframes.

3.4 Training

Wir verwenden für das Training des Convolutional Neural Networks das auf der Python-Implementierung von Theano basierende Deep Learning-Framework Lasagne. Im Gegensatz zu Caffe oder Tensorflow bietet diese Kombination genügend Flexibilität in der Gestaltung der Netzwerkarchitektur bei gleichzeitiger einfacher und komfortabler Programmierung. Zur GPU-Beschleunigung der Berechnungsvorgänge kommen auch hier das CUDA-Toolkit von NVIDIA und die cuDNN-Erweiterung zum Einsatz. Die Performance ist daher vergleichbar mit den Frameworks von Google oder dem BVLC.

Aufgrund von Einschränkungen der uns zur Verfügung stehenden Hardware haben wir uns für eine Netzwerkarchitektur entschieden, die der von Simonyan & Zisserman vorgestellten VGG-16 Architektur ähnelt [13]. Die Kombination aus Convolutions mit Linear Rectify-Aktivierung und anschließendem Max Pooling nach jedem Convolution-Block hat sich bereits vielfach bewährt und stellt durch ihre Einfachheit eine leicht zu modifizierende Basis für unseren Klassifizierungs-Task dar (Abbildung 2). Allerdings ist die geringe Input-Größe von nur 224×224 Pixeln in der ursprünglichen Implementierung für unsere Zwecke ungeeignet. Grund dafür ist die Vielzahl

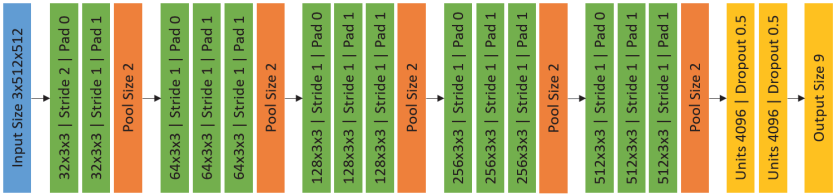


Abbildung 2:
Das in unserem Verfahren eingesetzte CNN ist an die bewährte VGG-16 Architektur angelehnt. (Blau = Input Layer, Grün = Convolutional Layer, Orange = Max Pooling Layer, Gelb = Dense Layer)

von Bildfehlern, die sich durch leichte Unschärfe, Artefakten in einzelnen Bildzeilen und/oder schwaches Rauschen manifestieren. Eine Reduktion der Bildgröße von 720×576 Bildpunkten in Standard PAL auf ein Achtel der ursprünglichen Auflösung sorgt verfahrensbedingt für eine Verbesserung der Bildqualität, bei der Fehler wie Unschärfe oder Rauschen verschwinden und führt damit zu weniger heterogenen Trainingsdaten. Bei einer Batch-Größe von 64 und einer Input-Auflösung von 512×512 Pixeln in drei Kanälen (RGB) liegt die Grenze für die von uns eingesetzte NVIDIA Titan X mit 12 GB RAM im akzeptablen Bereich und wir können davon ausgehen, dass die meisten Bildfehler in ihrer ursprünglichen Form erhalten bleiben.

Die Output-Größe des finalen Dense Layers liegt bei neun (Klassen) und bildet durch die Softmax-Aktivierung die Wahrscheinlichkeit der Zuordnung eines Eingangsbildes zur jeweiligen Qualitätsbewertung ab. Für den Gradientenabstieg kamen Nesterov Accelerated Gradient Updates mit einem Momentum von 0.9 zum Einsatz. Um den Fokus des neuronalen Netzes auf die wesentlichen Unterschiede der Trainingssamples zu legen, haben wir die von Ioffe & Szegedy vorgestellte Batch-Normalisierung vor der Aktivierung der Convolutional Layers eingesetzt [14]. Diese Vorgehensweise liefert nur unwesentlich bessere Ergebnisse bei der Klassifikation, beschleunigt den Lernprozess bis zur Convergence des Netzes allerdings erheblich.

Der Loss-Funktion kam besondere Bedeutung zu. In unserem Ansatz haben wir es mit einem Spezialfall der Klassifikation zu tun, bei dem benachbarte Klassen ähnlich zueinander sind. Das bedeutet, dass eine

Fehlklassifikation eines Frames mit der Ground-Truth-Bewertung von drei Sternen in die Klassen, die eine Bewertung von 2,5 oder 3,5 Sternen repräsentieren, weniger „falsch“ ist, als eine Einordnung in die Klassen „1 Stern“ oder „5 Sterne“. Aus diesem Grund haben wir uns zur Berechnung des Fehlers für eine Kombination aus Mean Squared Error (MSE) und einer Kostenfunktion auf Basis der Distanz aller möglichen Klassifikationen zur eigentlichen Ground-Truth-Annotation entschieden. Diese kostensensitive Erweiterung der Loss-Funktion sorgt in unserem Verfahren für eine signifikante Verbesserung der Fehlerrate und spiegelt das Spektrum der Subjektivität bei der Beurteilung von Bildqualität durch den Menschen deutlich besser wider. Sie multipliziert den MSE und verstärkt den Loss bei extremen Fehlklassifikationen.

Beim Training von neuronalen Netzen kann man üblicherweise beobachten, wie die Fehlerrate im Laufe des Trainings stetig sinkt. Allerdings kann das Absinken des Fehlers auch ein Hinweis auf Overfitting sein. Vor allem bei kleinen Trainingsdatensets ist dies oft bereits nach wenigen Epochen der Fall. Um den Zeitpunkt des Overfittings zu verzögern, haben wir die Trainingsdaten zur Laufzeit künstlich vergrößert. Üblicherweise geschieht das durch Augmentation des Datensets in Form von Bildtransformationen [11]. Hier bieten sich vor allem solche Umwandlungen an, die später auch im realen Anwendungsfall zu erwarten sind und das ursprüngliche Label nicht verfälschen. In unserem Fall eignen sich die klassische horizontale Spiegelung und der Random Crop besonders, da alle Bildfehler global für das gesamte Bild existieren und nicht an einzelnen inhaltlichen Details des Bildes gekoppelt sind. Wir haben diese Transformationen zufällig, mit einer Wahrscheinlichkeit von je 0.5, auf die Trainingsdaten zum Zeitpunkt des Ladens eines Batches angewendet.

4 Evaluation

Zur Evaluation nutzt man üblicherweise eine Teilmenge der Trainingsdaten, die vor dem Training vom Trainingsset getrennt werden und nicht als Samples zur Verfügung stehen. Wir haben uns für eine Teilmenge von zehn Prozent der Trainingsdaten entschieden. Das so entstandene Validierungsset enthielt 224 Keyframes – eine recht geringe Zahl, jedoch hätte die Erhöhung des Validierungsanteils gleichzei-

tig noch weniger Trainingssamples zur Folge gehabt. Wir gehen dennoch von einer repräsentativen Evaluation aus: Wir konnten beobachten, dass die Fehlerrate auch bei Variation der zum Validierungsset gehörenden Frames keine Unterschiede zeigte und das Training stets die gleichen Ergebnisse lieferte.
Wir haben die Ergebnisse unserer Evaluation in drei Runs unterteilt. In jedem davon kam das gleiche Netz und das gleiche Trainingsset zum Einsatz. Run 1 (Conv) enthält das CNN ohne Batch-Normalisierung und dem einfachen MSE als Loss-Funktion. Der zweite Run beinhaltet dagegen die Batch-Normalisierung der Convolutions (Conv+BNorm), der dritte Run verfügt darüber hinaus noch über die erweiterte Kostenfunktion aus MSE und kostensensitiver Klassifikation (Conv+BNorm+Cost). Tabelle 1 listet die Ergebnisse der Evaluation auf dem Validierungsset auf. Vergleichend dazu ist die Default 1 Class Accuracy angegeben, welche die maximale Genauigkeit symbolisiert, die erreicht wird, wenn alle Frames in ein und die selbe Klasse eingeordnet werden, also kein Training stattgefunden hat.

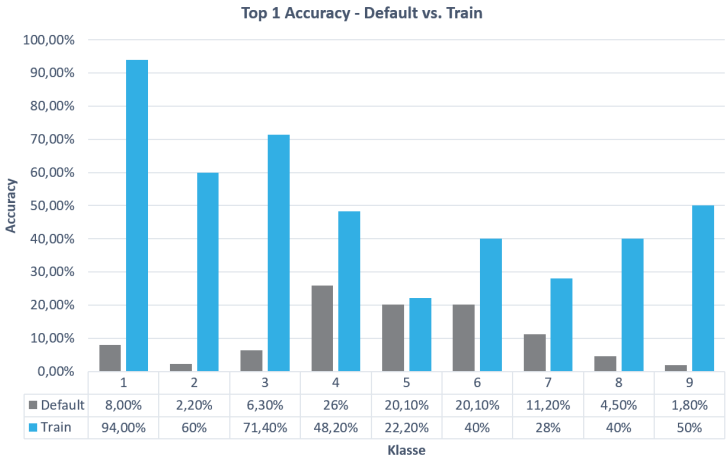
Tabelle 1: Die Evaluation zeigt deutlich den Einfluss der erweiterten Kostenfunktion auf die Top 1 und Top 3N Accuracy. Die Batch-Normalisierung sorgt für einen deutlich frühere Convergence, kostet aber pro Epoche deutlich mehr Zeit.

	Top 1 Acc.	Epoche	Top 3N Acc.	Epoche	Zeit/Epoche
Conv	37,1%	75	72,2%	90	**68 sek**
Conv+BNorm	40,1%	**49**	75,6%	**62**	85 sek
Conv+BNorm+Cost	**44,2%**	67	**83,5%**	73	85 sek
Default 1 Class	26,0%	-	66,1%	-	-

Wir konnten nach einem Training von jeweils 100 Epochen eine maximale Top 1 Accuracy von 44,2 Prozent erreichen. Ausschlaggebend war hier vor allem die erweiterte Loss-Funktion. Besonders die Klassen der Randbereiche ($1-3$ und $7-9$) profitieren sehr stark von dieser Methode und weisen eine Top 1 Accuracy von maximal 94 Prozent und eine deutliche Steigerung gegenüber der Default-Klassifizierung bei zufälligem Raten auf (Abbildung 3). Die Normalisierung der Batches in den Convolutions zeigte erwartungsgemäß keine große Steigerung der Genauigkeit, sorgte aber für einen deutlich früheren Zeitpunkt der besten Va-

lidierungsgenauigkeit. Der Preis dafür ist allerdings die erhöhte Dauer der Trainings einer Epoche, welche jedoch mit 85 Sekunden noch im akzeptablen Bereich liegt. Die Klassifikation eines Einzelbildes liegt im finalen Netz bei nur etwa 20 ms.

Die Top 1 Accuracy ist ein sehr restriktives Maß und spiegelt in unserem Fall auch die Güte des Verfahrens nur unzureichend wider. Wir haben daher ein weiteres Evaluationsmaß bestimmt. Bei der Top 3N Accuracy zählen wir neben den exakten Treffern in der Klassifikation auch solche hinzu, die in benachbarte Klassen eingeordnet wurden (N steht dabei für Nachbar). Das ist valide, da wir es in unserem Anwendungsfall mit einem Spezialfall der Klassifikation zu tun haben und davon ausgehen können, dass Unterschiede in benachbarten Klassen bezüglich der Bildqualität nur gering sind und im Rahmen der subjektiven Wahrnehmung des Menschen liegen. Die so ermittelte Genauigkeit bei der Klassifizierung der Validierungsdaten liegt bei 83,5 Prozent nach 73 Epochen. Auch hier ist die Steigerung durch die erweiterte Kostenfunktion sehr deutlich.



	1	2	3	4	5	6	7	8	9
Default	8,00%	2,20%	6,30%	26%	20,10%	20,10%	11,20%	4,50%	1,80%
Train	94,00%	60%	71,40%	48,20%	22,20%	40%	28%	40%	50%

Abbildung 3: Im Vergleich mit der Wahrscheinlichkeit einer korrekten Klassifizierung durch zufälliges Raten (Default) wird deutlich, dass vor allem Frames mit extrem guter oder extrem schlechter Qualität vom Training mit einer erweiterten Kostenfunktion profitieren (Train).

5 Zusammenfassung und Ausblick

Die in unserem Paper vorgestellte Methodik erlaubt eine automatisierte Qualitätsbeurteilung zu einem gegebenen (S)VHS-Archivbestand, die ähnlich einer subjektiven Bewertung durch einen Mean-Opinion-Score die Auswahl von geeignetem Archivmaterial für Videobeiträge erleichtert. Das Zusammenspiel aus Datensetaugmentation, größerer Input-Auflösung, Batch-Normalisierung und erweiterter Kostenfunktion stellt eine effektive Methode zur automatisierten Beurteilung der Bildqualität auf Basis des Mean-Opinion-Scores durch ein künstliches neuronales Netz dar. Die geringe Größe des Trainingsdatensets hat sicher einen Einfluss auf das Gesamtergebnis und sollte in weiteren Versuchen sukzessive gesteigert werden. Wir gehen davon aus, dass auch die Steigerung der Komplexität des neuronalen Netzes einen positiven Einfluss haben kann, allerdings ist es fraglich, ob die damit steigende Berechnungszeit und höheren Hardwareanforderungen einen solchen Einsatz rechtfertigen.

Es ist denkbar, zukünftig die Ergebnisse der nachträglichen Probandenbefragung stärker einzubinden und durch gezielte Verstärkung besonders auffälliger Bildstörungen die Heterogenität der Trainingssamples weiter zu steigern. Die Anbindung des vorgestellten Verfahrens an das Web-Interface einer Recommender-Plattform ist leicht über eine Kombination von JavaScript und eines Python-Servers möglich und auf gängiger Consumer-Hardware effizient lauffähig.

Literatur

1. M. Eibl, J.-M. Loebel und H. Reiterer, „Grand Challenge, Erhalt des digitalen Kulturerbes“, *Informatik-Spektrum*, Vol. 38, Nr. 4, S. 269–276, 2015.
2. A. Berger, M. Eibl, S. Heinich, R. Herms, S. Kahl, J. Kürsten, A. Kurze, R. Manthey, M. Rickert und M. Ritter, „ValidAX-Validierung der Frameworks AMOPA und XTRIEVAL“, *Chemnitzer Informatik-Berichte ; CSR-15-01*, 2015.
3. S. Müller, S. Kahl und M. Eibl, „Processing digitized (S) VHS archives: An approach of content-based retargeting for local TV stations“, in *Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia*. ACM, 2015, S. 259–262.

4. K. Seshadrinathan, R. Soundararajan, A. C. Bovik und L. K. Cormack, „Study of subjective and objective quality assessment of video“, *IEEE transactions on image processing*, Vol. 19, Nr. 6, S. 1427–1441, 2010.
5. M. H. Pinson und S. Wolf, „Comparing subjective video quality testing methodologies“, in *Visual Communications and Image Processing 2003*. International Society for Optics and Photonics, 2003, S. 573–582.
6. H. R. Sheikh, M. F. Sabir und A. C. Bovik, „A statistical evaluation of recent full reference image quality assessment algorithms“, *IEEE Transactions on image processing*, Vol. 15, Nr. 11, S. 3440–3451, 2006.
7. Z. Wang, A. C. Bovik, H. R. Sheikh und E. P. Simoncelli, „Image quality assessment: from error visibility to structural similarity“, *IEEE transactions on image processing*, Vol. 13, Nr. 4, S. 600–612, 2004.
8. L. Zhang, L. Zhang, X. Mou und D. Zhang, „FSIM: a feature similarity index for image quality assessment“, *IEEE transactions on Image Processing*, Vol. 20, Nr. 8, S. 2378–2386, 2011.
9. S. Iizuka, E. Simo-Serra und H. Ishikawa, „Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification“, *ACM Transactions on Graphics (TOG)*, Vol. 35, Nr. 4, S. 110, 2016.
10. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke und A. Rabinovich, „Going deeper with convolutions“, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, S. 1–9.
11. A. Krizhevsky, I. Sutskever und G. E. Hinton, „Imagenet classification with deep convolutional neural networks“, in *Advances in neural information processing systems*, 2012, S. 1097–1105.
12. M. Ritter, „Optimierung von Algorithmen zur Videoanalyse: Ein Analyseframework für die Anforderungen lokaler Fernsehsender“, Dissertation, Universitätsverlag der Technischen Universität Chemnitz, 2014.
13. K. Simonyan und A. Zisserman, „Very deep convolutional networks for large-scale image recognition“, *arXiv preprint arXiv:1409.1556*, 2014.
14. S. Ioffe und C. Szegedy, „Batch normalization: Accelerating deep network training by reducing internal covariate shift“, *arXiv preprint arXiv:1502.03167*, 2015.

Extended photometric stereo model

Thomas Stephan[1,2], Jürgen Dürrwang[3], Jan Burke[1], Stefan Werling[1,4] and Jürgen Beyerer[1,2]

[1] Fraunhofer IOSB
Fraunhoferstr. 1, Karlsruhe 76131, Germany
[2] Vision and Fusion Laboratory (IES), Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT)
Adenauerring 4, Karlsruhe 76131, Germany
[3] Institute for Energy-Efficient Mobility (IEEM)
University of Applied Science
International University Campus 3, Bruchsal 76646, Germany
[4] Baden-Württemberg Cooperative State University Mannheim
Coblitzallee 1-9, Mannheim 68163, Germany

Abstract Photometric stereo has shown its merits in industrial computer-aided optical inspection, but until now has only been used for defect detection as opposed to 3D reconstruction of the inspected surfaces. The reason is that in practice, the geometry of the measurement is not modelled accurately enough to infer surface shapes quantitatively. We derive a new extended photometric stereo model from physically based radiometric relations and effects. The new model is sufficiently accurate for surface reconstruction, nominal-actual comparison and reverse engineering. Thus the measurement of surface normals, which is routinely done on specular surfaces with deflectometry and offers very high sensitivity for surface defects, is now possible on scattering surfaces as well. The formalism is very general and can easily be adapted to different hardware set-ups.

Keywords Photometric stereo, shape from shading, 3D reconstruction, surface normals.

1 Introduction

The automation of quality assurance tasks plays an important role in industrial production processes, with very high potential for further

productivity improvements. One of the overarching goals is to achieve full quality monitoring where every produced component undergoes a nominal-actual comparison. New methods and measurement instruments are needed for many of these automation applications. Especially "aesthetic" object surfaces, whether dull or glossy – such as car bodies – often have to satisfy strict requirements.

Over the past decades, many computer-aided systems have been established to inspect and measure the 3D shape of surfaces, such as structured-light scanners [1] or deflectometric systems [2, 3]. 3D shape reconstruction can be divided into two main classes. Firstly, there are absolute shape measuring systems, e. g., structured-light scanners, stereo vision systems, or interferometric instruments. Secondly, there are systems that measure or estimate the surface normals of the shape, e. g., deflectometric systems or photometric stereo devices [4]. In devices measuring normals, the measurand is the surface slope (gradient). This creates very high sensitivity to local surface features and defects [3].

In contrast to deflectometric systems, photometric stereo devices are intrinsically able to reconstruct shape data from dull surfaces. The basic idea of photometric stereo or shape from shading was first formulated by Horn [5] in 1970, where an approach was presented that can recover shapes from shaded images. This early work provides the basis of many subsequent shape from shading and photometric stereo algorithms. Some restrictions of the original approach were later overcome, e. g., by extensions to glossy surfaces [6] or spatially varying reflectance [7, 8]. Other researchers focussed on improving the computational efficiency [9] and robustness [10]. Handling of unknown light conditions [11] or uncalibrated systems [12] were also topics of research.

However, the simplified assumptions in photometric modelling still prevent a low-uncertainty 3D recovery, so that photometric stereo 3D reconstruction systems still lack in industrial automated visual quality assurance. In this paper, we derive a more accurate photometric stereo model (sec. 2). It is based on the underlying radiometric relations and effects and thus is only limited by the precision of the measuring unit's hardware layout. In sec. 3.1 such an experimental setup for realising photometric stereo measurements is presented. Sec. 3.2 shows some evaluations using the derived model to demonstrate the performance of the proposed method. Possible extensions for further performance enhancement are proposed in sec. 4.

2 Extended photometric stereo model

In the past the basis of many photometric stereo approaches were limiting assumptions like Lambertian scattering, telecentric illumination and orthographic camera projection. Taking these assumptions, the photometric stereo equation can be easily written as

$$g = \rho \ \langle \boldsymbol{b}, \boldsymbol{n} \rangle \ , \tag{1}$$

where g is the observed intensity, ρ is the Lambertian surface reflection, $\boldsymbol{b}$ is the normalised direction toward the telecentric light source, $\boldsymbol{n}$ is the normalised surface normal, and $\langle \boldsymbol{b}, \boldsymbol{n} \rangle$ denotes the dot product of $\boldsymbol{b}$ and $\boldsymbol{n}$, and is hence proportional to the cosine of the angle between them (see fig. 1). The photometric stereo problem can then be addressed by changing illumination direction and solving the arising linear system of equations

$$\boldsymbol{g} = \rho \, \boldsymbol{B}^T \boldsymbol{n} \ , \tag{2}$$

where $\boldsymbol{g}$ is the vector of imaged pixel intensities and $\boldsymbol{B} = (\boldsymbol{b}_1 | \ldots | \boldsymbol{b}_N)$ is the matrix of the different illumination direction vectors $\boldsymbol{b}_i$.

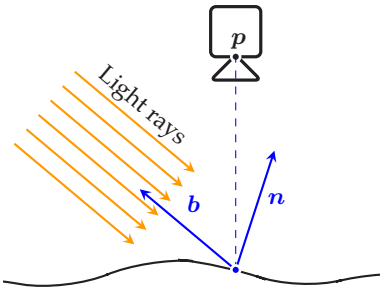


Figure 1: The basis of most photometric stereo models is a telecentric illumination, an orthographic camera projection, and a Lambertian surface. Here $\boldsymbol{b}$ is the direction of the light source and $\boldsymbol{n}$ is the surface normal vector.

2.1 Photometric model

The model assumptions described in sec. 2 are inflexible and frequently not applicable for visual inspection and quality assurance tasks. Thus we will extend the orthographic projection of the camera to a more general pinhole camera model, and the telecentric directed light will be changed to a point light source. Fig. 2 illustrates this configuration.

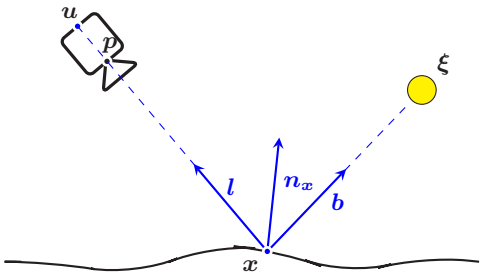


Figure 2: The configuration for the extended model consists of a pinhole camera at point $\boldsymbol{p}$, a point light source at $\boldsymbol{\xi}$, and a surface point $\boldsymbol{x}$ with its normalised surface normal $\boldsymbol{n_x}$.

The observed photometric scene consists of the pinhole camera whose projection centre is located at the scene point $\boldsymbol{p}$, the point light source located at $\boldsymbol{\xi}$, and the surface point $\boldsymbol{x}$. Thus, the observed surface point $\boldsymbol{x}$ is projected onto the camera pixel $\boldsymbol{u}$. In order to derive the extended photometric stereo model, three normalised vectors $\boldsymbol{l}, \boldsymbol{n_x}, \boldsymbol{b} \in \mathcal{S}^2 = \{\boldsymbol{r} \in \mathbb{R}^3 | \, \|\boldsymbol{r}\| = 1\}$ are introduced, where $\boldsymbol{n_x}$ is the normalised surface normal at point $\boldsymbol{x}$, $\boldsymbol{b}$ is directed from $\boldsymbol{x}$ towards the light source located at $\boldsymbol{\xi}$, and $\boldsymbol{l}$ is directed from $\boldsymbol{x}$ towards the projection centre $\boldsymbol{p}$ of the pinhole camera. Specifically, the constructions are

$$\boldsymbol{b} := \frac{\boldsymbol{\xi} - \boldsymbol{x}}{\|\boldsymbol{\xi} - \boldsymbol{x}\|} \quad \text{and} \quad \boldsymbol{l} := \frac{\boldsymbol{p} - \boldsymbol{x}}{\|\boldsymbol{p} - \boldsymbol{x}\|} = \frac{\boldsymbol{u} - \boldsymbol{p}}{\|\boldsymbol{u} - \boldsymbol{p}\|} \, . \tag{3}$$

Camera model

As already mentioned, the used camera model is a pinhole camera model, in which the image intensity $f(\boldsymbol{u})$ at pixel $\boldsymbol{u}$ can be either easily

modelled as linear concerning the incoming radiance,

$$f(\boldsymbol{u}) \propto L_i(\boldsymbol{p}, \boldsymbol{r}_i) \,, \tag{4}$$

or more generally as linear regarding the incoming radiance $L_i(\boldsymbol{p}, \boldsymbol{r}_i)$, integrated over the solid angle $\Omega_{\boldsymbol{u}} \subset \mathcal{S}^2$ spanned by the finite pixel projection. This is written as

$$f\left(\boldsymbol{u}\right) \propto \int_{\Omega_{\boldsymbol{u}}} L_i\left(\boldsymbol{p}, \boldsymbol{r}_i\right) \mathrm{d}\boldsymbol{r}_i \,. \tag{5}$$

Considering the projected pixel area $\mathcal{A}_{\boldsymbol{u}}$ on the scene surface and the conservation of radiance in geometric optics, the pixel intensity can also be calculated by integrating over the projected pixel $\boldsymbol{x}$ inside $\mathcal{A}_{\boldsymbol{u}}$ by integral substitution of (5):

$$f\left(\boldsymbol{u}\right) \propto \int_{\mathcal{A}_{\boldsymbol{u}}} L_i\left(\boldsymbol{x}, \boldsymbol{l}\right) \frac{1}{\left\|\boldsymbol{p} - \boldsymbol{x}\right\|^2} \mathrm{d}\boldsymbol{x} \,. \tag{6}$$

Surface reflection

The surface reflection can be described by the so called bidirectional reflectance distribution function (BRDF) $\rho(\boldsymbol{x}, \boldsymbol{r}_i, \boldsymbol{r}_o)$ that depends on the direction $\boldsymbol{r}_i \in \mathcal{S}^2$ of incoming light and the observed outgoing direction $\boldsymbol{r}_o \in \mathcal{S}^2$. The outgoing reflected radiance $L_o(\boldsymbol{x}, \boldsymbol{l})$ can be calculated as a weighted integration of incoming radiance over the hemisphere $\Omega \subset \mathcal{S}^2$

$$L_o(\boldsymbol{x}, \boldsymbol{l}) = \int_{\Omega \subset \mathcal{S}^2} \rho\left(\boldsymbol{x}, \boldsymbol{r}_i, \boldsymbol{l}\right) L_i\left(\boldsymbol{x}, \boldsymbol{r}_i\right) \left\langle -\boldsymbol{r}_i, \boldsymbol{n}_{\boldsymbol{x}} \right\rangle \mathrm{d}\boldsymbol{r}_i \,, \tag{7}$$

where $\langle -\boldsymbol{r}_i, \boldsymbol{n}_{\boldsymbol{x}} \rangle$ again follows the cosine of the angle between $-\boldsymbol{r}_i$ and $\boldsymbol{n}_{\boldsymbol{x}}$. This integral can also be expressed as an integration over an area $\mathcal{A}_{\boldsymbol{\xi}}$ by integral substitution

$$L_o(\boldsymbol{x}, \boldsymbol{l}) = \int_{\mathcal{A}_{\boldsymbol{\xi}}} \rho\left(\boldsymbol{x}, -\boldsymbol{b}, \boldsymbol{l}\right) L_i\left(\boldsymbol{x}, -\boldsymbol{b}\right) \left\langle \boldsymbol{b}, \boldsymbol{n}_{\boldsymbol{x}} \right\rangle \frac{1}{\left\|\boldsymbol{x} - \boldsymbol{\xi}\right\|^2} \mathrm{d}\boldsymbol{\xi} \,. \tag{8}$$

Light source

A point light is a light source emitting with finite radiance from an infinitesimal extent. This means that the radiance field $L_{\boldsymbol{\xi}}(\boldsymbol{x}, \boldsymbol{r})$ emitted

by the light source at $\boldsymbol{\xi}$ can only be modelled by using a Dirac delta function

$$L_{\boldsymbol{\xi}}\left(\boldsymbol{x}, \boldsymbol{r}\right) = I_{\boldsymbol{\xi}}\left(\boldsymbol{r}\right) \delta\left(\boldsymbol{x} - \boldsymbol{\xi}\right) \ , \tag{9}$$

where $I_{\boldsymbol{\xi}}\left(\boldsymbol{r}\right)$ is the finite intensity that is emitted into direction $\boldsymbol{r} \in \mathcal{S}^2$.

The resulting extended photometric stereo model can be derived by plugging (9) into (8) invoking the conservation of radiance and then plugging the result into (4). This leads to

$$f(\boldsymbol{u}) \propto \rho\left(\boldsymbol{x}, -\boldsymbol{b}, \boldsymbol{l}\right) I_{\boldsymbol{\xi}}\left(-\boldsymbol{b}\right) \left\langle \boldsymbol{b}, \boldsymbol{n}_{\boldsymbol{x}} \right\rangle \frac{1}{\left\| \boldsymbol{x} - \boldsymbol{\xi} \right\|^2} \ . \tag{10}$$

In general, of course, the position of the projected surface point $\boldsymbol{x}$ is unknown. It is located on the sight line of the camera pixel $\boldsymbol{u}$, which can be defined as

$$\boldsymbol{x}(\tau) := \boldsymbol{p} - \tau \boldsymbol{l} \ , \tag{11}$$

where $\tau \in [0, \infty)$.

Assuming that the light source position $\boldsymbol{\xi}$, its intensity function $I_{\boldsymbol{\xi}}(\,\cdot\,)$, the camera position $\boldsymbol{p}$, its geometric calibration and the BRDF $\rho(\,\cdot\,)$ are known, we now have a complete description of the measurement process. It still has four degrees of freedom, which are the path parameter τ, a proportionality factor σ and the two degrees of freedom of the surface normal $\boldsymbol{n}_{\boldsymbol{x}} \in \mathcal{S}^2$. In order to simplify the resulting model formula, the proportionality factor σ is combined with the surface normal $\boldsymbol{n}_{\boldsymbol{x}}$ to $\widetilde{\boldsymbol{n}}_{\boldsymbol{x}} = (n_1 \ n_2 \ n_3)^T := \sigma \boldsymbol{n}_{\boldsymbol{x}}$. Condensing these parameters into one parameter vector $\boldsymbol{\beta} = (\tau \ n_1 \ n_2 \ n_3)^T$ yields the extended photometric stereo model

$$f_{\boldsymbol{\xi}}(\boldsymbol{u}, \boldsymbol{\beta}) = \frac{I_{\boldsymbol{\xi}}\left(-\boldsymbol{b}(\tau)\right)}{\left\| \boldsymbol{x}(\tau) - \boldsymbol{\xi} \right\|^2} \, \rho(\boldsymbol{x}(\tau), -\boldsymbol{b}(\tau), \boldsymbol{l}) \ \left\langle \boldsymbol{b}(\tau), \widetilde{\boldsymbol{n}}_{\boldsymbol{x}} \right\rangle \ , \tag{12}$$

with

$$\boldsymbol{b}(\tau) := \frac{\boldsymbol{\xi} - \boldsymbol{x}(\tau)}{\left\| \boldsymbol{\xi} - \boldsymbol{x}(\tau) \right\|} \ . \tag{13}$$

2.2 Solving the photometric stereo problem

Introducing the extra variables as detailed above of course commits us to estimating the parameter vectors $\boldsymbol{\beta}$ for all pixels $\boldsymbol{u}$. Commonly – in the sense of photometric stereo – an image sequence $\boldsymbol{g}(\boldsymbol{u}) := \big(g_{\boldsymbol{\xi}_1}(\boldsymbol{u}) \ \dots \ g_{\boldsymbol{\xi}_N}(\boldsymbol{u})\big)^T$ is captured under different illumination conditions with changing light positions $\boldsymbol{\xi}_i$, where $i \in \{1, \dots, N\}$ with $N \geq 4$. The parameter vector $\boldsymbol{\beta}$ can be estimated by minimising the sum of squared differences (*least squares*)

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{u}} := \arg\min_{\boldsymbol{\beta}} \ \|\boldsymbol{g}(\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{u}, \boldsymbol{\beta})\|^2 \ , \tag{14}$$

where $\boldsymbol{f}(\boldsymbol{u}, \boldsymbol{\beta}) = \big(f_{\boldsymbol{\xi}_1}(\boldsymbol{u}, \boldsymbol{\beta}) \ \cdots \ f_{\boldsymbol{\xi}_N}(\boldsymbol{u}, \boldsymbol{\beta})\big)^T$ represents the extended photometric stereo model from (12) with different light positions $\boldsymbol{\xi}_i$. In practice, (14) can be solved by e. g. a Gauss-Newton algorithm.

Gauss-Newton algorithm

The Gauss-Newton algorithm is a method to solve non-linear least squares problems such as (14). To do so, the Jacobian matrix $\boldsymbol{J}_{\boldsymbol{\beta}}(\boldsymbol{u})$ of the residuum

$$\boldsymbol{r}(\boldsymbol{u}, \boldsymbol{\beta}) := \boldsymbol{g}(\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{u}, \boldsymbol{\beta}) \tag{15}$$

is needed for every pixel $\boldsymbol{u}$. The entries of the Jacobian matrix $\boldsymbol{J}_{\boldsymbol{\beta}}(\boldsymbol{u})$ are the partial derivatives of the residuum $\boldsymbol{r}(\boldsymbol{u}, \boldsymbol{\beta})$ with respect to the components of the parameter vector $\boldsymbol{\beta}$:

$$\big(\boldsymbol{J}_{\boldsymbol{\beta}}(\boldsymbol{u})\big)_{ij} = \frac{\partial r_i(\boldsymbol{u}, \boldsymbol{\beta})}{\partial \beta_j} \ . \tag{16}$$

With that, the iterative Gauss-Newton algorithm is given by

$$\boldsymbol{\beta}_{\boldsymbol{u}}^{(s+1)} = \boldsymbol{\beta}_{\boldsymbol{u}}^{(s)} - \Big(\boldsymbol{J}_{\boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}}^T(\boldsymbol{u}) \, \boldsymbol{J}_{\boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}}(\boldsymbol{u})\Big)^{-1} \boldsymbol{J}_{\boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}}^T(\boldsymbol{u}) \, \boldsymbol{r}(\boldsymbol{u}, \boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}) \ . \tag{17}$$

This equation must be computed for each pixel $\boldsymbol{u}$. The partial derivatives in (16) can be calculated numerically or in some special cases even analytically. In order to stabilise the Gauss-Newton algorithm, a

regularisation term can be introduced, which leads to the Levenberg-Marquardt formulation

$$\boldsymbol{\beta}_{\boldsymbol{u}}^{(s+1)} = \boldsymbol{\beta}_{\boldsymbol{u}}^{(s)} - \left(\boldsymbol{J}_{\boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}}^{T}(\boldsymbol{u})\, \boldsymbol{J}_{\boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}}(\boldsymbol{u}) + \lambda \boldsymbol{D}_{\boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}}^{T}(\boldsymbol{u})\, \boldsymbol{D}_{\boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}}(\boldsymbol{u}) \right)^{-1} \boldsymbol{J}_{\boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}}^{T}(\boldsymbol{u})\, \boldsymbol{r}(\boldsymbol{u}, \boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}) \,, \tag{18}$$

where λ is the regularisation parameter and $\boldsymbol{D}_{\boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}}^{T}(\boldsymbol{u})\, \boldsymbol{D}_{\boldsymbol{\beta}_{\boldsymbol{u}}^{(s)}}(\boldsymbol{u})$ is a positive definite matrix whose addition is intended to ensure decreasing residues, provided λ is adapted correctly. The larger λ is set, the smaller the step sizes in gradient direction will be, and thus, the slower the algorithm will converge. In our evaluations we have used $\lambda = 0.5$.

3 Evaluation

For an evaluation of the new formalism, an experimental setup was created that will be described in sec. 3.1. We still have two assumptions left in this setup: Firstly, the light source intensity function $I_{\boldsymbol{\xi}}(\,\cdot\,)$ from (9) has been modelled as a point source with $I_{\boldsymbol{\xi}}(\,\cdot\,) \equiv I_{\boldsymbol{\xi}}$, which means that the emitted intensity is constant over all outgoing directions. This is approximated in practice by making the sources as diffuse as possible. Secondly, the BRDF at each point is still modelled as a Lambertian surface, that is $\rho(\,\cdot\,) \equiv \rho$. In practice one can hardly determine the real BRDF – a goniometer would be needed. We will discuss in sec. 4 how this restriction could be overcome in the future. In order to evaluate the accuracy of the parameter estimation, a 3D reconstruction from the normal field is used.

3.1 Experimental setup

The aim for a well designed experimental setup is to determine the light positions $\boldsymbol{\xi}_i$, the intensity functions $I_{\boldsymbol{\xi}_i}(\,\cdot\,)$, and the camera sight lines $\boldsymbol{x}(\tau)$ for each pixel $\boldsymbol{u}$. Therefore small LED lights are mounted into a black hemisphere so that the positions of all LEDs are known. To approximate a uniform intensity function $I_{\boldsymbol{\xi}}(\,\cdot\,)$, a diffusion film is added in front of the light sources so that $I_{\boldsymbol{\xi}}(\,\cdot\,) \approx I_{\boldsymbol{\xi}}$. In order to overdetermine the minimisation formula (12), 48 LEDs are distributed over

the surface of the used hemisphere. The camera is mounted on top of the hemisphere. The experimental setup is shown in fig. 3.

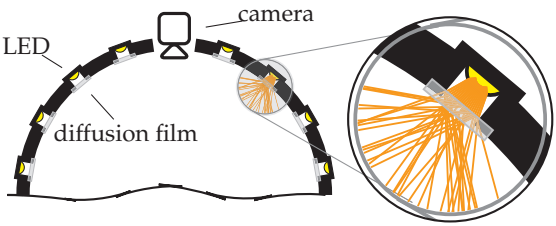
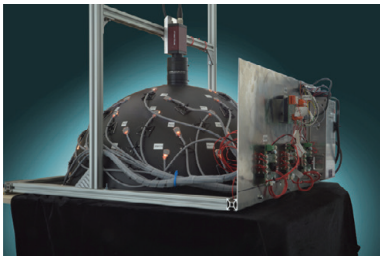


Figure 3: The experimental setup consists of a black hemisphere with mounted camera and LEDs. Inset: to achieve diffuse emittance characteristics, diffusion films were placed in front of each LED.

In order to determine the sight lines $\boldsymbol{x}(\tau)$, the camera was calibrated with standard geometrical calibration methods. Fig. 4 shows the final real setup.



(a) external view

(b) inner view

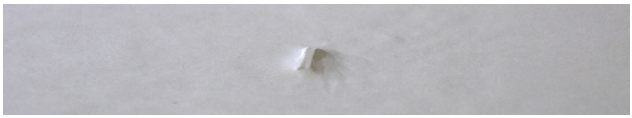
Figure 4: The final setup consists of the black hemisphere, the mounted camera, and the LEDs with control unit.

3.2 Experiments and results

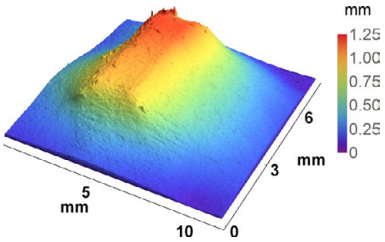
In order to assess the capability of the extended photometric stereo method with respect to industrial quality assurance tasks, we present two experiments.

Quantitative performance

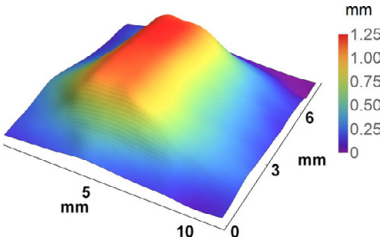
As a comparison with low-uncertainty reference data, a piece of paper with a small bump was measured with an established microscopy 3D reconstruction device and compared to our photometric stereo method. Fig. 5 shows the results.



(a) photo of original specimen



(b) microscopic reference measurement

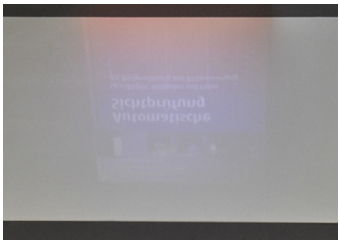


(c) photometric stereo measurement

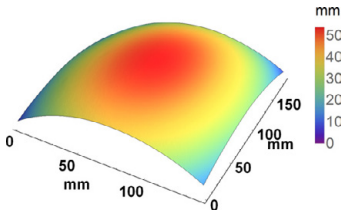
Figure 5: Comparison between microscopic measurement (b) and photometric measurement (c). The height and shape of the specimen match well between the two methods; the lateral resolution of the microscope is much higher.

Limitations of the implementation

As mentioned in sec. 3.1, our implementation of the extended photometric stereo model is still based on the Lambertian surface assumption. In case of mismatch with the actual BRDF, the accuracy of the reconstruction is impaired. For instance, the scattering properties of semi-specular sheet metal will create systematic errors, which can be seen in fig. 6. The specimen was a flat surface, which comes out as curved in the 3D reconstruction due to its non-Lambertian scattering.



(a) original specimen reflecting a scene



(b) 3D reconstruction

Figure 6: (a) Photo of partially specular, planar specimen with scene reflected by the surface; (b) reconstruction of surface assuming Lambertian BRDF. Note the large range of the z axis.

4 Conclusions

In this paper an extension of the classical photometric stereo approach was derived. Thanks to radiometrically and geometrically more accurate modelling, improved reconstructions result from the parameter estimation. The performance of the proposed methodology suggests that automatic quality assurance can be done with the shown experimental setup. The new method offers the potential to inspect large surfaces – compared to microscopic surface inspection – in a short time with low uncertainty, and to determine surface normals quantitatively. The proposed approach can be easily adapted to handle non-Lambertian surfaces as discussed below.

Future work

The extended photometric stereo model (12) requires knowledge or a very good guess of the inspected surface reflectance (BRDF). In the vast majority of surface inspection tasks, the BRDF is not known and cannot be measured without great expense of time and money. Therefore future work will deal with simultaneous estimation of surface normals as well as surface reflectance. This can be done by using a parameterised BRDF $\rho(\boldsymbol{b}(\tau), \boldsymbol{l}, \boldsymbol{\beta})$, whose parameters are also estimated in the least squares optimisation step (14).

References

1. R. Valkenburg and A. McIvor, "Accurate 3D measurement using a structured light system," in *Image and Vision Computing*, vol. 16, no. 2. Elsevier BV, Feb. 1998, pp. 99–110.
2. S. Werling, M. Mai, M. Heizmann, and J. Beyerer, "Inspection of specular and partially specular surfaces," *Metrology and Measurement Systems*, vol. 16, no. 3, pp. 415–431, 2009.
3. J. Beyerer, F. Puente León, and C. Frese, *Machine Vision – Automated Visual Inspection: Theory, Practice and Applications*. Springer-Verlag, 2016.
4. M. Heizmann, "Methoden der 3D-Vermessung von Oberflächen," in *Leitfaden zur Inspektion und Charakterisierung von Oberflächen mit Bildverarbeitung*, M. Sackewitz, Ed., vol. 16. Fraunhofer-Verlag, 2016, pp. 19–26.
5. B. Horn, "Shape from Shading: A Method for Obtaining the Shape of a Smooth Opaque Object from One View," Department of Electrical Engineering, MIT, Tech. Rep., 1970.
6. H.-S. Chung and J. Jia, "Efficient photometric stereo on glossy surfaces with wide specular lobes," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. Institute of Electrical & Electronics Engineers (IEEE), Jun. 2008, pp. 1–8.
7. N. Alldrin, T. Zickler, and D. Kriegman, "Photometric stereo with non-parametric and spatially-varying reflectance," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. Institute of Electrical & Electronics Engineers (IEEE), Jun. 2008, pp. 1–8.
8. A. Hertzmann and S. Seitz, "Example-based photometric stereo: shape reconstruction with general, varying BRDFs," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8. Institute of Electrical & Electronics Engineers (IEEE), Aug. 2005, pp. 1254–1264.
9. R. Szeliski, "Fast shape from shading," in *CVGIP: Image Understanding*, vol. 53, no. 2. Elsevier BV, Mar. 1991, pp. 129–153.
10. Q. Zheng and R. Chellappa, "Estimation of illuminant direction, albedo, and shape from shading," in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Institute of Electrical & Electronics Engineers (IEEE), 1991, pp. 540–545.
11. R. Basri, D. Jacobs, and I. Kemelmacher, "Photometric Stereo with General, Unknown Lighting," *Int J Comput Vision*, vol. 72, no. 3, pp. 239–257, 2006.
12. T. Papadhimitri and P. Favaro, "A new perspective on uncalibrated photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1474–1481.

3D reconstruction by a combined structure tensor and Hough transform light field approach

Alessandro Vianello[1,2], Giulio Manfredi[1], Maximilian Diebold[1] and Bernd Jähne[1]

[1] Heidelberg Collaboratory for Image Processing (HCI) at IWR, Heidelberg University, Germany
[2] Robert Bosch GmbH, Robert Bosch Campus 1, Renningen, Germany

Abstract Disparity estimation using the structure tensor is a local approach to determine orientation in Epipolar Plane Images. A global extension would lead to more precise and robust estimations. In this work, a novel algorithm for 3D reconstruction from linear light fields is porposed. It uses a modified Progressive Probabilistic Hough Transform, in combination with the structure tensor, to extract orientations from Epipolar Plane Images edge maps, allowing to achieve high quality disparity maps.

Keywords Light field, Hough transform, structure tensor.

1 Introduction

One of the most important tasks in computer vision and image processing is 3D geometry reconstruction. In recent years, techniques based on *light field* analysis have been widely developed. A 4D light field can be expressed as a collection of several pinhole cameras, with coplanar image planes. Thus, a camera can move linearly in front of a scene, or a linear camera array can capture scene points. These points correspond to a linear trace in the Epipolar Plane Image (EPI), where the slope of the trajectory determines the scene point's distance to the cameras. In literature, various approaches have been proposed to analyze EPIs. In the work of Bolles [1], salient lines are extracted by finding zero crossings

with a Difference of Gaussians filter and then by merging collinear segments. Criminisi et al. [2] propose to extract EPI regions by using photo-consistency either in 2D (*EPI-strips*) or in 3D (*EPI-tubes*). More recently, Wanner [3] used the structure tensor (ST) to estimate the local slope of each pixel in the EPI, obtaining a coarse depth map which is then refined by means of a global optimization. Their method was continued by Diebold [4, 5], who introduced two variants of the ST. In the first, the so called *modified structure tensor*, the inner Gaussian smoothing is replaced by a derivation filter in the x-direction, i. e., the one parallel to the camera motion, in order to be robust against intensity changes along feature paths. In the second variant, termed *2.5D structure tensor*, an additional smoothing in the y-direction, i. e., the one perpendicular to the EPIs, is introduced. Eventually, Kim et al. [6] proposed a method which computes depth estimates around object boundaries, i. e., in the scene's high textured areas, by testing all the possible disparity hypotheses and choosing the one that leads to the best color constancy along the EPI-line.

In this work a new method is introduced, which uses a coarse disparity map, generated with the local ST, together with an edge map of the EPI, to extract feature paths by using a modified version of the Hough transform (HT). The result is a sparse but very accurate disparity map. Differently from the ST, which provides a local evaluation of orientation in EPIs, the HT considers all points along orientations. Therefore, this approach can be considered a *semi-global* method.

2 Structure tensor

For a given image I, smoothed by an inner Gaussian filter $\mathcal{G}_\sigma$ at an *outer scale* σ [7], in order to suppress noise, the ST is defined as:

$$S_{\rho,\,\sigma}(I) = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \mathcal{G}_\sigma * \begin{pmatrix} \left(\frac{\partial \hat{I}}{\partial x}\right)^2 & \frac{\partial \hat{I}}{\partial x} \cdot \frac{\partial \hat{I}}{\partial y} \\ \frac{\partial \hat{I}}{\partial x} \cdot \frac{\partial \hat{I}}{\partial y} & \left(\frac{\partial \hat{I}}{\partial y}\right)^2 \end{pmatrix}, \tag{1}$$

where $\hat{I} = \mathcal{G}_\sigma * I$ is the smoothed image, and the $x-y$ gradients are computed at an *inner scale* ρ [7]. Two of the most common gradient filters are the 3×3 Sobel and Scharr filters. An alternative, which combines the

initial smoothing and differentiation steps, is the derivative of Gaussian (or Gaussian gradient) filter.

The disparity is achieved by using the ST components:

$$d = \tan\theta = \tan\left(\frac{1}{2}\arctan\left(\frac{2S_{12}}{S_{11}-S_{22}}\right)\right), \tag{2}$$

where θ is the orientation of the EPI-line.

Aside from the disparity, a reliability measure $c \in [0,1]$, called coherence, is used to indicate the confidence of the estimated estimation:

$$c = \sqrt{\frac{(S_{11}-S_{22})^2 + 4(S_{12})^2}{(S_{11}+S_{22})^2}}. \tag{3}$$

3 Hough transform

The Hough Transform is an elegant method for estimating parametrized line segments in an image. This approach can locate regular curves in an image, e. g., straight lines, circles, parabolas, and ellipses. In practice, only image points classified as edge points are considered, where the classification can be obtained with any edge detector which yields a binary edge map.

The line detection case is quite straightforward. As shown in figure 1, a line can be expressed in polar coordinates as:

$$\rho = x\cos\theta + y\sin\theta\,, \tag{4}$$

where ρ is the perpendicular distance from the origin to the line (vector **d** in figure 1), and θ is the angle between the x-axis and the **d** vector. With this parametrization, each line l_i in the image can be associated with a pair (ρ_i, θ_i), and the (ρ, θ) plane is called *Hough space* (or *accumulator*). In order to identify lines, the Hough space is discretized in so called *cells* and initially populated with zeros. Then each edge point *votes*, which means it increments by 1 the cell having coordinates (ρ_i, θ_i) with $|\theta_i| \leq \theta_{\max}$. Once voting is complete, cells whose values are local maxima or peaks define the parameters for the lines in the image. Line detection in EPIs by means of the HT has the advantage of being robust against noise and unaffected by occlusions of the feature paths. Moreover, the HT is tolerant to gaps in the edges, because even a partial line can be reconstructed if it has enough support.

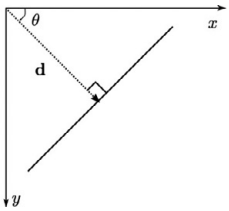


Figure 1: Parametrization of a line in polar coordinates, where $\|\mathbf{d}\| = \rho$.

4 Proposed algorithm

Given its characteristics, a HT based approach seems ideal to treat linear light fields. The sampling of θ is performed by taking the arctangent of linearly spaced disparity values. The *disparity resolution* can be computed from the height of the EPI, i. e., by the number of views N, and is defined as $\Delta d = 1/(N-1)$. A preliminary step to identify lines using the HT is to generate a binary edge map of the EPI. Similarly to Criminisi et al. [2] a Canny edge detector is used. In order to determine the extension of a line (i. e., where it is actually visible), the chosen approach adapts a particular implementation of the HT, the *Progressive Probabilistic Hough Transform* (PPHT) [8], to the characteristics of an EPI. The choice of the PPHT was mainly guided by speed considerations. In fact, for the standard HT, equation (4) has to be solved for every value of θ for every edge point. Therefore, the voting process can be a very costly operation. The advantage of the PPHT is that voting is restricted to a subset of all edge points: to detect a line only as many points have to vote as are required to bring the value of the corresponding accumulator cell above a threshold.

In the following, the algorithm and all its features are described.

Outline In the algorithm, edge points vote in an order defined by a random probability distribution. Once a line has been detected, it is processed in two steps. In the *first step* the end points are determined based on the edge map: the longest segment is found which is either continuous or has gaps of a given maximum length. In the *second step*, all line points remove their votes (if any) from the accumulator and are deleted from the edge map, so that they will not vote anymore.

Although applying the original algorithm is possible, it was decided to leverage the local orientation provided from the ST in order to:

(*a*) reduce the voting range of edge points;

(*b*) control the deletion of points from the edge map in the 2^{nd} step;

(*c*) determine the end points of lines in the disparity map.

Feature *a* To speed up computation and remove noise in the Hough space, it is desirable to restrict the angle range over which an edge point p_i votes to a region around the ST orientation θ'_i. If the coherence c_i at that point is below a threshold c_{th}, p_i will vote over the whole θ range. On the other hand, as the coherence grows, the range can be decreased. So if $c_i \geq c_{th}$, the size of the search range is determined by a linear function of the coherence at that point. More formally, a point p_i votes over $|\theta - \theta'_i| \leq \Delta\theta(c_i)$ with:

$$\Delta\theta(c_i) = \Delta\theta_{\min} + (\theta_{\max} - \Delta\theta_{\min})\frac{c_i - 1}{c_{th} - 1}, \tag{5}$$

where $\Delta\theta_{\min}$ defines the minimum size of the angle range, and $\theta_{\max} = 45°$.

Feature *b* A common situation is that background lines (lower disparity), which are occluded at some point, are detected before the foreground line marking the occlusion boundary (higher disparity). Since in the second step of the algorithm each detected line is deleted from the edge map, these background lines will also delete points from the boundary line, making its detection more difficult. The adopted solution is to delete a point in the edge map only if the ST's disparity is smaller than the detected line's disparity by a margin.

Feature *c* The last feature is needed in cases where lower disparity lines are erroneously propagated over an occlusion boundary (i. e., in a foreground region) due to the sparsity of the EPI edge map. This happens when the maximum line gap is larger than the distance of two neighboring foreground lines, so that these lines end up "supporting" the background line. To detect this scenario, line points for which

the disparity of the line lies below the ST's one by a given margin are counted, and if their number exceeds a threshold the line is marked as "suspect" and saved for further processing.

Line score The *line score* is defined as:

$$\text{line score} = \frac{1}{2}\left(\frac{\text{supported points}}{\text{line length}} + \frac{\text{line length}}{\text{EPI height}}\right), \tag{6}$$

where *supported points* is the number of line points in the edge map. This can be used as a reliability measure, in a similar way to the coherence of the ST (e. g., to merge disparity maps belonging to different refocusing steps and to filter out lines with low scores). As for the coherence, its value is in the range of $[0, 1]$.

Handling of suspect lines At this stage, the disparity map already contains all non-suspect lines, while suspect lines have been saved in a list. This list is sorted by decreasing disparity (i. e., from foreground to background lines), and lines are drawn in the disparity map from a given point until a line with a higher disparity is met (this should be the occlusion point). The problem is finding the point from which to start propagating the line. To this end, the difference between line's and ST's disparity is checked at the two end points of a line, to determine if the end points yield the correct disparity. Thus, three cases arise:

1. Both points are correct: the line is propagated from both points.
2. Only one point is correct: the line is propagated from this point.
3. None of the points is correct: the algorithm walks the line until a point is met where local orientation and line orientation are similar. The line is then propagated from this point in both directions.

5 Experimental evaluation

To assess the performance of the algorithms, 201 synthetic EPIs of constant disparity in the range $\pm 1\,px$, linearly spaced by $0.01\,px$, were generated. Each EPI has an height of 101 pixels, representing 101 views. An

important remark is that the slope of lines in EPIs is not actually determined by a rotation but by a shift of the pixels between views, i. e., by the disparity. This means that orientation can be detected reliably only for lines with disparities in the range $[-1\,px, +1\,px]$; outside this range a line "degenerates into a sequence of disconnected sections" [1, p. 18]. The EPIs are processed with the proposed algorithm and the disparity values of the center row are compared with the ground truth to determine the estimation error. For each disparity value, 50 EPIs were created to make the results statistically reliable, giving a total of $N = 201 \times 50$ EPIs. Additionally, to evaluate the robustness of the algorithms, zero-mean Gaussian white noise of variance σ_n^2 was added to the EPIs; the noise variance refers to intensities in the range $[0, 1]$. The proposed HT approach is compared with the ST, which is implemented with three derivative filters: Gaussian gradient ($[5 \times 5]$), Scharr ($[3 \times 3]$), and Sobel ($[5 \times 5]$). The same Gaussian gradient filter is used to compute the ST, which is then combined with the HT. Moreover, also the modified ST from Diebold [5] was compared. An inner scale for the gradients $\rho = 0.75$ was used, whereas the outer scale σ was set to 1.5. Figure 2 shows the mean errors for each disparity value, for both the noise-free and the noised EPIs ($\sigma_n^2 = 0.01\,px^2$). These plots give a measure of the estimations' biases, showing that the classic ST with Gaussian gradient and of the HT exhibit the smallest bias. Moreover, the noised EPIs case demonstrates the robustness against noise of the HT approach, which dramatically outperforms the others.

To get a measure of the overall accuracy, the *root-mean-square error* (RMSE) over all N EPIs is computed. The results are reported in table 1, for noise-free and noised EPIs. In the noise-free case it can be observed that the HT performs worse than the Gaussian gradient. This is due to the fact that the ST produces continuous orientation estimates. On the other hand, the HT searches for lines with a disparity chosen among all the possible discretized disparities. The discretization depends from the *disparity resolution*, as explained at be beginning of section 4. In the experiments, the disparity resolution is $0.01[px]$. Therefore, the resulting RMSE of 0.0079 agrees with this resolution value. On the other hand, for the case of noised EPIs (which better represents real EPIs), this discretization effect is negligible, and the HT approach gives the best result.

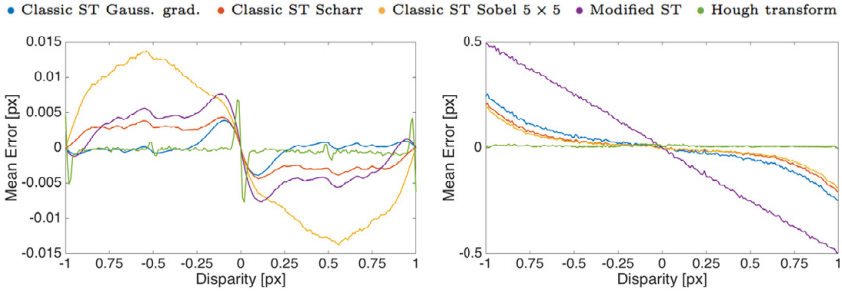


Figure 2: Mean disparity error of the estimates: noise-free EPIs **(left)** and noised EPIs ($\sigma_n^2 = 0.01\,px^2$) **(right)**.

Table 1: RMSE for the different methods, without and with noise.

	RMSE $[px]$	
	noise-less	noisy
Classic ST Gauss. Grad.	**0.0022**	0.2933
Classic ST Scharr	0.0037	0.2394
Classic ST Sobel	0.0114	0.207
Modified ST	0.005	0.4434
Hough Transform	0.0079	**0.1064**

5.1 Light field datasets evaluation

To evaluate the computed disparity maps the *Peak Signal-to-Noise Ratio* (PSNR) [9] is used. In the following, the algorithms are evaluated on synthetic datasets generated with Blender, as well as a real light field. The proposed HT algorithm is compared with two classic structure tensors (Gaussian gradient and Scharr), as well as the modified ST and the 2.5D ST from Diebold [4, 5].

Synthetic datasets Two synthetic datasets of 101 views (i. e., the *Buddha* and *Clutter* dataset) are evaluated. Camera position and baseline were chosen to fit a $2\,px$ disparity range. The PSNR of the center view reconstructions can be found in table 2, where it can be observed that the HT gives the best result. Among the tensor based methods, the 2.5D variant gives the highest PSNR; this is expected as the additional

smoothing of the tensor components removes noise, although it does not mean that the disparity map is more accurate. Figure 3 shows the reconstructed disparities for the center view. These results are visually complicate to evaluate, especially because of the sparsity of the HT. For this reason, the resulting point cloud for the *Buddha* dataset are also provided in figure 3. Here it can be observed that the ST estimation is worse at occlusion boundaries (e. g., around the wooden plank and the column) and at regions with little or no texture (some parts of the dice). While for texture-less areas there is simply no feature path in the EPI, occlusions represent a problem for the ST, which averages between the foreground and background disparities. On the contrary, the HT approach, with its better edge localization properties, gives sharp depth discontinuities and less noise in the texture-less areas.

In order to analyze noise robustness, for the *Buddha* dataset Gaussian noise of variance $\sigma_n^2 = 0.01\,px^2$ is added to each view in the same way as in section 5. Disparity maps where computed with the classic ST and with the HT. The ST gives a PSNR of 8.7099, while for the HT the PSNR is 16.8098. These results confirm that, as observed in section 5, the HT is more robust to noise than the ST.

Table 2: PSNR for the synthetic datasets. For all the ST approaches a coherence threshold of 0.9 has been applied. For the HT a line score threshold of 0.7 is used.

	Classic ST	ST Scharr	ST 2.5D	Modified ST	HT
Buddha	30.21	30.49	30.55	28.72	**30.94**
Clutter	25.01	24.66	26.29	24.59	**27.46**

Real dataset For the evaluation of real data, a Buddha's head sculpture was captured with a single camera mounted on a translation stage. As reference, the Buddha's head was measured with a structured-light scanner. For this dataset, 71 images were acquired, with a baseline of $4\,mm$ and a resulting disparity range of $1.7\,px$. It is important to notice that the reliability of the evaluation is affected by the correctness of the alignment of the estimated and reference point clouds. Numerical results are presented in table 3. The modified ST has now the best score, while the HT scores better than the classic ST but worse than its 2.5D

variant. Unfortunately, as previously stated, a better PSNR does not necessarily mean that the disparity map is better, since it cannot measure aspects like the quality of depth discontinuities and the noise in texture-less areas. The center view disparities are showed in figure 5, and the point clouds in figure 6. Here it is possible to visualize the improvements of the HT reconstruction, which again is more precise and less noisy than the other methods.

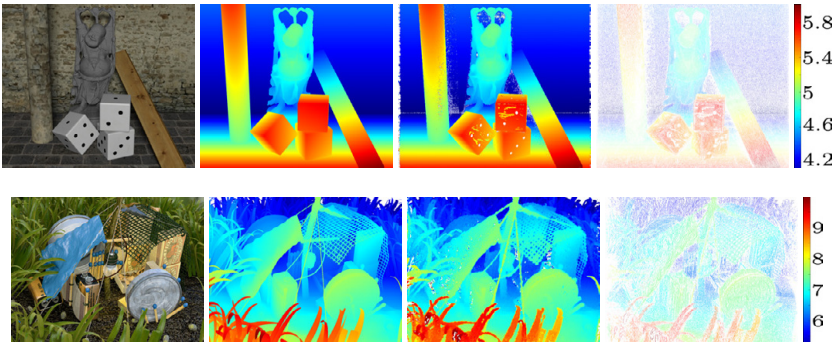


Figure 3: Results for the *Buddha* **(top row)** and *Clutter* **(bottom row)** datasets. **From left to right:** center view, ground truth disparity, classic ST 2.5D disparity, and HT disparity (with line score threshold 0.7).

(a) Classic ST 2.5D (b) Hough Transform

Figure 4: Point clouds for the *Buddha* dataset. The HT approach gives better results, especially at depth discontinuities.

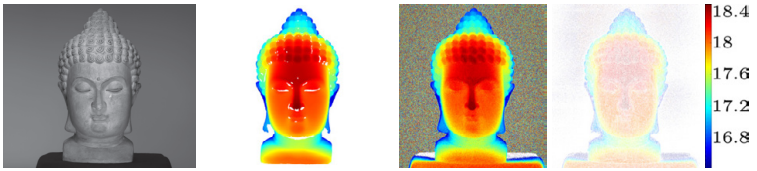
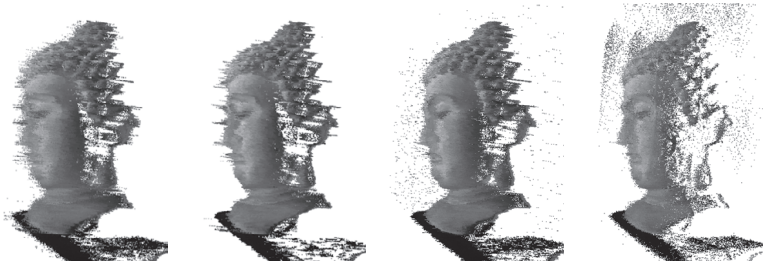


Figure 5: Results for the Buddha's head dataset. **From left to right:** center view, ground truth disparity (from structured-light scanner), modified ST disparity, and HT disparity.



(a) Classic ST (b) Classic ST 2.5D (c) Modified ST (d) HT

Figure 6: Point clouds for the Buddha's head real dataset.

Table 3: PSNR for Budda's head real dataset. For all the ST approaches a coherence threshold of 0.9 has been applied. For the HT a line score threshold of 0.7 is used.

Classic ST	ST Scharr	ST 2.5D	Modified ST	HT
27.69	26.65	29.79	**30.81**	29.17

6 Conclusions and future work

A new semi-global approach for 3D reconstruction from linear light field was presented. This method retrieves reliable EPI lines by exploiting both the benefits of local and global slope estimation. The proposed

approach was compared with three variants of the ST: classic, modified and 2.5D. On synthetic EPI images, in the case of realistic noisy data, the HT approach provides the best results. Eventually, all the methods were evaluated on real and synthetic light fields, confirming the better quality reconstruction of the HT.

In the future, the algorithm should be further improved with respect to occlusion handling, by computing the intersection points directly from the lines' equations. Furthermore, the proposed approach is directly applicable to circular light field, by adapting the parametrization space to circular EPIs.

References

1. R. Bolles, H. Baker, and D. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *International Journal of Computer Vision*, 1987.
2. A. Criminisi, S. Kang, R. Swaminathan, R. Szeliski, and P. Anandan, "Extracting layers and analyzing their specular properties using epipolar-plane-image analysis," *Computer vision and image understanding*, 2005.
3. S. Wanner, J. Fehr, and B. Jähne, "Generating EPI representations of 4D light fields with a single lens focused plenoptic camera," in *Advances in Visual Computing*, 2011.
4. M. Diebold, "Light-Field Imaging and Heterogeneous Light Fields," Ph.D. dissertation, Heidelberg University, 2016. [Online]. Available: http://www.ub.uni-heidelberg.de/archiv/20560
5. M. Diebold, B. Jähne, and A. Gatto, "Heterogeneous Light Fields," in *CVPR*, 2016.
6. C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene Reconstruction from High Spatio-Angular Resolution Light Field," in *Proc. SIGGRAPH*, 2013.
7. S. Wanner and B. Goldlücke, "Globally consistent depth labeling of 4D light fields," in *CVPR*, 2012.
8. C. Galambos, J. Matas, and J. Kittler, "Progressive Probabilistic Hough Transform for line detection," in *CVPR*, 1999.
9. PSNR. [Online]. Available: http://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio

Ein neuartiges multispektrales 3D-Bildaufnahmesystem

A novel multispectral 3D imaging system

Chen Zhang[1], Maik Rosenberger[1] und Gunther Notni[1,2]

[1] Technische Universität Ilmenau, Fachgebiet Qualitätssicherung und Industrielle Bildverarbeitung,
Gustav-Kirchhoff-Platz 2, 98693 Ilmenau
[2] Fraunhofer-Institut für Angewandte Optik und Feinmechnik,
Albert-Einstein-Straße 7, 07745 Jena

Zusammenfassung Der Gegenstand dieser Arbeit ist die Entwicklung eines multispektralen 3D-Bildaufnahmesystems mit einer erhöhten Anzahl der spektralen Kanäle. Dafür wurde ein aktives Stereo-Filterradkamera-System aufgebaut, und ein Ansatz zur geometrisch hochgenauen Fusion der Multispektralbilder beider Kameras mittels 3D-Daten wurde entwickelt. Mit diesem System wird eine zuverlässige Fusion von 3D-Punktwolken und Bilddaten aus maximal 23 spektralen Kanälen realisiert, und eine Verkürzung der Aufnahmezeit lässt sich durch simultane Erfassung des Spektrums mit Stereo-Kameras ermöglichen.

Schlagwörter Optische 3D-Messtechnik, Filterradkamera, Bilddatenfusion.

Abstract This work is aimed at the development of a multispectral 3D imaging system with a major number of spectral channels. For this purpose, an active stereo system based on filter wheel cameras was established, and an approach for the fusion of multispectral images of both cameras via 3D data in high precision was developed. This system realizes a reliable fusion of 3D point cloud and image data from at most 23 spectral channels, and the stereo spectral capturing enables a reduction of acquisition duration.

Keywords Optical 3D measurement, filter wheel camera, image data fusion.

1 Einleitung und Stand der Technik

Die mehrkanalige Bildaufnahme spielt eine immer zunehmende Rolle in den gegenwärtigen Bildverarbeitungsapplikationen. Diese Technologie ermöglicht die Detektion von in den herkömmlichen Bildaufnahmeverfahren unsichtbaren Merkmalen, und mit einer erhöhten spektralen Auflösung noch die Materialcharakterisierung anhand spektraler Signaturen. Ein weiteres wichtiges Teilgebiet der Bildverarbeitung ist das 3D-Imaging, welches das Objekt bzw. die Szene im dreidimensionalen Raum darstellt. Durch die Kombination von diesen beiden Technologien, womit sowohl die 3D-Form als auch die optischen Eigenschaften des Objektes gleichzeitig gewonnen werden können, können sich zahlreiche neuartige Anwendungsmöglichkeiten in der biomedizinischen Bildverarbeitung, in der Dokumentation und multimedialen Visualisierung von Kulturgütern sowie in dem industriellen Bereich, vor allem in der optischen Sortierung und Inspektion eröffnen. Diese Arbeit zielt im Folgenden auf einen prinzipiellen Lösungsansatz für eine stabile und hochgenaue Fusion von 3D-Punktwolken und multispektralen Bilddaten, wobei eine erhöhte Anzahl der spektralen Kanäle bei einer kurzen Aufnahmezeit angestrebt wird.

Zur automatisierten Fusion von 3D- und spektralen Bilddaten existiert schon eine Vielzahl von Arbeiten. Herkömmlich können die 3D-Punktwolke und das Spektralbild mit unterschiedlichen Bildaufnahmegeräten erfasst und zueinander registriert werden. Die Transformationsberechnung in der Bildregistrierung kann entweder mittels einer Co-Kalibrierung zwischen beiden Geräten [1] oder einer merkmalsbasierten Korrespondenzanalyse zwischen dem 3D-Modell und dem 2D-Spektralbild [2] realisiert werden. Derartige Ansätze können eine grundsätzliche Funktionsfähigkeit aufweisen, unterliegen aber einigen Beschränkungen. In dem erstgenannten Ansatz wird ein komplexer Kalibrierprozess benötigt, und im zweiten Ansatz existieren große Schwierigkeiten bei Objekten mit einer geringen Menge von Merkmalen. Außerdem ist damit nur eine begrenzte Genauigkeit bei der Datenfusion zu erzielen. In den oben genannten Arbeiten beträgt der mittlere Registrierungsfehler in der Bildebene mehr als einen Pixel.

Aus diesen Gründen könnte ein Ansatz, in welchem die 3D- und multispektrale Bildaufnahme mit dem gleichen Gerät durchzuführen sind, wesentliche Vorteile aufweisen. In [3] wurde eine Kamera sowohl

zum Laserschnittverfahren als auch zur multispektralen Bildaufnahme mit multispektraler Beleuchtung mittels Monochromator verwendet. Jedoch hat das Prinzip der multispektralen Beleuchtung den Nachteil, dass die Aufnahme durch Umgebungslicht stark beeinflusst wird. Zur Erzielung besserer spektralen Charakteristik wird ein sensorikbasiertes Prinzip für die multispektrale Bildaufnahme benötigt. Unter allen Möglichkeiten ist hinsichtlich der Ortsauflösung nur das Filterrad-Prinzip für hochgenaue 3D-Messungen geeignet. In [4] und [5] wurde ein Filterrad mit optischen Interferenz-Bandpassfiltern in ein triangulationsbasiertes 3D-Scanningssystem integriert, so dass sieben bzw. zehn Kanalbilder aus dem sichtbaren Bereichen auf 3D-Punktwolke gemappt werden können. Die Unschärfe aufgrund Längsfehler wird basierend auf der Punktverwaschungsfunktion (PSF) durch Bildrestaurationsverfahren korrigiert, jedoch bleibt in diesen Arbeiten der transversale Abbildungsfehler unkompensiert. Der Transversalfehler kann mehrere Pixel betragen [6] und damit die Genauigkeit der Datenfusion verschlechtern.

Diese Recherche über den Stand der Technik zeigt, dass die existierenden Systeme entweder hinsichtlich der geometrischen Genauigkeit begrenzt sind, eine nicht ausreichende spektrale Eigenschaft haben, oder nur eine beschränkte Anzahl an spektralen Kanälen bieten können. In dieser Arbeit wird angestrebt, einen bezüglich dieser Kriterien optimierten Lösungsansatz zu entwickeln.

2 Konzeption des neuartigen multispektralen 3D-Bildaufnahmesystems

2.1 Systementwurf

Diese Arbeit folgt dem Grundansatz mit dem Filterrad-Prinzip. Die Beschränkung dieses Prinzips liegt in der niedrigen Anzahl der spektralen Kanäle, wie in [4] und [5]. Ein gängiger Lösungsansatz ist, die Größe des Filterrads auszuweiten, um eine erhöhte Anzahl der Filterpositionen zu schaffen, aber dies führt zur Ausdehnung des gesamten Systemaufbaus und Vergrößerung der Aufnahmezeit. Angesichts dieser Probleme wurde das Systemkonzept in Abbildung 1 entwickelt, welches aus zwei Filterradkameras C_1 und C_2 und einem DLP-Projektor für Strei-

fenprojektion besteht. Dabei werden das aktive stereoskopische 3D-Verfahren und die filterradbasierte multispektrale Bildaufnahme in einem System vereint. Der Stereo-Filterradkamera-Aufbau ermöglicht es, dass sich der gesamte Spektralbereich $[\lambda_1,...,\lambda_n]$ in zwei Teile $[\lambda_1,...,\lambda_m]$ und $[\lambda_{m+1},...,\lambda_n]$ aufteilen lässt. Diese zwei Unterbereiche werden von beiden Kameras simultan erfasst, wobei die Aufnahmezeit mit der Erhöhung der Anzahl der spektralen Kanäle unverändert bleibt. Die Streifenprojektion in einem Referenzkanal λ_{ref} gewährleistet eine robuste Korrelation zwischen Bildkoordinaten beider Kameras, was sowohl eine hohe 3D-Messgenauigkeit als auch eine fehlerfreie Fusion von Stero-Multispektralbildern ermöglicht.

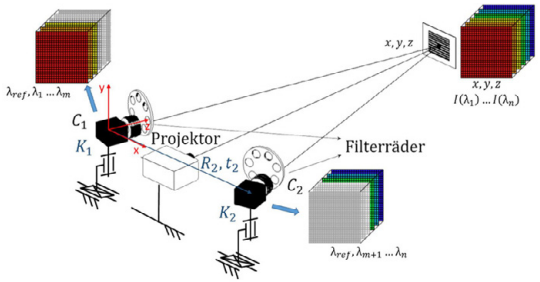


Abbildung 1: Systemkonzept.

Der experimentelle Systemaufbau ist in Abbildung 2 angezeigt. Die Grundelemente sind zwei Filterradkameras „Smart Spectral Imager 2.0“ [7] jeweils mit zwölf Filterpositionen und einer maximalen Drehgeschwindigkeit von fünf Umdrehungen pro Sekunde. Die linke Kamera wird mit zwölf Filtern zwischen 400 nm und 950 nm in Schritt von 50 nm bestückt, und die rechte Kamera enthält neben dem Filter von 550 nm als Referenzkanal für die 3D-Rekonstruktion noch drei Filter von 975 nm, 1000 nm und 1050 nm. Insgesamt sind 15 spektrale Kanäle in diesem System enthalten, wobei eine höchste Kapazität von 23 Kanälen vorliegt. Zur Beleuchtung für die multispektrale Bildaufnahme dienen drei Halogen-Lampen.

Die Filterradkamera verfügt über eine Autofokus-Funktion mittels Sensorverstellung zur Schärfekorrektur. Im Vergleich zu dem Bildrestaurationsverfahren in [8] mit bester Performanz unter PSF-basierten

Ansätzen weist diese mechanische Korrektur folgende Vorteile auf. Zuerst ist damit ein höherer Schärfegrad bei verstärkter Unschärfe hinsichtlich des großen Spektralbereiches zu erzielen, welche Schwierigkeiten bei der Umsetzung des Kalibrieralgorithmus in [8] durch erschwerte Detektion der Referenzmarken verursacht und sich nur bis zum bestimmten Maß numerisch kompensieren lässt. Zweitens wird die irreguläre Bildverzerrung aufgrund der ortsabhängigen Freiform-PSF in [8] vermieden, damit ist eine höhere Genauigkeit bei der Kompensation des Transversalfehlers zu erreichen.

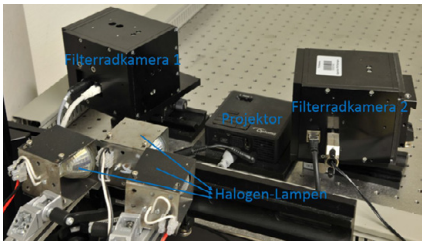


Abbildung 2: Systemaufbau.

2.2 Funktionsprinzip und Ablaufplan

In Abbildung 3 und Abbildung 4 werden das Funktionsprinzip und der Ablaufplan der multispektralen 3D-Bildaufnahme illustriert. Die Streifenprojektion und die multispektrale Bildaufnahme unter Halogenlampen-Beleuchtung erfolgen sequentiell. Für die 3D-Erfassung werden die aperiodischen sinusförmigen Streifenmuster projiziert, denn mit diesen Mustern ist eine Reduktion der 3D-Messzeit durch Senkung der Musteranzahl im Vergleich zu dem Phasenschiebe-Verfahren zu erzielen [9]. Die Auswertung der Streifenbildsequenzen für die Korrespondenzanalyse zwischen Stereo-Kameras erfolgt dann durch die intensitätswertbasierte Kreuzkorrelation:

$$r = \frac{\sum_{i=1}^{N} (I_{1,i} - \overline{I_1})(I_{2,i} - \overline{I_2})}{\sqrt{\sum_{i=1}^{N} (I_{1,i} - \overline{I_1})^2} \cdot \sqrt{\sum_{i=1}^{N} (I_{2,i} - \overline{I_2})^2}}, \tag{1}$$

wobei N die Anzahl der projizierten Muster ist, $I_{k,i}$ der Intensitätswert an dem Bildpunkt (u, v) auf dem i-ten Bild der k-ten Kamera, und $\overline{I}_k$ der mittlere Intensitätswert dieses Bildpunktes. Die Erzielung der Subpixelgenauigkeit an der rechten Kamera erfolgt durch Maximierung des Korrelationskoeffizientes r mit Interpolation des Intensitätswertes in jedem einzelnen Bild der rechten Bildsequenz.

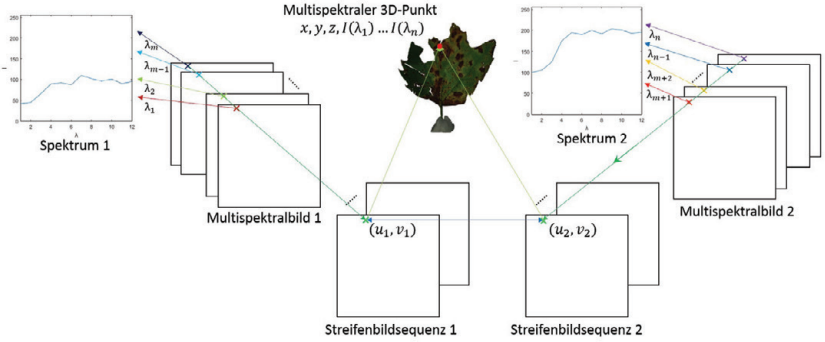


Abbildung 3: Funktionsprinzip des Systems.

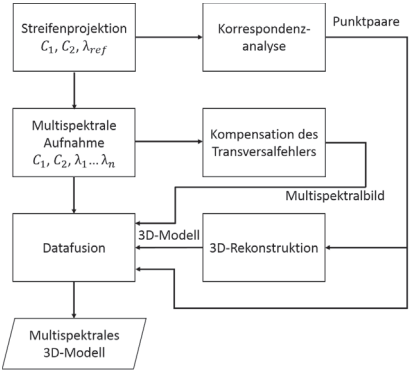


Abbildung 4: Ablaufplan.

Bei der Datenfusion werden zunächst die Bildpunkte im Referenzkanal λ_{ref} durch Kompensation der relativen Transversalfehler in beiden

Filterradkameras jeweils mit einem Unterteil der spektralen Daten verknüpft. Anhand der Bestimmung der korrespondierenden Bildpunktpaare mittels Streifenmuster werden diese beiden Teile des Spektrums weiter mit den zugehörigen 3D-Punkten kombiniert, so dass jeder 3D-Punkt über das vollständige Spektrum $I(\lambda_1) \ldots I(\lambda_n)$ verfügt. In den Punktpaaren wird die Subpixelgenauigkeit an der rechten Kamera erzielt, weshalb die Spektralwerte aus dem rechten Multispektralbild interpoliert werden.

3 Kompensation des Abbildungsfehlers und Systemkalibrierung

3.1 Kompensation des Transversalfehlers

Der relative Abbildungsfehler in der Filterradkamera lässt sich in eine Längs- und eine Querkomponente zerlegen. Da die Unschärfe wegen Längsfehlers durch die Sensorverstellung eliminiert wird, bleibt nur der transversale Teil des Fehlers numerisch zu kompensieren.

Im verzeichnungsfreien Fall setzt sich der Transversalfehler aus der chromatischen Aberration und der Filteraberration zusammen. Die Analyse in einer vorherigen Arbeit des Autors [10] weist nach, dass die Filteraberration mit dem variablen Sensor-Apertur-Abstand die Kombination einer Skalierung mit einer Translation ist. Zur Modellierung der chromatischen Aberration stehen verschiedene Ansätze zur Verfügung. Typischerweise gibt es dazu das affine Modell [11] und das Radial-Tangential-Modell in [12]. Durch Kombination dieser Modelle mit der Filteraberration ergeben sich dann ein affines Korrekturmodell und folgendes Polynom-Modell mit p_1 bis p_6:

$$\begin{pmatrix} x_d \\ y_d \end{pmatrix} = \begin{pmatrix} p_1 + p_2 x_s + p_3 r^2 x_s + p_4(3{x_s}^2 + {y_s}^2) + 2p_5 x_s y_s \\ p_6 + p_2 y_s + p_3 r^2 y_s + 2p_4 x_s y_s + p_5({x_s}^2 + 3{y_s}^2) \end{pmatrix}, \quad (2)$$

wobei (x_s, y_s) und (x_d, y_d) die Koordinaten des ursprünglichen und verzerrten Bildpunktes bezüglich des Bildhauptpunktes sind, und r der Abstand des ursprünglichen Bildpunktes zum Hauptpunkt ist. Unter Betrachtung der Wellenlängenabhängigkeit der Objektivverzeichnung kann eine kanalweisen Verzeichnungskorrektur mit dem Verfahren in [13] als Vorverarbeitung umgesetzt werden.

Die Modellauswahl soll an die Systemkonfiguration angepasst sein. Ein zu schlichtes Modell kann den Fehler nicht mit ausreichender Genauigkeit beschreiben, demgegenüber besteht aber das Risiko bei einem Modell mit exzessiver Komplexität, dass in dem Modell-Fitting Instabilität auftreten kann. Die Untersuchung in [14] zeigt, dass der Transversalfehler mit dem affinen Korrekturmodell schon sehr genau beschrieben werden kann. Das Polynom-Modell und die kanalweise Verzeichnungskorrektur trägt nur geringfügig zur Erhöhung der Genauigkeit der Fehlerkompensation bei, jedoch benötigen einen hohen Rechenaufwand. Aus diesem Grund wird in dieser Arbeit das affine Modell implementiert, und die Wellenlängenabhängigkeit der Verzeichnung wird in der Fehlermodellierung vernachlässigt.

3.2 Systemkalibrierung

Der gesamte Kalibrierprozess lässt sich lediglich mit einem Schachbrett-Kalibriertarget durchführen. Für die 3D-Kalibrierung wird die Methode in [15] mit zusätzlichen tangentialen Komponenten im Verzeichnungsmodell verwendet. Das Kalibriertarget wird in verschiedenen Positionen und Orientierungen bei allen spektralen Kanälen aufgenommen. Im Referenzkanal werden die Stereoparameter für die 3D-Rekonstruktion ermittelt. Mit diesen Aufnahmen werden bei den restlichen Kanälen zugleich die Koeffizienten im affinen Korrekturmodell für die Kompensation des Transversalfehlers bestimmt.

4 Evaluation und Demonstration des Systems

Zur Systemdemonstration wurde ein Programm im Matlab-Framework erstellt, wobei die Eckpunktextraktion aus Target-Aufnahmen und die 3D-Kalibrierung mithilfe der Computer Vision System Toolbox realisiert werden. Zur Kompensation der spektralen Ungleichmäßigkeit in der Beleuchtung und der Empfindlichkeitskurve des Bildsensors wird ein Intensitätsausgleich zwischen allen spektralen Kanälen durch eine kanalweise Anpassung der Integrationszeit umgesetzt, wobei die Anpassungskoeffizienten mit einem Spektralon ermittelt wurden.

4.1 Evaluation der 3D-Formerfassung

Für die Evaluation der 3D-Messgenauigkeit wurde die Aufnahme des Prisma-Prüfnormals in Abbildung 5 ausgewertet, wobei die Güte der Flächenrekonstruktion aus 3D-Punkten bewertet wird. Bei dem Fitting von Flächen F1 bis F5 ergibt sich eine mittlere Standardabweichung von $22\,\mu m$. Mit dieser Kennzahl lässt sich eine gute 3D-Messstabilität prinzipiell bestätigen.

Anhand des Prüfnormals wird noch ein Skalierungsfaktor zur 3D-Koordinatenkorrektur bestimmt, wobei die Summe von den Längen L1 bis L4 überprüft wird. Das Verhältnis von deren Sollwert zu dem aus der Punktwolke berechneten Istwert wird als Skalierungsfaktor ermittelt und für weitere 3D-Aufnahmen implementiert.

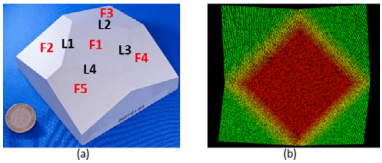


Abbildung 5: Prüfnormal und 3D-Aufnahme.

4.2 Evaluation der Fehlerkompensation

Die Kompensation des Transversalfehlers wurde mit weiteren Target-Aufnahmen evaluiert, wobei der mittlere Restfehler nach der Bildkorrektur anhand der Eckpunkte ermittelt wurde. Die Ergebnisse in Abbildung 6 zeigen, dass sich der Restfehler mit Erhöhung der Wellenlänge vergrößert. Dieser Effekt wird durch die verstärkte Unschärfe des Bildes verursacht, welche auf der Sensorik beruht, weil Photonen mit längerer Wellenlänge größere Eindringtiefe haben und in der unteren Schicht in die Nachbarzellen diffundieren können.

In der Evaluation wird prinzipiell eine hohe Genauigkeit aufgewiesen. Bei der schlechtesten Performanz bei dem Kanal von $1050\,nm$ beträgt der Restfehler ca. 0,1 Pixel, welcher ca. $12\,\mu m$ in der Objektebene entspricht. Ein Mittelwert von 0,06 Pixeln zwischen allen Kanälen wird erreicht. Mit diesen Kennwerten lässt sich der Algorithmus der Bildkorrektur verifizieren.

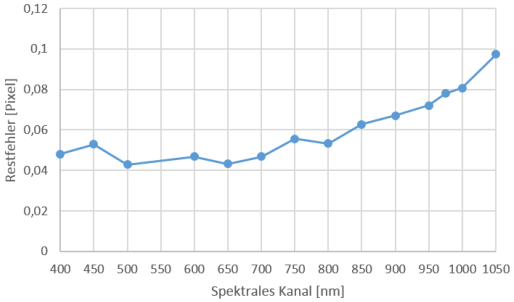


Abbildung 6: Restfehler nach der Kompensation des Transversalfehlers.

4.3 Multispektrale 3D-Aufnahme

Zur Demonstration des Systems wurde eine Aufnahme eines Ahornblattes durchgeführt. Das multispektrale 3D-Punktwolkenmodell ist in Abbildung 7 gezeigt. Die Datenstruktur dafür wird analog zu dem ply-Format gestaltet, in der die 3D-Punkte neben RGB-Werten noch weitere Spektralwerte besitzen. Aus den Daten im VIS-Bereich lässt sich ein RGB-Modell im ply-Format exportieren. Zur Visualisierung des 3D-Modells in einem einzelnen spektralen Kanal werden die Spektralwerte in alle drei Farbkanäle des ply-Formats geschrieben, um einen Grauwert-Effekt zu erzeugen.

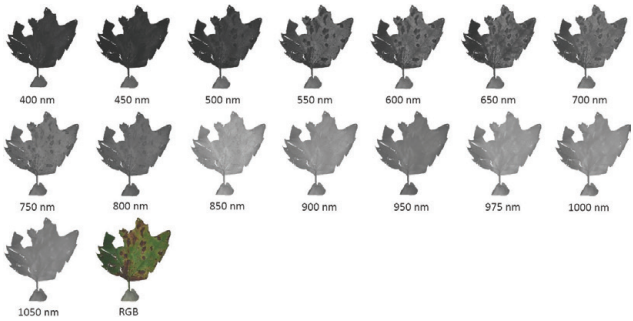


Abbildung 7: Multispektrales 3D-Ahornblattmodell.

5 Zusammenfassung und Ausblick

In dieser Arbeit wird ein Stereo-Filterradkamera-System zur multispektralen 3D-Bildaufnahme vorgestellt. Mit dem Ansatz zur Bilddatenfusion wird eine effiziente, robuste und präzise multimodale Bilderfassung realisiert. Durch die Evaluation wird eine hohe Genauigkeit bei der Kompensation des Abbildungsfehlers der Filterradkamera nachgewiesen, womit eine zuverlässige Fusion von 3D- und multispektralen Bilddaten prinzipiell gewährleistet werden kann. In der zukünftigen Arbeit ist ein radiometrischer Ausgleich zwischen Stereo-Kameras durchzuführen, wobei der Einfluss des Betrachtungswinkels durch eine 3D-basierte Intensitätskorrektur zu kompensieren ist.

Danksagung

Diese Arbeit ist aus dem Graduiertenforschungsprojekt „Beitrag zur hyperspektralen 3D-Oberflächenerfassung und -verarbeitung für die industrielle Bildverarbeitung“ mit dem Förderkennzeichen 03ZZ0410 entstanden. Das Projekt wird vom Bundesministerium für Bildung und Forschung (BMBF) Deutschlands im Rahmen der Innovationsallianz 3Dsensation gefördert. Die Autoren danken dem Fördermittelgeber.

Literatur

1. M. H. Kim, T. A. Harvey, D. S. Kittle, H. Rushmeier, J. Dorsey, R. O. Prum und D. J. Brady, „3D imaging spectroscopy for measuring hyperspectral patterns on solid objects“, *ACM Transactions on Graphics (TOG)*, Vol. 31, Nr. 38, 2012.
2. X. Zhang, A. Zhang und X. Meng, „Automatic fusion of hyperspectral images and laser scans using feature points“, *Journal of Sensors*, Vol. 2015, Nr. 415361, 2015.
3. V. C. Paquit, K. W. Tobin, J. R. Price und F. Mèriaudeau, „3D and multispectral imaging for subcutaneous veins detection“, *Optics Express*, Vol. 17, S. 11 360–11 365, 2009.
4. A. Mansori, A. Lathuilière, F. S. Marzani, Y. Voisin und P. Gouton, „Toward a 3D multispectral scanner: An application to multimedia“, *IEEE Multimedia 2007*, Vol. 14, Nr. 1, S. 40–47, 2007.

5. G. Mączkowski, R. Sitnik und J. Krzesłowski, „Data acquisition enhancement in shape and multispectral color measurements of 3D objects“, *Proceedings of the 5th international conference on Image and Signal Processing*, S. 27–35, 2012.
6. M. Rosenberger und G. Linß, „Multispectral image correction for geometric measurements“, *Journal of physics*, Vol. 588, 2015.
7. M. Rosenberger, M. Preißler, R. Fütterer, C. Zhang, R. Celestre und G. Notni, „Development and characterization of a high speed linear moving stage for multispectral measurements“, *IMEKO TC1-TC7-TC13 Joint Symposium 2016*, 2016.
8. J. Brauers, C. Seiler und T. Aach, „Direct PSF estimation using an random noise target“, *Proc. SPIE 7537, Digital Photography VI*, Nr. 75370B, 2010.
9. S. Heist, P. Kühmstedt, A. Tünnermann und G. Notni, „Theoretical considerations on aperiodic sinusoidal fringes in comparison to phase-shifted sinusoidal fringes for high-speed three-dimensional shape measurement“, *Applied Optics*, Vol. 54, Nr. 35, 2015.
10. C. Zhang, M. Rosenberger, A. Breitbarth und G. Notni, „A novel 3D multispectral vision system based on filter wheel cameras“, *2016 IEEE International Conference on Imaging Systems and Techniques*, 2016.
11. J. Brauers, N. Schulte und T. Aach, „Multispectral filter-wheel cameras: geometric distortion model and compensation algorithms“, *IEEE Transactions on Image Processing*, Vol. 17, S. 2368–2380, 2008.
12. J. Klein, J. Brauers und T. Aach, „Spatio-spectral modeling and compensation of transversal chromatic aberrations in multispectral imaging“, *Journal of Imaging Science and Technology*, Vol. 55, Nr. 060502, 2011.
13. C. Bräuer-Burchardt, „A new methodology for determination and correction of lens distortion in 3D measuring systems using fringe projection“, *Pattern Recognition*, Vol. 3663, S. 200–207, 2005.
14. W. Wang, „Untersuchung zur Kalibrierung einer multispektralen Kamera nach dem Filterrad-Prinzip“, Master's Thesis, Technische Universität Ilmenau, 2016.
15. Z. Zhang, „Flexible camera calibration by viewing a plane from unknown orientations“, *Proceedings of IEEE International Conference on Computer Vision*, S. 666–673, 1999.

Compressive shape from focus based on a linear measurement model

Ding Luo[1,2], Thomas Längle[2] and Jürgen Beyerer[1,2]

[1] Karlsruhe Institute of Technology, Vision and Fusion Laboratory, Adenauerring 4, 76131 Karlsruhe
[2] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation, Fraunhoferstraße 1, 76131 Karlsruhe

Abstract Estimation accuracy of conventional shape from focus techniques is strongly coupled with the number of images in the focal stack, limiting the measurement speed. In this article, a novel compressive shape from focus scheme is proposed with an exemplary algorithm based on modified Laplacian operator and principal component analysis. Simulation with synthetic focal stacks have demonstrated comparable results to the conventional method. Test with 6 compressively captured images achieves same level of performance to that of the conventional method with 100 images.

Keywords Shape from focus, modified Laplacian, compressive sensing, PCA.

1 Introduction

Depth estimation based on an imaging system has been a widely studied topic in the area of computer vision and image processing. Generally, existing methods can be classified into active methods and passive methods. Active methods involve the projection of an optical probe unto the target scene, often in the form of laser beam or illumination pattern [1]. The 3D profile of the target scene is reconstructed with the information in the scattering/reflection of the optical probe captured by the imaging system. The requirement of the additional projection/illumination system will increase the complexity and cost of the active

methods, inevitably limiting their applicability. In situations where physical interaction with the scene is not allowed, passive methods are applied by taking images of the scene without additional illumination. Various depth cues in the captured images have been proposed by researchers, including stereopsis [2], shading [3], focus [4], etc., which are used to reconstruct the 3D information. In this paper, the usage of focus as a cue for depth measurement will be studied and discussed.

Research focuses in this area are mainly placed upon two topics, the design of robust focus measure operators and the development of estimation algorithms. Pertuz et al. [5] made an extensive survey and comparison of popular focus measure operators for shape from focus. Apart from the operators listed in the above survey, more complex operators are being developed constantly not only for shape from focus but also for sharpness estimation as a more general topic, such as the S_3 operator by Vu et al. [6], which utilizes both spatial and spectral information in color images. Conventional estimation algorithms involve finding the maximum focus position from the focal stack for each pixel. A widely accepted method is to take a Gaussian model as proposed by Nayar et al. [5]. Alternatively, other fitting methods have also been studied, such as quadratic and polynomial fits [7]. With the recent development of machine learning and optimization algorithms, more sophisticated methods have been proposed by breaking the isoplanatic restriction [7], such as surface fitting and optimization by neural networks [8], and total variation regularization [9]. It can be seen from the listed literature that the design of focus measure and the development of estimation algorithm are often conducted simultaneously in a holistic manner in order to improve the performance of the overall method.

Unlike shape from defocus techniques where the blur kernel is assumed known, shape from focus techniques generally require a minimum number of image samples along the focal axis in order to perform robust estimation, which are achieved by either shifting the focal plane or changing the relative distance between the camera and the scene. When large numbers of images are required, such shift/movement commonly leads to slow measurement speed and bulky systems. Additionally, the large number of the images, which are needed for evaluation, adds to the data transfer and computational cost. To tackle this problem, a novel shape from focus scheme is proposed and discussed. Although only the algorithm is described in detail and simulated in this

paper, potential hardware implementations will be introduced briefly in the last section.

In Section 2, theoretical background is given with a focus on the linear measurement model, which serves as the core of this new shape from focus (SFF) scheme. Section 3 describes the proposed algorithm with an example and the results from simulation tests are presented in Section 4.

2 Linear measurement model

Various real-world signals can be viewed as an n-dimensional vector $\mathbf{x} \in \mathbb{R}^n$, such as sound, image, etc. In a linear measurement model, each measurement of the target signal is a linear combination of all values in the vector $\mathbf{x}$. The complete measurements of the signal can be written as an m-dimensional vector $\mathbf{y} = A\mathbf{x} \in \mathbb{R}^m$ with an $m \times n$ measurement matrix A.

The ultimate goal of the linear measurement model, like any other measurement systems, is to retrieve the signal $\mathbf{x}$ and the information it is carrying. Formulation of the linear measurement model as a linear system naturally leads to a classical problem of linear algebra: conditions for solving the equation $\mathbf{y} = A\mathbf{x}$. In this context, this question is equivalent to what kind of measurements are needed in order to recover the signal.

Although prevented by classical theory of linear algebra, recent developments in compressive sensing have shown that an underdetermined linear system can be uniquely solved provided sufficient prior knowledge [10]. In the case of compressive sensing, such prior knowledge refers to the assumption of sparsity. However, this is not the only possible prior knowledge. From a more general perspective, the underdetermined linear system with prior information represents a linear manifold learning problem where the prior information acts as the boundary of the manifold to be learned by its low-dimensional projection. The fundamental philosophy behind solutions of such problems is that the information embedded inside the high dimensional manifold is intrinsically of low dimension. In the case of compressive sensing, the unknown manifold is limited to hyperplanes spanned by limited number of axes which correspond to the sparsity assumption.

The significance of the linear measurement model to conventional

SFF approaches is that the number of images required in the focal stack can be effectively compressed if each image can act as a linear combination of all originally required images in the focal stack. The focus measure stack can then be retrieved from the focus measure calculated from the compressed images. It should be noted that this is only possible when the focus measure operator is linear, which is seldom true for modern focus measure operators. Fortunately, most of focus measure operators are composed of several sub-operators, and as long as there is at least one linear sub-operator before all nonlinear sub-operators, the reconstruction can be inserted. In other words, the first sub-operator applied on the compressed images must be linear.

The algorithm for the reconstruction of the focus measure stack depends on the prior knowledge, i. e., the focus measure operator. On one hand, when the focus measure curve has a defined peak, recent compressive sensing algorithms can be incorporated for the recovery of the whole curve. On the other hand, if a training process is allowed or possible focus measure curves can be assumed, conventional methods like principle component analysis (PCA) can be applied in this scheme to yield the measurement/compressing matrix and the reconstruction matrix.

3 Proposed algorithm

To explain this idea in a concrete and clear manner, an exemplary algorithm is presented in this section. The schematic of the algorithm is illustrated in Fig. 1.

The measurement matrix forming the compressed images and the reconstruction matrix for decompression are designed by a training process. In this process, conventional SFF procedures are implemented on a sample focal stack so that the focus measure curve for each pixel is calculated. All the focus measure curves are then assembled, with which PCA is conducted. The largest components are combined to construct the measurement matrix for the compressed images and the reconstruction matrix is simply the transpose of the measurement matrix. The focus measure curve can be reconstructed by multiplying the focus measure of compressed images with the reconstruction matrix:

$$x_R = A^T y = A^T A x\,, \tag{1}$$

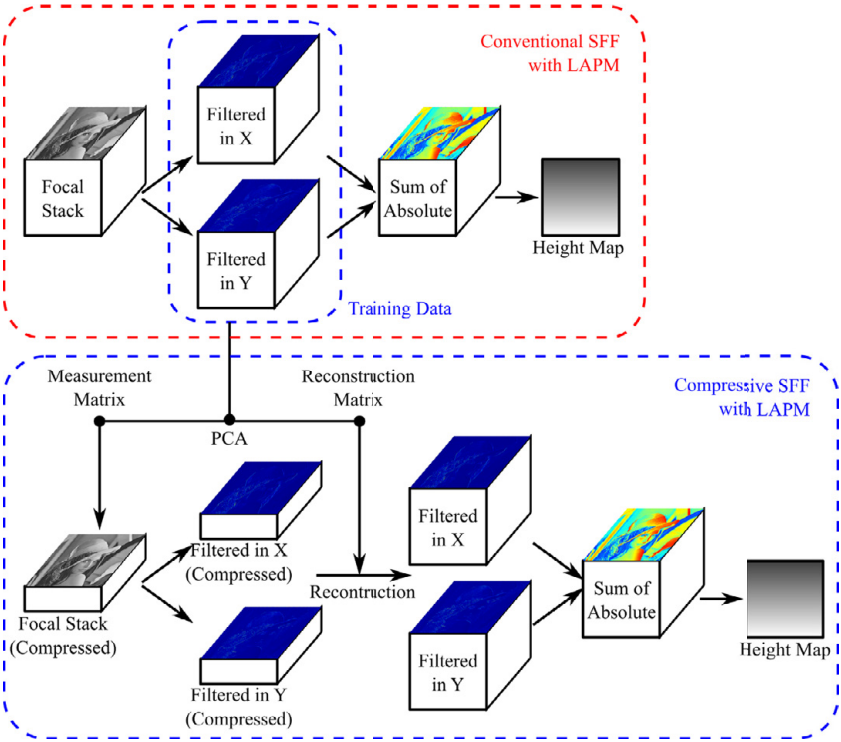


Figure 1: Schematic of compressive SFF with LAPM operator.

where x is the original focus measure curve and A is the measurement matrix for compression.

The widely accepted modified Laplacian operator (LAPM) is selected for calculation of the focus measure [4]. It consists of two sub-operators. Firstly, a one dimensional Laplacian filter is constructed as $f = (-1, 2, -1)$ and used to filter the image in both X and Y directions respectively. Secondly, the absolute value of two filtered images are summed as the final focus measure value. Apparently the 1D filtering operation as a convolution is linear while taking the absolute value is non-linear. Therefore the training and reconstruction step must be inserted before taking the absolute value. From the recovered datacubes

of filtering in X and Y directions, the final focus measure value can be computed through the sum of the two absolute values. Then for each pixel, the axial focus measure curve is smoothed before the maximum value is located to estimate the axial depth.

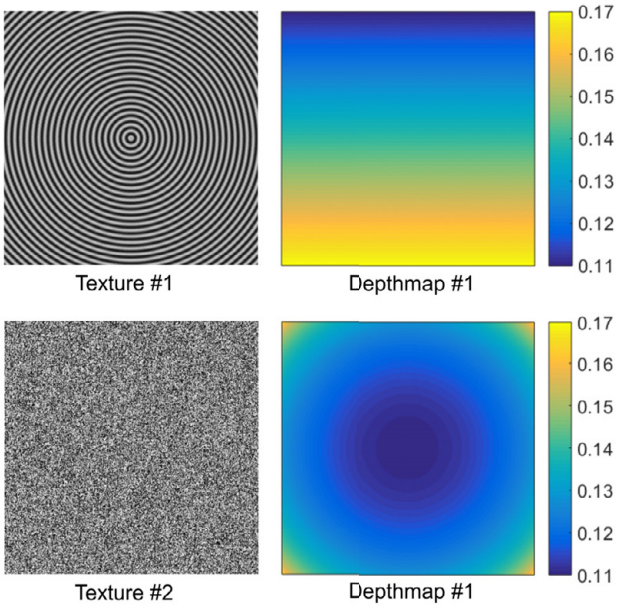


Figure 2: Textures and depthmaps used to sythesize focal stacks.

4 Simulation result

To demonstrate the applicability of the proposed algorithm, simulation is implemented in Matlab with a series of datasets synthetically generated through programs provided by Pertuz et al. in their survey study [5]. The generation of the focal stacks is based on a non-linear, shift variant model of defocus. All estimation results shown in this section are smoothed with a mean filter (window size = 5).

To investigate the influence of training dataset on the compressive SFF (CSFF) result, two texture maps and two depthmaps are combined to form four different datasets. Texture #1 is a structured concentric pattern while texture #2 is a random pattern. Depthmap #1 is a linear ramp and depthmap #2 is part of a sphere. The four datasets #1−#4 are formed with combination of texture and depthmap in the following order: #1 and #1, #1 and #2, #2 and #1, #2 and #2.

Table 1: RMS error showing influence of training set on testing result.

	Set #1	Set #2	Set #3	Set #4
No Training	0.38	0.45	1.48	0.97
Trained #1	NA	0.12	0.36	0.19
Trained #2	0.28	NA	0.90	0.19
Trained #3	1.24	0.56	NA	0.11
Trained #4	2.09	0.42	1.68	NA
Trained All Sets	1.07	0.38	0.39	0.11

Results of CSFF are compared with those of conventional SFF using the root-mean-square (RMS) error with respect to the ground-truth depthmaps, which are listed in Table 1. For the conventional method, a focal stack of 61 images is generated in each case and for the CSFF method the 61 images are compressed into 6 images. The row labeled as no training represents the conventional case whereas the other rows are labeled with their corresponding training set, which is used to generate the measurement matrix and the reconstruction matrix. It can be seen from Table 1 that the choice of training set has an influence on the testing result. In general, the compressive results are comparable to the conventional results but requires much smaller number of compressively captured images. The test result of set #1 with training set #2 and the test result of set #4 with training set #1 are illustrated in Fig. 3.

To investigate the number of images needed for SFF, a series of focal stacks with different numbers of images is synthesized based on texture #1 and depthmap #1 (same combination as dataset #1 used in previous simulations). As expected for the conventional CSS scheme, when the number of images increases, the RMS error decreases, indicating better estimation result. This is due to the fact that the simulated imaging system for image synthesizing has a limited depth of field defined by

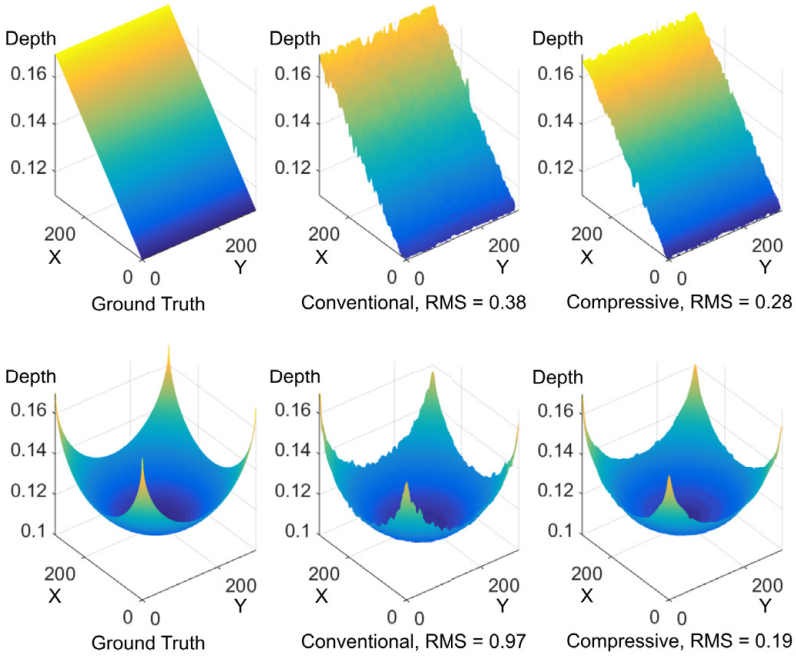


Figure 3: Depth estimation results with conventional SFF and compressive SFF. The upper row illustrates result of set #1 with training set #2 and the lower row illustrates result of set #4 with training set #1.

the blurring kernel. When the step between two adjacent images is too large, the areas with depth in the interval between two focal planes will never get the chance to be imaged sharply and thus cannot be estimated robustly. With the conventional SFF scheme, the minimum number of images needed for robust estimation depends largely on the depth of field of the imaging system, which determines the width of the peak in the focus measure curve when using an operator like LAPM. Generally speaking, when the step size is larger than the width of the focus measure curve, artifacts will start to appear in the estimation result. The dependency of estimation accuracy on the number of images in focal stack is illustrated by the blue solid curve in Fig. 4.

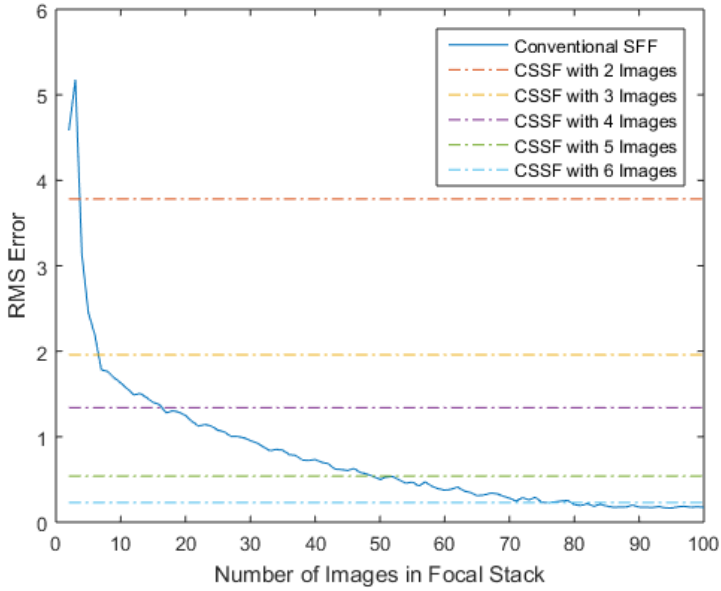


Figure 4: Dependency of estimation accuracy on the number of input images.

On the contrary, in CSFF, regardless of the number of compressive images to be captured, each image acts as a linear combination of all focal planes within of measurement range, and thus contains information from all focal positions in an encoded manner. A training dataset based on texture #1 and depthmap #2 is synthesized with 100 images. The number of compressive images is solely determined by the number of largest principal components to be selected for the construction of the measurement matrix. As shown in Fig. 4, CSFF allows much less images to be captured to achieve same level of estimation accuracy as the conventional method. As the information contained in the largest principal components is related with the rank of the matrix, it is preferred to have a matrix with low rank. This means that the width of the focus measure curve should be larger, i. e., the imaging system should have a larger depth of field. However, as the width gets larger, the relative magnitude of the focus measure peak gets smaller, effectively reducing

the SNR of the measurement. Therefore, a balance must be made between these two factors to generate best estimation performance.

5 Conclusion

In this article, a novel scheme of compressive shape from focus is presented and simulated. Based on linear measurement model, CSFF compressively captures several images, each as a linear combination of all possible focal planes within the measurement range. It has been shown in the simulation that the estimation error of CSFF is comparable to the conventional method using same number of images as the number of images in the training set for CSFF. With datasets synthesized in this article, CSFF with 6 compressive images yields similar performance to the conventional method with a focal stack of 100 images.

Hardware implementation of this scheme could take many different forms. To begin with, auto-focus mechanisms in cameras can be modified so that the focal plane is shifted with a varying speed within one exposure according to the measurement matrix. Customized objective based on liquid lens can be developed to guarantee the speed of focus shifting. On the other hand, if the texture is colorless, hyperchromatic objective can be designed and coupled with tunable illumination to achieve same effect.

References

1. M. Schaffer, M. Grosse, B. Harendt, and R. Kowarschik, "High-speed three-dimensional shape measurements of objects with laser speckles and acousto-optical deflection," *Opt. Lett.*, vol. 36, no. 16, pp. 3097–3099, Aug 2011. [Online]. Available: http://ol.osa.org/abstract.cfm?URI=ol-36-16-3097
2. D. Marr and T. Poggio, "Cooperative computation of stereo disparity," Cambridge, MA, USA, Tech. Rep., 1976.
3. R. Zhang, P.-S. Tsai, J. Cryer, and M. Shah, "Shape-from-shading: a survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 8, pp. 690–706, Aug 1999.

4. S. Nayar and Y. Nakagawa, "Shape from focus," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 16, no. 8, pp. 824–831, Aug 1994.

5. S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," *Pattern Recogn.*, vol. 46, no. 5, pp. 1415–1432, May 2013. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2012.11.011

6. C. Vu, T. Phan, and D. Chandler, "$\mathbf{S}_3$: A spectral and spatial measure of local perceived sharpness in natural images," *Image Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 934–945, March 2012.

7. M. Subbarao and T. Choi, "Accurate recovery of three-dimensional shape from image focus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 3, pp. 266–274, Mar. 1995. [Online]. Available: http://dx.doi.org/10.1109/34.368191

8. M. Asif and T.-S. Choi, "Shape from focus using multilayer feedforward neural networks," *Image Processing, IEEE Transactions on*, vol. 10, no. 11, pp. 1670–1675, Nov 2001.

9. M. Mahmood, "Shape from focus by total variation," in *IVMSP Workshop, 2013 IEEE 11th*, June 2013, pp. 1–4.

10. E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006. [Online]. Available: http://dx.doi.org/10.1002/cpa.20124

Height adaptive shading correction for line-scan stereo imaging based multi-spectral reflectance measurements

Timo Eckhard and Markus Schnitzlein

Chromasens GmbH,
Max-Stromeyer-Str. 116, D78467 Konstanz

Abstract Imaging-based device independent spectral and color measurement can be achieved using a multi-spectral line-scan camera system. However, the measurement is typically limited to flat surfaces, scanned at a fixed working distance to avoid image channel misalignment. A recently proposed method uses the stereo imaging principle to overcome the misalignment problem. We extend this approach by a novel height-adaptive shading correction to account for the dependence of scene illumination on scanning object height with respect to the camera, which deteriorates the spectral and color measurement performance. The correction is based on a fit function obtained from a specifically designed white reference step target. We support our proposal with experimental evaluation, illustrating the performance gain that can be achieved.

Keywords Multi-spectral imaging, line-scan imaging, stereo imaging.

1 Introduction

Imaging based reflectance measurement is a major challenge in various domains of the producing industry. Conventional reflectance measurement devices, such as spectrophotometers, only permit point measurements and cannot be used for in-line or 100 % inspection. Also, highly textured surfaces impose limitations to point measurements, as the readings of such devices typically correspond to an average region of a rather large measurement aperture. In [1], a line-scanning based

multi-spectral imaging system for device independent reflectance and color measurement was proposed, aiming to overcome the aforementioned limitations. Initially, measurement was limited to planar scanning surfaces due to channel misalignment that results from non-planar objects, degrading the measurement performance of the system significantly. A solution to overcome the channel misalignment problem was proposed in [2, 3], where the authors made use of the stereo imaging principle [4] to estimate image disparity in order to correct the misalignment.

To achieve norm-conform spectral reflectance measurement in the scope of stereo line-scan multi-spectral imaging, an additional processing step is required. Essentially, the non-uniformity of scene-illumination has to be corrected adaptively with respect to the relative scene object height on a pixel-by-pixel basis. In this work, we propose a novel calibration routine to achieve this required shading correction. For that, image data of a specifically designed step target with reference white material and known reflectance at 10 height levels is acquired. A 3D surface is then fitted to the data to model the scene-illumination non-uniformity as a function of pixel location and relative scene object height. At last, the surface function is discretized to obtain a lookup table that can be used for scene-adaptive shading correction with low computational cost.

In what follows, we introduce the stereoscopic line-scanning based multi-spectral imaging principle and the data processing framework used. In Section 2, we explain the method proposed for height-adaptive shading correction and illustrate the calibration target and process that were developed for the given task. Results of the experimental evaluation of the proposed approach are shown in Section 3. At last, we summarize and conclude the work in Section 4.

1.1 Stereoscopic line-scanning based multi-spectral imaging principle

A line-scan sensor contains a single sensor line, or various sensor lines with additional color filters (i. e., red, green and blue color filters). In line-scan imaging, either camera or object has to be moved with respect to the other. Synchronously with the movement, image lines are ac-

quired successively and concatenated to obtain a 2D image. By placing multiple line-scan cameras with additional color filters in front of the lenses next to each other with aligned scan lines, the number of spectral image bands can be increased for the object region visible in all camera images. The corresponding imaging principle has been developed and commercialized by Chromasens GmbH in form of a multi-spectral line-scan camera product line called *truePIXA* [5]. On the data processing side, a simple geometrical transformation is required to create a multi-channel image from multiple color filtered RGB images. However, accurate measurement is limited to scanning objects that correspond in shape to the geometrical calibration target used, which is typically flat. Otherwise channel misalignment occurs.

Another application for the parallel arrangement of line-scan cameras is stereoscopic imaging, which is possible due to the inherent difference in perspective for the parallel arrangement of cameras. Chromasens GmbH developed the product line *3DPIXA*, which uses this imaging principle for 3D object measurements [6].

For this work, a prototype camera system has been developed, which consists of a 3DPIXA with additional color filters in front of each lens of the stereo camera system. The imaging principle of this system is illustrated in Figure 1, and a photographic image of a prototype desktop scanner is shown in Figure 2. The optical resolution of this system is $30\,\mu m$ per pixel and the height resolution $6\,\mu m$. The scanning width of the camera system is $215\,mm$, while the length of a scan is freely adjustable. The camera is mounted at 0° with respect to the scan surface and two line-lights are placed at 45° opposed to each other, illuminating the scan line of the camera. The scanning movement is implemented by translating the scan object under the camera using a motorized linear stage.

1.2 Data processing framework

The data processing required for device-independent color and spectral measurement consists of various computational steps, illustrated in form of a block diagram in Figure 3. We adapt a summary of these steps from our previous work [3]:

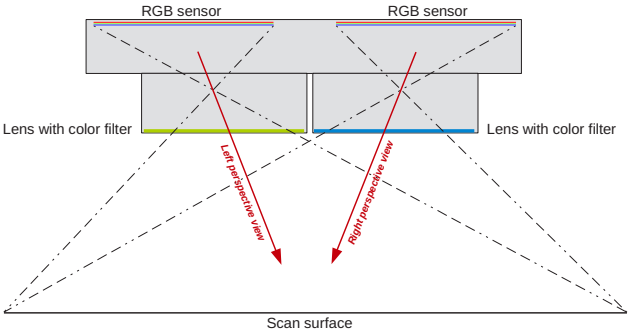


Figure 1: Line-scan multi-spectral stereo imaging principle. Illustration adapted from [3].

- **Image rectification:** image rectification is a spatial image deformation process required to correct distortion and ensure that pixels corresponding to the same physical location in the scene from right and left stereo image lie in the same image row. This condition reduces the correspondence problem from a search in two spatial image dimensions to a search only along the image dimension corresponding to the scanning line, which makes determining 3-dimensional structure very efficient. The model used in this work is based on polynomial functions.

- **Feature image computation:** refers to the process of computing grayscale images from color filtered RGB stereo images. The feature image transformation [1, 2] accounts for the color filtering and is required for correlation based stereo imaging to simplify the stereo matching problem. In this work, we used an information preserving transformation model that was created empirically from the data of the color calibration chart. The transformation of the left image is given as a weighted sum of R, G and B value of each image pixel and the weight is determined as the first principal component from calibration data. In a similar manner, the grayscale version of the right image is computed, however, using weights that minimize differences between left and right image [1].

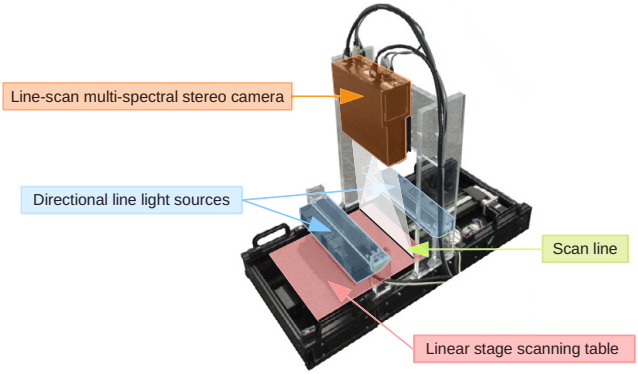


Figure 2: Prototype stereo line-scan multi-spectral camera system. Illustration adapted from [3].

- **Disparity image computation:** a disparity image encodes the pixel-wise difference between pixels corresponding to the same physical location in left and right image. The disparity measure is proportional to the relative depth of the image scene, i. e., the distance between camera system and scene objects.

  To find image correspondence from rectified images, we use block-matching based normalized cross-correlation, a well-established method from computation stereo [4]. A polynomial interpolation is used to achieve sub-pixel accuracy.

- **Adaptive shading correction:** is discussed in Section 2.

- **Image registration:** corresponding scene object points from right and left image of a stereo system are displaced by the amount of pixel disparity. For spectral estimation based color measurement, image channels from both left and right image need to be combined to a single, pixel-wise registered multi-channel image. Using image disparity, left and right image can be registered with sub-pixel accuracy to a 6-channel multi-spectral image cube. The required image resampling is performed using a shape-preserving piecewise cubic interpolation [7].

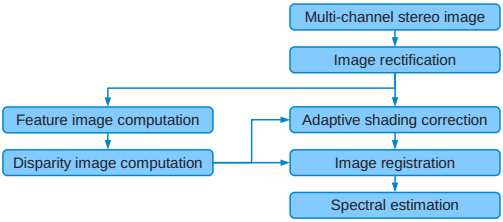


Figure 3: Data processing pipeline for 3D spectral imaging.

- **Spectral estimation:** using a regression based model, spectral reflectance factor data can be estimated from camera responses of the 6-channel image on a pixel-by-pixel basis. The model is created using camera response data of a predefined color calibration chart with 660 color patches and corresponding spectral reflectance factors. This data was measured using a Konica Minolta FD7 spectrophotometer in accordance with ISO13655:2009-M2 norm [8].

2 Height-adaptive shading correction

In conventional line-scan imaging, shading correction is the process of accounting for lens vignetting, non-homogeneous scene illumination along the scan line, and sensor photo response non-uniformity. From a simplistic point of view, shading correction is the process that makes the image data of a spatially homogeneous scan target appear spatially homogeneous in the image data.

An adequate correction for the aforementioned effects is achieved by applying a pixel-wise scaling factor, channel-wise line-by-line to the entire image. In case of planar scanning objects, the correction factor can be computed from image data of a spatially homogeneous reference white sheet, scanned at working distance. If non-planar objects are scanned, the correction factor is a function of scene object height and pixel position.

The consequence of not applying shading correction as a preprocessing step to device independent spectral and color measurement is a bias in the measurement results and resulting from that, a measurement error.

2.1 Computational model

Let $I(x, y, c)$ be the pixel at location (x, y) of the c-th image channel of a spatially registered multi-spectral image cube. $D(x, y)$ denotes image disparity at pixel location (x, y) and let $f(x, D(x, y), c)$ be the height-adaptive shading correction function. Note that f is a function of image disparity D and only x-coordinate, not y-coordinate. This is specific to line sensors, for which the correction factor is only determined by the sensor pixel location and the scene object disparity.

A height-adaptive shading corrected pixel intensity $I'(x, y, c)$ is obtained by the following function

$$I'(x, y, c) = I(x, y, c)f(x, D(x, y), c). \tag{1}$$

The shading correction function f can be modeled empirically using camera response data from scanned image of a white reference step target. Using this target, scaling factors at discrete height steps over the entire *field-of-view* of the camera can be extracted.

Our target contains 10 steps over a height range of approximately $20\,mm$ with a width corresponding to the *field-of-view* of the camera, and is produced with spot-face milled aluminum and sandblasted surface finishing. A photographic illustration of the target is shown in Figure 4. Each step consists of a spatially homogeneous reference white stripe region, and an uncovered region of the sandblasted aluminum surface.

From the white region, the scaling factor is extracted. From the uncovered region, exact image disparity at every pixel location can be computed using the stereo imaging approach. The uncovered region is required, as the homogeneous white reference stripe does not contain enough texture for correlation based stereo imaging, and we did not want to make assumptions on the exact geometrical shape of the target due tolerances in the fabrication process.

In Figure 5 (left), an example of the empirical data extracted from the step target is shown for the red channel of the left stereo image. The discrete data from the step target is approximated with a two dimensional 4th degree polynomial surface function. An example of a resulting surface for the red channel of the left image is shown in Figure 5 (right).

Figure 4: White reference step target.

When applying height adaptive shading correction, f has to be evaluated for every channel and pixel in the image with $D(x, y)$. To reduce computational complexity of this operation, lookup tables are used instead of evaluating f.

3 Experimental evaluation

To experimentally evaluate the performance of the proposed height adaptive shading correction approach, we considered the application of device independent spectral and colorimetric measurement. Reference spectral reflectance data and CIE-L*a*b* color coordinates of a printed color chart with 660 patches were measured using a Konica Minolta FD7 handspectrophotometer. This chart was then scanned at 5 equally spaced height steps in the range of $-1\,cm$ to $+1\,cm$ around the nominal working distance of the camera, using our prototype scanner. Using the data processing pipeline described in Section 1.2, we obtained spectral reflectance and CIE-L*a*b* color coordinates for every pixel of the scanned images. Pixel-wise data corresponding to the measurement aperture of the spectrophotometer at each color patch was extracted and averaged in an automated fashion.

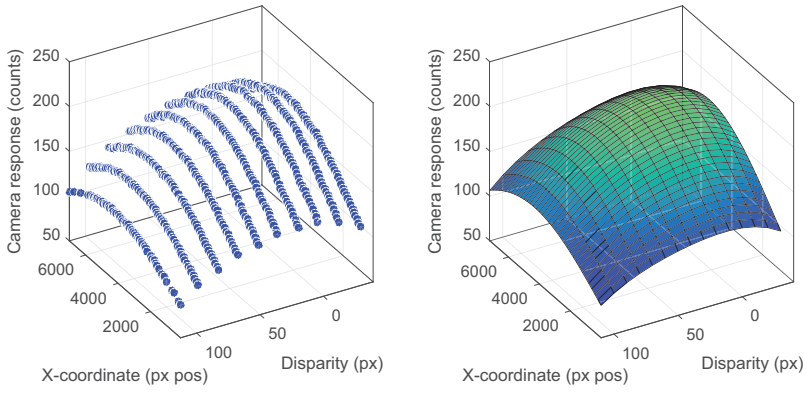


Figure 5: Discrete camera response data extracted from the red channel of the left image of the step target reference white (left); Corresponding surface fit (right).

The reference data was then compared with the data from the prototype scanner. Differences were quantified by spectral root mean square error (*RMSE*) and color difference in accordance with the latest recommendation of the International Commission on Illumination (CIE), using the CIEDE2000 color-difference formula (ΔE_{00}) [9].

We considered the case of measurement without and with height adaptive shading correction. First order statistics of the numerical results are presented in Table 1 and 2. From Table 1, corresponding to results obtained without applying height adaptive shading correction, we can see that both spectral and colorimetric error increase with the height at which the color target was scanned. Even for small height deviation, the range of errors is within the range of noticeable difference.

When height adaptive shading correction is applied (see Table 2), spectral and colorimetric errors do not increase with height. There is still considerable residual error, which can partly be explained by the underdetermined nature of the mapping performed to estimate high-dimensional spectral reflectance data from low-dimensional multispectral image data. As shown in previous studies, using a system with a larger number of image channels can improve this situation [1].

At last, we would like to point out that this evaluation only considered the case of a varying height of a flat scanning surface. In this case, the surface normal was still aligned with the optical axis of the camera system (0°) and the light source at approximately 45°. The effect of deviation of these geometrical conditions on spectral and color measurement was not considered yet.

Table 1: Spectral and color estimation performance ***without*** height-adaptive shading correction.

	RMSE				ΔE_{00}			
	Avg.	**Std.**	**Min.**	**Max.**	**Avg.**	**Std.**	**Min.**	**Max.**
Height 1	0.0163	0.0092	0.0033	0.0589	1.08	1.140	0.06	8.77
Height 2	0.0614	0.0399	0.0061	0.1878	3.26	0.806	1.02	7.21
Height 3	0.0954	0.0632	0.0035	0.3156	4.94	1.17	1.84	7.98
Height 4	0.0846	0.0556	0.0033	0.2965	4.49	1.08	1.52	7.98
Height 5	0.0962	0.1773	0.0042	1.1314	5.81	10.46	0.71	76.70

Table 2: Spectral and color estimation performance ***with*** height-adaptive shading correction.

	RMSE				ΔE_{00}			
	Avg.	**Std.**	**Min.**	**Max.**	**Avg.**	**Std.**	**Min.**	**Max.**
Height 0	0.0163	0.0092	0.0033	0.0589	1.08	1.14	0.06	8.77
Height 1	0.0178	0.0100	0.0032	0.0618	1.00	0.76	0.10	6.90
Height 2	0.0172	0.0097	0.0031	0.0594	0.99	0.87	0.14	11.69
Height 3	0.0171	0.0095	0.0026	0.0592	1.03	0.86	0.08	8.07
Height 4	0.0170	0.0094	0.0033	0.0593	1.06	0.86	0.07	5.80

4 Summary and conclusions

We used a previously proposed modified line-scanning based stereo imaging system for multi-spectral imaging and device independent spectral and colorimetric measurements. The measurement principle incorporated in this system allows for simultaneous 3D and spectral image acquisition. To account for the scene illumination dependence on scene object height, a height-adaptive shading approach was proposed.

The method developed is based on fitting a correction function to empirical data extracted from a reference white step target. To reduce computational load when applying the correction, lookup tables were used. Experimental evaluation of device independent spectral and color measurement performance showed that height adaptive shading correction is a necessary requirement to avoid measurement bias for 3D measurement objects.

References

1. T. Eckhard, "Design considerations of line-scan multi-spectral imaging systems – Application in spectral and color measurement," dissertation, University of Granada, 2015.
2. T. Eckhard, J. Eckhard, E. M. Valero, and J. Hernández-Andrés, "Scene-adaptive registration of line-scan multi-spectral image data for non-planar scanning objects," in *Color and Imaging Conference*, vol. 2015, no. 1. Society for Imaging Science and Technology, 2015, pp. 46–51.
3. T. Eckhard and M. Schnitzlein, "Stereo line-scan imaging based multi-spectral color measurement on non-flat scanning objects," in *Farbworkshop 2016*. ZBS Zentrum für Bild- und Signalverarbeitung e. V. Ilmenau, 2016.
4. M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 25, no. 8, pp. 993–1008, 2003.
5. Chromasens. (2016) Multi-spectral line-scan camera truePIXA. [Online]. Available: http://www.chromasens.de/en/multi-spectral-camera-truepixa
6. ——. (2016) Line-scan camera 3DPIXA. [Online]. Available: http://www.chromasens.de/en/3d-line-scan-camera-3dpixa
7. F. N. Fritsch and R. E. Carlson, "Monotone piecewise cubic interpolation," *SIAM Journal on Numerical Analysis*, vol. 17, no. 2, pp. 238–246, 1980.
8. DIN-Norm, "ISO 13655: 2009 Graphische Technik – Spektrale Messung und farbmetrische Berechnung für graphische Objekte."
9. C. I. de lÉclaraige, "Industrial colour-difference evaluation," 1995.

Line-scan stereo for 3D ground reconstruction

Doris Antensteiner, Bernhard Blaschitz, Clemens Eisserer, Reinhold Huber-Mörk, Johannes Ruisz, Svorad Štolc, Kristián Valentín

AIT Austrian Institute of Technology GmbH
Digital Safety & Security Department
Donau-City-Straße 1, Vienna, Austria

Abstract We present an image acquisition setup and data processing pipeline for 3D ground reconstruction used in applications for road surface inspection. To obtain a larger overlapping surface region the optical axes of the sensors are verged. We describe a calibration procedure correcting various distortions inherent to this specific optical setup. We employed the so called Stochastic Binary Local Descriptor (STABLE) descriptor capable of dense stereo matching for high-performance vision. Additionally, we describe an optimization problem formulation where the cost from STABLE is forming a data term which is extended by adding a regularization term incorporating prior knowledge. Energy minimization is solved by a primal optimization approach and enhances details significantly when compared to a purely data-dependent solution. The accuracy of the depth reconstruction method is assessed making use of a synthetic ground truth. Furthermore, we also present results of processed real-world data from road surface measurements.

Keywords Linescan stereo camera, synthetic road surface, stereo matching, 3D ground reconstruction, STABLE descriptor.

1 Introduction

Line-scanning is a popular method to acquire images of moving objects, especially in machine vision applications. From moving platforms, like air- or spaceborne scanners, the so called pushbroom principle is used to acquire sensor lines while moving along a predefined path [1]. We utilize this acquisition principle, extended to binocular stereo, for an

application in ground reconstruction from a vehicular platform. The application area is the inspection of road surface conditions. We will describe how to obtain depth information from stereo pairs, e. g., pairs of images taken concurrently from slightly displaced positions.

In stereo imaging the range for each pixel is obtained from the estimated disparity, i. e., the displacement between corresponding points observed in two (or more) images. The epipolar constraint in a stereo vision system states that a point in one image is found along the corresponding epipolar line in the other image. Epipolar rectification in area-scan stereo pairs aligns epipolar lines to images lines, thus reducing the correspondence estimation to a search oriented along an expected disparity range in image lines. In a line-scan stereo system one mechanically adjusts this geometrical constraints such that epipolar lines correspond to sensor lines. Estimation of disparities is then performed along sensor lines. The STochAstic Binary Local dEscriptor (STABLE) for disparity estimation is discussed in [2].

To incorporate prior knowledge, i. e., a smoothness assumption, and to cope with missing data we employed an optimization problem formulation where the cost from STABLE is forming our data term which is extended by adding a regularization term based on total variation [3]. An energy minimization procedure is iteratively solved by a primal optimization approach. This enhances details significantly when compared to a purely data-dependent solution.

The rest of this paper is organized as follows. Sec. 2 presents the image acquisition setup and discusses calibration issues. Sec. 3 shortly describes STABLE and introduces the regularization approach. Quantitative results for synthetic data and illustrative results for road surface data are presented in Sec. 4. Finally, Sec. 5 summarizes and provides and outlook.

2 Image acquisition

We used one or two line-scan stereo cameras which are sensitive in the visible spectrum for acquisitions of the ground surface while the acquisition device is moving. In line-scan stereo the surface is acquired using either one long image sensor line shared by two lenses or two collinearly arranged line-scan image sensors observing the same surface line patch.

Our setup uses two collinearly arranged line-scan sensors observing the surface from two different viewpoints. The optical axes are verged in order to obtain a larger overlapping region.

2.1 Verged stereo geometry

Verging of the optical axis has two drawbacks. First, the pixel resolution decreases from one side of the sensor to another and differently for both views. Secondly, the limited depth of field might result in sharpness reduction depending on optical parameters of the lens employed. Geometric calibration of the sensor lines ensures a constant pixel size at the expected working distance.

The depth of field was estimated to be at the order of magnitude of $\pm 6.16\,mm$ for an f-number of 5.6, a magnification of 0.1 and a sensor pixel size of $10\,\mu m$. This appeared to be suited to compensate for the varying distance due to verging and the expected depth variation. For f-numbers of 1.4 or 2.8 we would obtain a depth of field of $\pm 1.54\,mm$ or $\pm 3.08\,mm$, respectively.

In calibration, the alignment of the sensor lines is required to fulfill the property that the plane spanned by the left optical axis and left sensor line is coplanar with the plane spanned by the right optical axis and right sensor line. This property is important in order to fulfill the epipolar constraint at each depth and requires a calibration procedure which ensures an overlap of the sensor lines at a number of distances.

2.2 Image rectification

Because of the verged optical axes, a perspective distortion *per camera line* is introduced to the images. By taking calibration images of parallel lines orthogonal to the transport direction (Fig. 1) and comparing them to the original, the corresponding perspectivity can be calculated.

In order to estimate the angle α between the optical axis and the ground plane, we used a calibration pattern of equally distributed parallel lines orthogonal to the transport direction, see Fig. 1.

Different real-world lengths x and y are mapped to equal lengths x' and y', namely exactly the left and right half of the image line. This is caused by the natural symmetry of the optical system with its *field of view* (FOV) of the camera lens of 2β, which in our case is approx. $11.7°$.

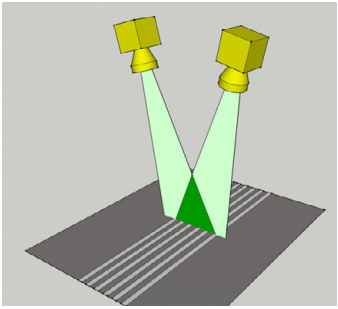
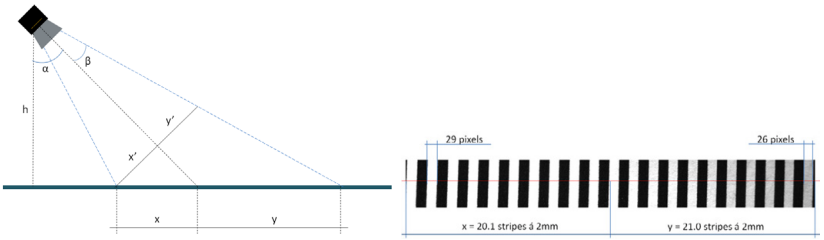


Figure 1: The calibration pattern consists of parallel lines orthogonal to the camera movement. These lines of equal distance and thickness (e. g., $2\,mm$) allow accurate estimation of the distance in real-world coverd by a camera line.



(a) Camera setup (b) Calibration pattern

Figure 2: (a) Schema of the camera setup; real-world distances x and y appear equally sized (x' and y') within the image line, (b) Sample image of calibration pattern taken from an angle of about 12° from the left side.

We can set up the following equations for x and y:

$$\begin{aligned} x &= h\tan(\alpha) - h\tan(\alpha-\beta) \\ y &= h\tan(\alpha+\beta) - h\tan(\alpha) \\ \Rightarrow \frac{x}{y} &= \frac{\tan(\alpha)-\tan(\alpha-\beta)}{\tan(\alpha+\beta)-\tan(\alpha)}\,, \end{aligned} \tag{1}$$

where h is the normal distance of the camera to the ground and β is the half of the FOV. Please note that in Equation (1) height h was cancelled

which means it is not essential for computation of α.

Solving for α we obtain

$$\tan(\alpha) = \left(\frac{y-x}{y+x}\right)\frac{1}{\tan(\beta)}\,. \tag{2}$$

Positive values of α correspond to a setup with the camera on the left and negative values mean camera on the right. Eq. (2) also shows that the *ratio* of x and y is essential, not the absolute values of the real-world dimensions.

As for the example in Fig. 2 we can calculate α with the given values $x = 20.1$, $y = 21.0$ and $\beta = 5.85°$ and get $\alpha = 12.17°$, which is quite an accurate result.

Direct computation of the perspective distortion

It is also possible to compute the perspectivity directly without knowing the angles: Let y' be the homogeneous coordinates of 1D features (cf. Fig. 2(b)) in a scanned image line of a linear pattern y, then $\binom{y'}{1} = H \cdot \binom{y}{1}$ for a homography $H \in \mathbb{R}^{2\times 2}$, which can be computed by solving the overdetermined system

$$-h_{21}y'y + h_{11}y + h_{12} = y' \tag{3}$$

and h_{22} can be assumed $= 1$. The linear undistortion mapping H^{-1} is then applied to every line of the image, equivalently to the well-known 2D case. It is also possible to compute the angle between the cameras of a stereo linescan system using the LP fundamental matrix, see [1].

3 Stereo image processing

In order to obtain depth information from stereo image pairs, corresponding points are typically identified via block matching, i. e., comparison of image patches. Measures of block similarity include direct comparison of pixel intensities using similarity metrics such as the sum of absolute differences (SAD), the sum of squared errors (SSE), the normalized cross-correlation (NCC). Alternatively, comparison is based on measuring distances between feature descriptors such as STABLE.

While for descriptors such as SURF or SIFT some vector metrics in high-dimensional spaces are commonly used to quantify descriptor similarity, for binary descriptors the Hamming distance is typically applied.

3.1 The STABLE descriptor

We consider an image patch $\mathbf{p}$ with a size of $X \times Y$ pixels. The operation β derives the i-th descriptor bit $d_i \in \mathbf{d}$ from patch $\mathbf{p}$ as follows:

$$\beta(\mathbf{p}, i) = \begin{cases} 1 & \text{if } (\mathbf{p} * f_i) > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where f_i is a filter mask of equal size as the image patch $\mathbf{p}$. We refer to the operation β as the *binarized convolution*. The filter dictionary $\mathbf{f}$ contains K sparse filter masks f_i. Each entry in f_i is either 0, 1 or -1. The descriptor $\mathbf{d}$ is a K-dimensional bitmask which is obtained for a given image patch $\mathbf{p}$ using

$$\mathbf{d}(\mathbf{p}) = \sum_{i=1}^{K} 2^{i-1} \beta(\mathbf{p}, i). \tag{5}$$

A set of sparse filter masks from a dictionary are applied to the same image patch and, depending on the number and individual signs of the filter mask entries, a number of pixels is contributing to each descriptor bit.

A more efficient implementation of STABLE, avoiding binarized convolution with K sparse feature filters, uses a single index filter mask $\mathbf{g}$, details are found in [2].

3.2 Stereo matching

In stereo matching we consider a discrete range of disparities $[r_1 \dots r_P]$ for which the descriptors corresponding to each image pixel position are compared using the Hamming distance. This results in a cost stack C of dimension $M \times N \times P$, where M and N are the image dimensions and P is the number of evaluated disparities.

Furthermore, we consider a hierarchical representation of the cost stack, i. e., a Gaussian pyramid is constructed, reducing the stack in the

spatial domain preserving the number of disparity hypotheses. We denote the stack at the i-th pyramid level by $C_i, i = 1 \dots i_{max}$, where C_1 is the stack at original resolution. The pyramid decomposition is used in the regularization of our cost stack.

3.3 Regularization

Regularization of the cost stack C is suggested as a hierarchical and iterative procedure. A regularized solution S is sequentially updated using local neighborhood information, i. e., horizontally and vertically neighboring pixels. The regularization is penalizing large disparity variations in local neighborhoods. This property is propagated through iteration and embedded in a hierarchical framework. The algorithm starts with the cost stack $C_{i_{max}}$ from STABLE matching at the coarsest pyramid level i_{max}. An initial solution for the coarsest pyramid level is given by the minimum of the cost stack at each pixel position (x, y) for the disparities $z = 1 \dots P$ as follows:

$$S^1_{i_{max}}(x, y) = \underset{z=1\dots P}{\arg\min}\, C_{i_{max}}(x, y, z). \tag{6}$$

The solution at the i-th pyramid level and j-th iteration is then iteratively refined:

$$S_i^{j+1}(x, y) = \begin{cases} \underset{z=1\dots P}{\arg\min}\,(C_i(x, y, z) + \lambda L^j(x, y, z)) & \forall (x, y) \in \Omega^{j+1}, \\ S_i^j(x, y) & \text{otherwise,} \end{cases} \tag{7}$$

where $L(\cdot)$ is the regularizer function, λ controls the regularization strength, and Ω^{j+1} is the set of pixel coordinates to be updated in this iteration. Finally, $S_1^{j_{max}}$ is accepted as the regularized solution at the original resolution.

Unlike common procedures in regularization, where the direction of optimization, i. e., along lines or along columns, is changed between adjacent iteration steps we suggest updating pixels in two complementary groups in each step. In order to ensure convergence we suggest a checkerboard-like update pattern which switches between adjacent iterations. The set of pixel coordinates updated in j-th iteration is defined as follows:

$$\Omega^j = \{(x, y) \mid x + y + j \text{ is even}\}. \tag{8}$$

The Pseudo-Huber loss [4] is used to regularize the information originating from matching costs at each pixel position and its neighborhood:

$$L^j(x,y,z) = \sum_{(\hat{x},\hat{y})\in\mathcal{N}_4(x,y)} \delta^2\Big(\sqrt{1+|S^j(\hat{x},\hat{y})-z|^2/\delta^2}-1\Big), \qquad (9)$$

where $(\hat{x},\hat{y})$ are pixel positions within the 4-neighborhood $\mathcal{N}_4(x,y)$. The Pseudo-Huber loss function has the nice property of having linear (i. e., total variance) behavior towards large values, while being quadratic around zero. The parameter δ governs the range of the quadratic behavior around zero.

4 Results

We will present the depth reconstruction performance for synthetic data with ground truth as well as for real-world data of a road surface.

4.1 Synthetic data

We used 3D surface models to evaluate our algorithm in a quantitative way. The models consist of real-world surfaces that were captured with a camera and a calculated depth map [5]. These scenes were rendered using POV-Ray [6], with a virtual construction similar to our real-world setup. Two cameras were set up to look under the angle towards a central point on a surface located at a defined distance.

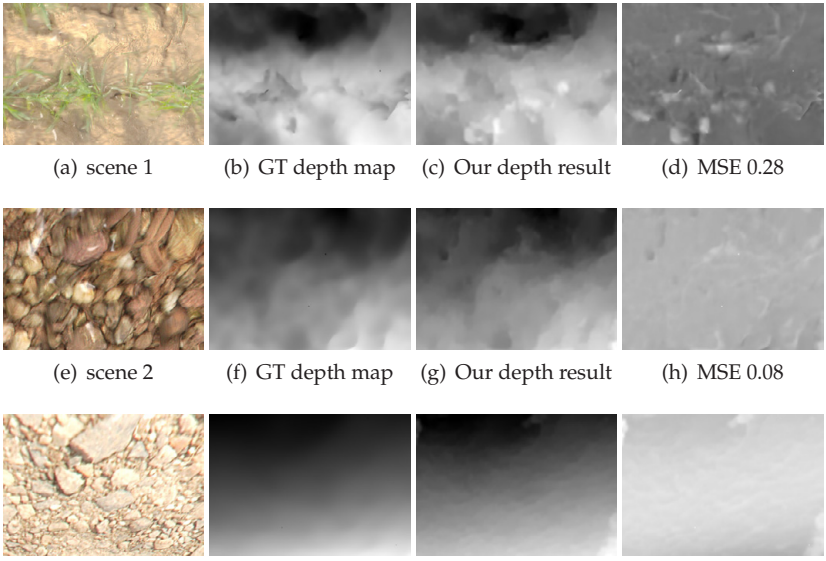
Fig. 3 shows the surface images, ground truth depth maps, our results and an error map. Error rates were measured in disparity values by the mean squared error (MSE), the mean absolute error (MAE) and bad1, which represents the number of pixels with an error $err > 1$ of the disparity estimation result R compared to the total amount of pixels. The surfaces cover disparity ranges from 9.31 (scene 2) to 18.0 (scene 3), where the range is defined as $|\max(R) - \min(R)|$. The results shown in Table 1 were achieved by a 25×25 STABLE descriptor (see Sec. 3.1) with a total variation regularizer (see Sec. 3.3).

4.2 Road surface data

We present results on real-world data acquired driving the system on a freeway. We provide results for STABLE with different descriptor length

and illustrative examples for selected features found during road surface survey. The purpose of this survey is to assess 3D road surfaces as poor road conditions lead to increased wear and tear on vehicles and also impact the surface water transport, noise emission, etc.

Fig. 4 shows the performance of STABLE with descriptor length ranging from 16 bits to 112 bits, which is the maximum bit count possible for the 15×15 matching window. On the one hand, the 16 bit long descriptor still provides quite noisy results, using 32 or 64 bit descriptors improves the reconstruction quality significantly. On the other hand, increasing the size of the descriptor to full 112 bits does not seem to improve the result any further.



(a) scene 1 (b) GT depth map (c) Our depth result (d) MSE 0.28

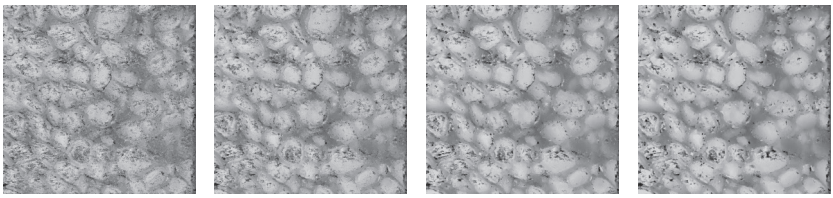
(e) scene 2 (f) GT depth map (g) Our depth result (h) MSE 0.08

(i) scene 3 (j) GT depth map (k) Our depth result (l) MSE 0.59

Figure 3: First column (a, e, i): 3D plot of the original road surfaces. Second column (b, f, j): ground truth (GT) depth maps. Third column (c, g, k): our depth results. Fourth column (d, h, l): error maps, where white indicates no depth error. Detailed error rates are shown in Table 1.

Table 1: Quantitative reconstruction results for rendered road surfaces (in disparity units).

	MSE	MAE	bad1	disparity range	Error MAE [%]
scene 1	0.28	0.37	0.0585	10.3	2.7 %
scene 2	0.08	0.17	0.0008	9.31	1.8 %
scene 3	0.59	0.58	0.1709	18.0	3.2 %



(a) 16 bits (b) 32 bits (c) 64 bits (d) 128 bits

Figure 4: Depth reconstruction quality obtained by 15×15 STABLE with different bit counts (16, 32, 64, and 112, respectively) without regularization.

Sample images from road survey

Due to the lack of ground truth we refer to a manual annotation of interesting properties which are visible to human observers and show the derived 3D reconstruction from which these properties become clearly visible. In most of the results there is a vertically oriented 3D structure visible. This stems from diamond grinding, which is a pavement preservation technique to remove surface irregularities in order to reduce noise and increase safety. We applied postprocessing based on total variation regularization [3], as described in Sec. 3.3, in order to obtain smoother 3D renderings.

Fig. 5 (a) shows an image of a ground concrete road surface with an expansion joint. The grinding stripes, as well as the expansion joint, are visible in the disparity map in Fig. 5 (b). The brighter the disparity the closer the observed object point is to the observer. A 3D rendering of the portion around the expansion joint is provided in Fig. 5 (c). Figs. 5 (d) to (f) show a ground concrete pavement with a small hole. A larger break

out of the surface is shown in Figs. 5 (g) to (i). Finally two grinding lanes of different depth are provided in Figs. 5 (j) to (l). Additionally, in the upper left corner there is material, presumably a chewing gum, observed in the area of the deeper grinding.

5 Conclusions

We have presented the hardware details and algorithms for a line-scan stereo system for close-range surface observation from a mobile platform. Illustrative qualitative results are provided on real-world data and a quantitative evaluation of our approach was shown on synthetic data with ground truth information. Quantitative results showed an average MAE of $2,57\,\%$ (see 1) in relation to the disparity range of the ground truth surface. The visual evaluation additionally shows that the results are well suited for the task of road surface assessment, especially when compared to the state-of-the-art procedures. Further work will include stabilization of the setup and speedup of the computation, especially the regularization step.

References

1. R. Gupta and R. I. Hartley, "Linear pushbroom cameras," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 9, pp. 963–975, 1997.
2. S. Štolc, K. Valentín, and R. Huber-Mörk, "STABLE: Stochastic binary local descriptor for highperformance dense stereo matching," in *Proceedings of IS&T International Symposium on Electronic Imaging: Machine Vision Applications IX*, Feb. 2016.
3. L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.
4. P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Transactions on Image Processing*, vol. 6, no. 2, pp. 298–311, Feb. 1997.
5. Siebencorgie, "Scanned surface collection," http://siebencorgie.jimdo.com/, 2016, [Online; accessed 22-Sept-2016].
6. P. of Vision Pty. Ltd., "Scanned surface collection," http://www.povray.org/download/, 2016, [Online; accessed 22-Sept-2016].

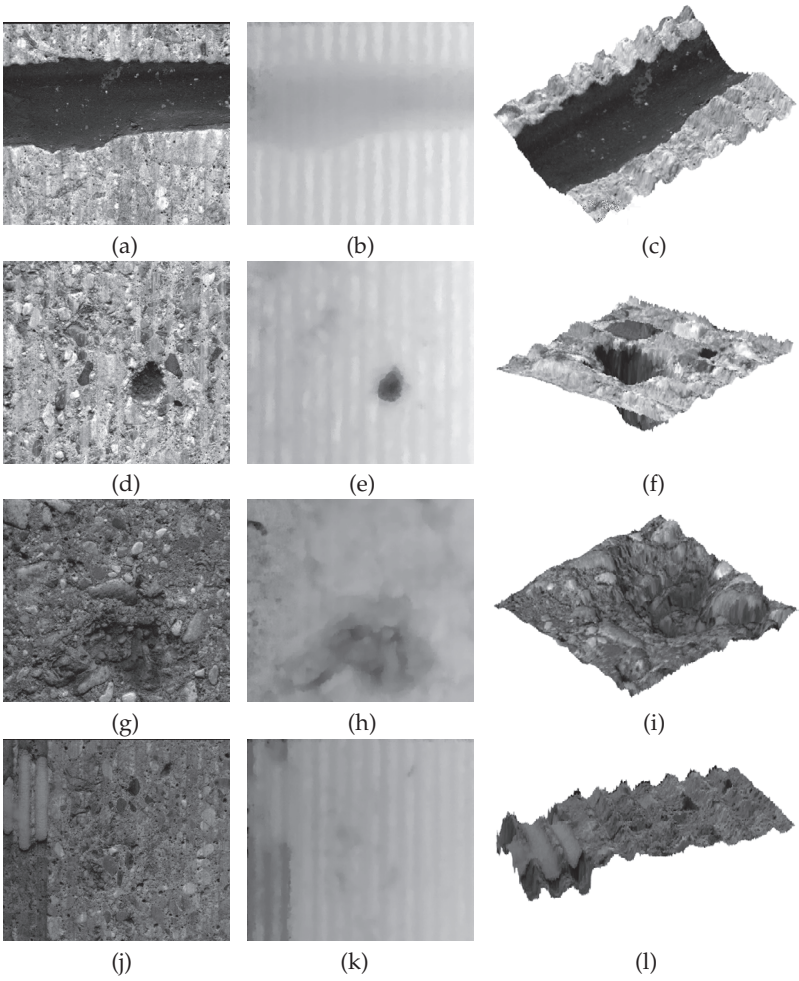


Figure 5: Illustrative results for road surface data (disparity maps smoothed by postprocessing): ground concrete surface with expansion joint (a) image, (b) disparity and (c) 3D rendering (cutout); ground concrete surface with a hole (d) image, (e) disparity and (f) 3D rendering (cutout); unground concrete surface with a larger break out region (g) image, (h) disparity and (i) 3D rendering (cutout); grinding at different depth and a chewing gum like object observed in a portion of the deeper grinding (j) image, (k) disparity and (l) 3D rendering (cutout).

The SPHERE project: Sleep monitoring using computer vision

Manuel Martinez and Rainer Stiefelhagen

Karlsuhe Institute of Technology, Institute for Anthropomatics and Robotics,
Computer Vision for Human Computer Interaction (cv:hci) Lab,
Vincenz-Priessnitz-Str. 3, 76131 Karlsruhe, Germany
{manuel.martinez, rainer.stiefelhagen}@kit.edu

Abstract The aim of the SPHERE project is to develop a compact device that monitors sleep using computer vision. Our system attaches to the ceiling above the bed and integrates infrared and depth cameras, alongside other auxiliary sensors. We installed systems in a nursery home and a sleep laboratory, allowing us to evaluate algorithms that analyze respiration, sleep position and agitation. Compared to other sleep monitoring modalities, computer vision is non-intrusive, and provides a holistic understanding of the bed environment, enabling better alarm systems and cleaner sleep summaries.

Keywords Sleep monitoring, computer vision, respiration.

1 Introduction

Lack of sufficient quality sleep can lead to mental and physical health problems, diminished awareness status and reduced quality of life. This is aggravated on elderly patients, as the ability to sleep deteriorates with age.

There are effective ways to treat most sleep disorders, but they need to be diagnosed first. To perform such diagnoses, patients spend the night at the hospital while being monitored by up to a hundred contact sensors. The sensor report (named polysomnogram) is then reviewed manually. This protocol has significant human and material costs, resulting in waiting lists of several months long.

In nursery homes it is not feasible to throughly monitor the sleep quality of every resident, as a result, most sleep disorders stay undiagnosed. Night watch nurses already need to deal with residents falling out of the bed, having panic attacks, etc. To assist in their task, modern nursery homes have all sorts of intelligent sensors (infrared, or pressure sensors usually) to detect if a bed is occupied or empty, or if a person has fallen out of it. The aim is to minimize the duration between an accident and the arrival of the assistance. The sensors are very sensitive to avoid missing actual dangerous events, therefore they often trigger false alarms, and thus are a source of alarm fatigue, which lowers their usefulness.

We created the SPHERE project with the aim to help to improve the diagnose rate of sleep disorders by monitoring sleep quality indicators, and to generate more reliable alarms for accidents. By using Computer Vision we can explore the bed and its surroundings, allowing us to develop holistic algorithms that provide better understanding of the situation.

We have developed a Medical Recording Device (MRD) [1] with several cameras alongside other auxiliary sensors that works autonomously. We collaborate with a nursery home and the sleep laboratory from the ThoraxKlinikum Heidelberg (THX), who is interested in portable monitoring systems.

Our collaborations are a great advantage over alternative studies which rely on simulated patients and environments. Our experiments and results reveal a huge performance gap between simulated and real scenarios, the latter clearly being more challenging. We leverage on this advantage by learning from the data we collected, and design our algorithms accordingly.

In this paper we describe in detail the problem of continuously monitoring respiration rate on real patients, and then we show state-of-the-art algorithms for sleep position and agitation quantification, which we use to provide nightly sleep quality summaries.

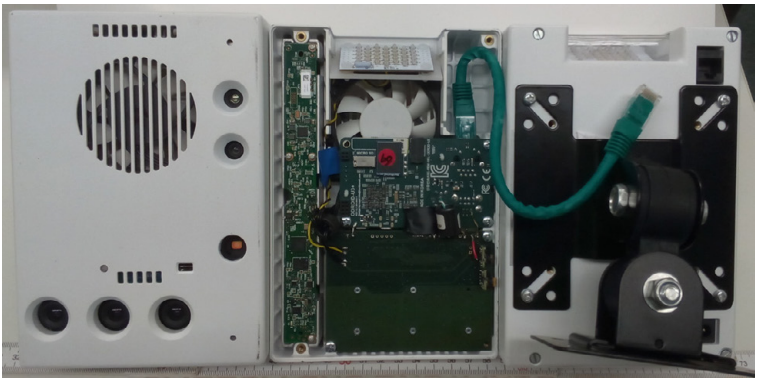


Figure 1: Left to right: front view of our Medical Recording Device, view of the internals, and rear view (including VESA mount). The device has five PCBs of which three have been custom designed for the project.

2 Medical recording device

Our Medical Recording Device (MRD) [1] is compact ($120 \times 180 \times 55\,\text{mm}^3$) and attaches to the ceiling above the bed via a standard VESA mount. It integrates depth [2, 3], stereo and mono cameras; stereo microphone, temperature, pressure, humidity and light sensors; a 4-core ARM CPU with storage, ethernet, WiFi, and bluetooth.

As we need to record in the darkness, we use infrared sensible image sensors with active infrared illumination. The infrared light is projected to the ceiling which reflects over the patient, providing uniformly illuminated images.

The device requires a fan for cooling, but its speed is dynamically controlled and the device is virtually silent by night. It is CE compliant, in order to be installed in hospitals and nursery homes.

We use the hardware compression engine to compress the grayscale images in h264 format, while the quadcore is used to compress the depth maps from the camera using a custom lossless codec.

Our management software recovers automatically from any major flaw, our 7 installed units have accumulated more than 5,000 hours of continuous use without malfunction.

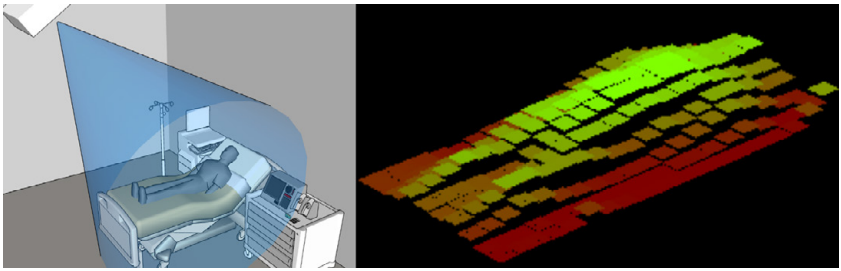


Figure 2: Left: our Medical Recording Device (MRD) on the ceiling of an Intensive Care Unit (ICU). Right: the Bed Aligned Map (BAM), a low resolution, height-based descriptor aligned to the bed. Obtained from a depth camera, it is privacy conscious and robust to light and orientation variations.

3 Bed aligned maps

We capture spatial information from a depth camera and use the bed position, which is automatically estimated, to align the point cloud. The bed matress is divided in equally sized cells, and the mean cell height above the mattress is stored, as seen in Fig. 2. We call the resulting low dimensional descriptor Bed Aligned Map (BAM) [4].

Resolution is an important trade-off for BAM. Lower resolution provides better depth estimates, minor storage requirements, and better privacy protection. However, too low of a resolution may discard important spatial information. Unless stated otherwise, we use $10\,\text{cm} \times 10\,\text{cm}$ cell BAMS, which translates to a descriptor size between 8×20 for the most narrow bed in our database, to 13×20 for the widest.

BAMs are scale, orientation, light and alignment independent, while occlusions are generally filtered out by a raytracing algorithm. This not only makes our algorithms robust to the common ailments of Computer Vision, but also reduces the amount of data we need to collect to train our machine learning tools, as the differences induced by varying scenarios are reduced.

Furthermore, storing BAMs has practical advantages over storing RAW image data: it reduces the storage requirement, which is substantial when performing long term sleep monitoring, and BAMs are ethically friendly, as the subjects are not recognizable.

4 Respiration analysis

Against the general impression, respiration is a very complex signal to retrieve in real conditions. The respiration control system is semi-autonomic: the muscles involved can be voluntarily controlled, but the autonomous system will take care as soon the voluntary control stops. This is important as our respiration pattern changes when we speak, or move our body or become agitated. There exist multitude of events that alter our breathing, some are very common (*e. g.*, snoring, coughing), and some are less common but important nevertheless, like obstructive apnoea. In obstructive apnoea the upper airways are blocked and the diaphragm moves the air from the lungs to the stomach and back, resulting in chest motion but no air exchange.

If an instant breath rate measurement is required in an hospital, it is usually taken by a nurse. The patient will be *told not to move or talk* for a while, and the nurse will count the number of chest excursions during a set period of time. In case of coughing or agitation, the nurse will *repeat the test*.

On polysomnograms, respiration is monitored using no less than 5 sensors: a **thermistor** placed under the nose measures the temperature differential, a **barometer** placed under the nose measures the pressure differential, a **chest band** measures the extension of the thorax, an **abdomen band** measures the extension of the abdomen, finally, a **video stream** records the full session.

Measuring the respiration rate can be as simple as counting chest excursions, but only on very simple scenarios. In SPHERE we want to evaluate how well can we *continuously* estimate breath rate using a depth camera to measure chest excursions in an unconstrained scenario.

4.1 Methodology

We use a dataset obtained from the ThoraxKlinik Heidelberg containing 99 recorded polysomnogram sessions (81 different patients). We took 40 samples for each night, generating a total of 3,960 samples, each being 30 seconds long. Several samples contain challenging situations: empty beds, patients sitting, changing sleep positions, having apnoeas, etc. We use the thermistor signal as a reference for evaluation purposes.

From our MRD, we use a grayscale camera (752×480@10 Hz) and

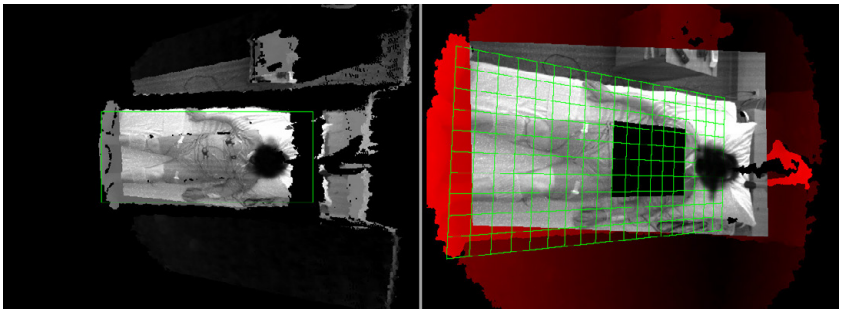


Figure 3: Left: Bed Aligned Map (BAM) generated depth map with infrared image overimposed. Right: Raw disparity image with infrared camera superimposed. In green are displayed the BAM grid used for alignment. The black square is the Region of Interest selected (the same for all patients). Face is hidden to preserve privacy.

a depth camera (PS1080 based, 640×480@30 Hz). Three different bed sizes are used in the study.

To obtain the breathing rate we calculate the Power Spectral Density (PSD) with 8x interpolation. On 30 second windows, this gives us a resolution of 0.25 Breaths Per Minute (BPM). The breathing rate reported is given by the position of the largest peak.

Our evaluation is performed using an acceptance curve: on the **x** axis we have our acceptance threshold, and on the **y** axis we plot the percentage of samples that provide estimates within the threshold distance to our reference (the thermistor signal).

4.2 Breathing rate recognition from images

In a previous work [5], we showed analytically that the signal-to-noise ratio for the breathing signal is inversely proportional to the 4th power of the distance when captured by cameras. Most studies place the depth camera at distances between 70 cm and 1 m [6] to the chest, at those distances, no signal processing is needed to obtain a clean signal. However, as we need to attach the camera to the ceiling to capture the whole environment without obstructions, our distance to the chest is around 4 meters. At 4 meters, the breathing signal we record is 256 times (24 dB)

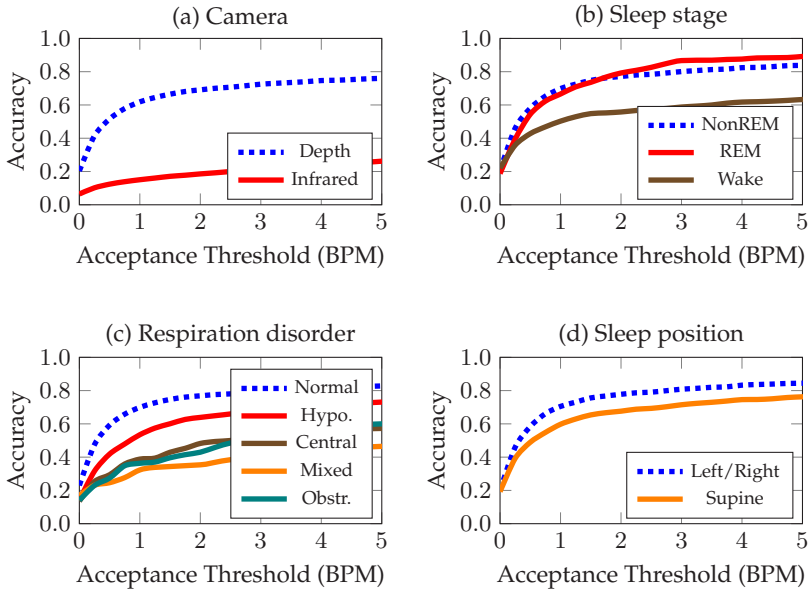


Figure 4: The acceptance curves for our breath rate recognition algorithm show the ratio of samples providing an estimate within the accepted threshold. We evaluate the algorithm under different sensor modalities (a), sleep stages (b), respiration disorders (c) and sleep positions (d).

weaker than at 1 meter, therefore it is crucial to perform signal processing to recover the breathing signal from the background noise.

Our previous approach used PCA combined with Durbin-Watson filters to fuse trajectories [5]. This approach aggressively discards noisy samples to create a very clean signal estimate, however at 4 meters all samples are noisy, and the approach fails to produce an estimate at all.

We use a simpler fusion strategy. First, we create a trajectory for each image pixel. Second, we filter out the pixels using a Region-of-Interest (RoI). Third, we discard trajectories with significant discontinuities. Fourth, we obtain the PSD of each trajectory, and discard pixels whose power outside our interest band ($3 - 30$ BPM) is larger than inside. Last, we aggregate the remaining PSD to create a single PSD estimation.

In our evaluation, we found that depth cameras perform significantly better than infrared cameras (see Fig. 4(a)).

4.3 Effects of sleep stage, respiration disorders, sleep position

Due to the semi-autonomous nature of breathing, it is easier to recognize breathing rate if the patient is sleeping (see Fig. 4(b)). We appreciate a strong impact of sleep disorders in our estimations (Fig. 4(c)). Sleep position has a surprising impact on our estimates: it is more difficult to estimate breathing rate if the patient is in supine position (Fig. 4(d)). Our experience with the dataset suggests that our method is as reliable in supine position than in left or right positions, however there is a larger incidence of sleep disorders when sleeping in supine position, inducing a bias in the measurement.

Our findings confirm that measuring breath rate is reliable if the subject is relaxed and breaths normally, but a chest movement estimator is not sufficient by itself to diagnose respiration disorders.

4.4 Developing a confidence metric

As currently exposed, our system simply reports the location of the maximum peak of the PSD, therefore it provides a breathing estimate in all cases (even if there is no patient in the bed).

We use a simple metric to rate the confidence of our measurements. We consider the ideal measurement to be a PSD consisting of a single, powerful peak, while the worst measurement would provide an almost uniform PSD with no discernible peak. The power of the signal is not a good measure, as the breathing signal may be very weak, and a distractor signal might be very powerful, therefore we normalize the PSD before rating it. Then we compare the normalized PSD to a uniform PSD using the Earth Mover's Distance [7], which is the natural metric to use when comparing histograms. A low distance would imply that our measurement is similar to the uniform PSD, and thus, not very reliable. Conversely, a large distance implies higher reliability.

By applying a simple threshold on such reliability metric, our estimates using the depth camera coincide with the thermistor with a correlation 0.998 (p-value < 0.0001), having a confidence interval of 0.383 Breaths per Minute for a 95 % confidence level.

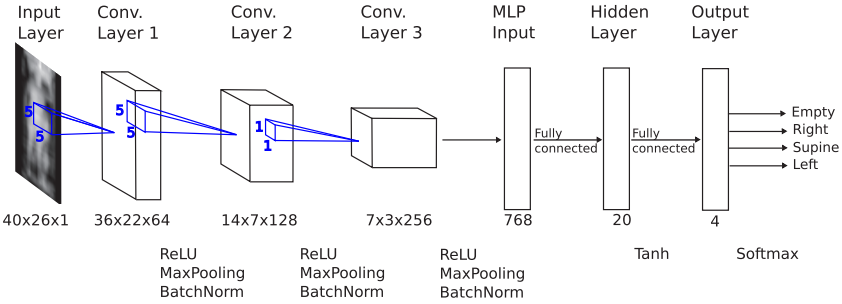


Figure 5: CNN architecture used for sleep position classification. It follows a conventional architecture of three convolutional and two fully connected layers.

Table 1: Confusion matrix for the gravity-based chest sensor worn in the sleep laboratory (left) and our approach based on BAMs and CNNs (right).

	Empty	Right	Supine	Left
Empty	98.61	0.00	0.69	0.69
Right	1.36	93.49	6.24	0.44
Supine	0.00	1.01	96.98	2.00
Left	0.86	0.45	13.24	85.45

(a) Chest Sensor

	Empty	Right	Supine	Left
Empty	98.40	0.00	0.00	1.60
Right	0.40	93.20	4.40	2.00
Supine	0.40	1.60	94.00	4.00
Left	0.00	1.20	4.80	94.00

(b) BAM + CNN

5 Sleep position

Sleep position monitoring is crucial in nursery homes. Medication and illnesses may prevent patients from changing sleep position themselves, and this causes pressure ulcers. If patients do not change sleep position themselves, nurses must move them. But keeping track of the sleep position of all patients is complicated.

We use the deep convolutional neural network pictured in Fig. 5 to classify sleep position from a single BAM into one of the following four classes: “Empty bed”, “Left”, “Supine”, or “Right”. Evaluating on the 81 patients from our sleep laboratory dataset, our algorithm achieves an average accuracy of 93.0 % with a Matthews Correlation Coefficient (MCC) of 0.86. Therefore we outperform the gravity sensor used in the sleep laboratory, which is directly attached to the patient’s chest and uses an accelerometer to localize the gravity vector, and has an accuracy of only 91.9 % with a MCC of 0.84 on the same dataset (see Table. 1).

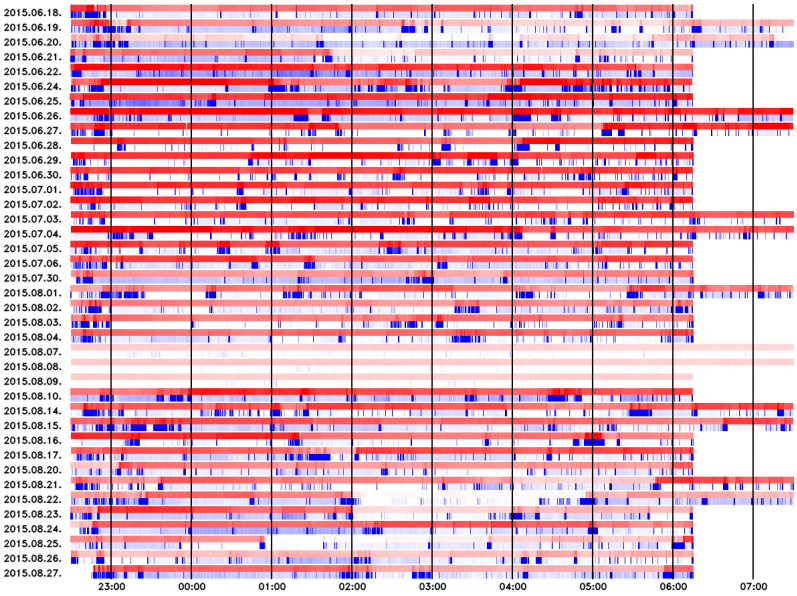


Figure 6: 39 night sleep summary of a nursery home resident. The red bar indicates bed occupancy, showing how the patient spent long hours outside the bed towards the end of the study. The blue bar indicates agitated periods. Best viewed in color.

6 Long term sleep summaries

One of the goals of the SPHERE project is to help assessing the long term sleep quality of nursery home residents. Towards this goal, we generate nightly summaries of the patient sleep using two objective metrics based on BAM: bed occupancy, and agitation.

We define bed occupancy as the volume occupied above the bed mattress. It is simply calculated by adding together all BAM cells. This indicator registers exactly when the patient goes to the bed and wakes up, and helps to quantize the amount of sleep.

We use a custom designed agitation metric to complement bed occupancy. Agitation is a strong health indicator, however there is no objective gold standard to measure it. We suggest to use the absolute

variation of BAMs within one second as an agitation measure, which has already shown compelling results [1, 4].

Both metrics together can summarize a large amount of information in a compact view (see Fig. 6).

7 Conclusions

We have presented the SPHERE project, whose aim is to develop a sleep monitoring system using computer vision. We developed a Medical Recording Device (MRD). It is compact but integrates a wide variety of sensors, including depth and infrarred cameras and it is CE certified. Seven units currently installed in real locations have accumulated more than 5,000 hours without incident.

While the MRD is the hardware backbone of the project, the software backbone is the Bed Aligned Map (BAM), a compact image descriptor based on depth that provides alignment and is robust to common image aliments (light, position, rotation, scale). Using BAMs obtained from the MRD, we have shown algorithms that estimate breath rate, sleep position, agitation and bed occupancy.

More importantly, SPHERE has been designed to be evaluated in real scenarios instead of simulated ones. This has revealed how tasks that were considered simple are actually very challenging when performed in unconstrained scenarios (*e. g.*, breath rate estimation).

We hope that SPHERE represents a big step forward towards the development of automated and non-intrusive sleep monitoring devices that can be deployed in nursery homes and assisted living installations.

Acknowledgements: This work is supported by the German Federal Ministry of Education and Research (BMBF) within the SPHERE project. We want to thank our partners at Thoraxklinik Heidelberg, Evangelische Heimstiftung and Videmo for their invaluable aid.

References

1. M. Martinez and R. Stiefelhagen, "Automated Multi-Camera System for Long Term Behavioral Monitoring in Intensive Care Units," in *Machine Vision Applications (MVA)*, 2013.

2. ——, "Kinect Unleashed: Getting Control over High Resolution Depth Maps," in *MVA*, 2013, pp. 247–250.
3. ——, "Kinect Unbiased," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 5791–5795.
4. M. Martinez, B. Schauerte, and R. Stiefelhagen, ""BAM!" Depth-Based Body Analysis in Critical Care," in *Computer Analysis and Image Patterns (CAIP)*, 2013, pp. 465–472.
5. M. Martinez and R. Stiefelhagen, "Breath Rate Monitoring During Sleep using Near-IR Imagery and PCA," in *International Conference on Pattern Recognition (ICPR)*, 2012.
6. N. Burba, M. Bolas, D. M. Krum, and E. A. Suma, "Unobtrusive measurement of subtle nonverbal behaviors with the Microsoft Kinect," in *International Workshop on Ambient Information Technologies*, 2012.
7. Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision (IJCV)*, vol. 40, no. 2, pp. 99–121, 2000.

Registrierung stabiler Merkmale zur Regularisierung des optischen Flusses bei der erscheinungsbasierten Schätzung der 3D-Kopfpose

Registration of stable features for optical flow regularization for appearance-based 3D head pose estimation

Sebastian Vater und Fernando Puente León

Karlsruher Institut für Technologie,
Institut für Industrielle Informationstechnik,
Hertzstraße 16, 76187 Karlsruhe

Zusammenfassung Die erscheinungsbasierte Kopfposenschätzung mittels des optischen Flusses leidet insbesondere bei großen Kopfdrehungen in die Bildebene hinein aufgrund des Aperturproblems unter Stabilitätsproblemen. In dieser Arbeit wird zur Lösung des Problems eine unabhängige Schätzung der Kopfbewegung durch Berechung einer projektiven Transformation zwischen aufeinanderfolgenden Frames bestimmt. Basierend auf der Registrierung stabiler Merkmale und einer bewegungsabhängigen Regularisierung des optischen Flusses wird die Schätzung stabilisiert.

Schlagwörter Erscheinungsbasierte Kopfposenschätzung, optischer Fluss, Regularisierung, projektive Transformation, stabile Merkmale.

Abstract Appearance-based head pose estimation by means of optical flow computation suffers particularly under large out-of-plane head rotations from stability problems due to the aperture problem. In this work an independent estimation of head pose is performed by computation of a projective transformation between consecutive frames to cope with this problem. The estimation is stabilized based on the registration of stable features and a motion-dependent regularization of the optical flow.

Keywords Appearance-based head pose estimation, optical flow, regularization, projective transformation, stable features.

1 Einleitung

Die dreidimensionale Kopfposenschätzung stellt einen wichtigen Aspekt der Mensch-Maschine-Interaktion dar. Sie findet Anwendung in der Gesichtsdetektion sowie Gesichtserkennung, in der Emotionserkennung, der Bildregistrierung und ist notwendig für eine kopfposeninvariante Blickrichtungsschätzung.

Dieser Beitrag beschäftigt sich mit der erscheinungsbasierten Kopfposenschätzung mit Hilfe von einfacher Hardware, wie etwa Webcams, um eine möglichst breite Anwendung zu ermöglichen. Um aus 2D-Bilddaten die Kopfpose zu schätzen, ist die Berechnung des optischen Flusses [1] ein häufiger Ansatz [2, 3]. Ziel dabei ist es, die vollständige Bewegung des Kopfes aus einer 2D-Bildsequenz in Echtzeit zu gewinnen.

Ein Problem bei der Nutzung einfacher Hardware ist das inhärente Fehlen von Tiefeninformation sowie eine schlechte Bildqualität. Insbesondere Ersteres führt zum Aperturproblem bei der Berechnung des optischen Flusses, welches eine Ambiguität unterschiedlicher Bewegungsschätzungen zur Folge hat [4]. Diese Ambiguitäten sowie kleine Bildgradienten in homogenen Bildregionen führen zu Singularitäten in der Hesse-Matrix in den Gleichungen zur Berechung des optischen Flusses. Das Invertieren der Hesse-Matrix bei der Lösung dieser Gleichungen führt dann zu einem schlecht konditionierten Problem.

Um die hohe Konditionszahl der Hesse-Matrix zu reduzieren, soll in dieser Arbeit ein bewegungsabhängiger Regularisierungsterm in die Berechnung der Hesse-Matrix integriert werden. Dabei soll durch eine auf der Verfolgung stabiler Merkmale basierende, unabhängige Be-

wegungsschätzung Information über die Bewegung direkt in die Regularisierung mit einfließen, um die Regularisierung an die aktuelle Kopfbewegung anzupassen. Dazu wird eine projektive Transformation (Homographie) aus der Zuordnung von Punktepaaren in aufeinanderfolgenden Frames berechnet, aus welcher dann separate Regularisierungsparameter für einzelne Bewegungsrichtungen bestimmt werden. Es werden Kriterien zur geeigneten Auswahl von Punktepaaren erörtert und die Berechnung der Regularisierungsparameter aus der Transformationsschätzung anhand von Schätzergebnissen der Kopfpose für das *Boston University head pose dataset* [5] diskutiert.

2 Stand der Wissenschaft

Bestehende Systeme basierend auf der Berechnung des optischen Flusses, die parametrische Modelle zur Beschreibung der Geometrie des Kopfes und veränderlicher Beleuchtungsbedingungen nutzen, wurden in [2, 5] vorgestellt. In [3] wird eine Stabilisierung der Kopfposenberechnung durch Bildregistrierung durch eine Aktualisierung des Objektmodells basierend auf vorigen Schätzergebnissen durchgegeführt, während [6] ein adaptives Modell nutzt, welches durch eine kontinuierlich aktualisierte Gaußverteilung repräsentiert wird, um die Schätzung robuster zu gestalten.

Existierende Ansätze, die eine Regularisierung des optischen Flusses vorschlagen, verwenden einen skalarwertigen Regularisierungsparameter, um den optischen Fluss zu dämpfen [3, 5]. Dabei ist der Regularisierungsparameter monoton fallend mit steigender Iterationszahl bei der Bestimmung des optischen Flusses, allerdings unabhängig vom aktuellen Bewegungszustand des Kopfes. Die Autoren in [7] verzichten auf eine Berechnung des optischen Flusses und nutzen eine Merkmalsregistrierung und Kalman-Filter zur Bestimmung der Kopfpose.

3 Erscheinungsbasierte 3D-Kopfposenschätzung

Für die 3D-Kopfposenschätzung wird angenommen, dass die Erscheinung eines Bildausschnittes $g(\mathbf{u})$, wobei $\mathbf{u} = [u, v]^{\mathrm{T}}$ eine Pixelkoordinate beschreibt, sich nicht erheblich zwischen zwei aufeinanderfolgenden Frames ändert. In jedem Bildregistrierungsschritt sollen dann die

Bewegungsparameter $\mathbf{p} = [r_\mathrm{x}, r_\mathrm{y}, r_\mathrm{z}, t_\mathrm{x}, t_\mathrm{y}, t_\mathrm{z}]^\mathrm{T}$, wobei r für die Rotationen um die x-, y-, und z-Achse (*roll*, *yaw*, *pitch*) und t für Translationen steht, zwischen zwei Frames gewonnen werden. Unter der Annahme, dass die Änderung in $g(\mathbf{u})$ zwischen den Frames $N-1$ und N durch eine Transformation $f(\,\cdot\,)$ beschrieben werden kann, besteht der Registrierungsschritt darin, $\mathbf{p}$ so zu bestimmen, dass $g_{N+1}\left(f(\mathbf{u}, \mathbf{p})\right) = g_N(\mathbf{u}')$ gilt, wobei $\mathbf{u}' = f(\mathbf{u}, \mathbf{p})$ den Koordinaten von $\mathbf{u}$ im neuen Frame entspricht.

3.1 Posenschätzung durch Berechnung des optischen Flusses

Bei der Kopfposenschätzung beschreibt der Parametervektor $\mathbf{p}$ die Kopfpose. Bei Verwendung eines durch die 3D-Punkte $\mathbf{x} = [x, y, z]^\mathrm{T}$ beschriebenen 3D-Starrkörpermodells zur Modellierung des Kopfes lässt sich der Zusammenhang zwischen Pixelkoordinaten und Pose schreiben als:

$$\mathbf{u}' = \begin{pmatrix} u' \\ v' \end{pmatrix} = P\left(W\left(\mathbf{x}, \mathbf{p}\right)\right) , \tag{1}$$

wobei P eine 3D-2D-Abbildung mit Hilfe eines Lochkameramodells darstellt.

Der Vektor $\mathbf{p}$ kann dann durch Minimierung des Fehlers zwischen einem Objektmodell $g_0(\mathbf{u})$ und $g_{N+1}\left(P(W(\mathbf{u}, \mathbf{p}))\right)$ gefunden werden. Das Minimierungsproblem lässt sich dann mit dem *Compositional Image Alignment*-Algorithmus für inkrementelle Änderungen der Pose auf einem Bildausschnitt $\Omega = g(\omega)$ wie folgt definieren:

$$\begin{aligned} \Delta\mathbf{p} &= \min_{\Delta\mathbf{p}} \mathrm{E}(\Delta\mathbf{p}) \\ &= \min_{\Delta\mathbf{p}} \sum_{\mathbf{u} \in g_0(\omega)} \Big(g_{N+1}\Big(P\big(W\big(W\left(\mathbf{x}, \Delta\mathbf{p}\right), \mathbf{p}\big)\big)\Big) - g_0(\mathbf{u})\Big)^2 . \end{aligned}$$

Auflösen des quadratischen Problems führt mit dem Objektmodell $g_0(\mathbf{u})$ auf

$$\Delta\mathbf{p} = \mathbf{H}^{-1} \sum_{\mathbf{u} \in g_0(\omega)} \left(\nabla g_{N+1}\big(P\left(W\left(\mathbf{x}, \mathbf{p}\right)\right)\big) \left(\frac{\partial P(W)}{\partial \mathbf{p}}\right)^\mathrm{T} \right) \cdot g_\mathrm{Err} ,$$

mit $g_{\mathrm{Err}} = \Big(g_{N+1}\Big(P\big(W\big(W\left(\mathbf{x}, \Delta\mathbf{p}\right), \mathbf{p}\big)\big)\Big) - g_0(\mathbf{u})\Big)$, wobei sich die Hesse-Matrix $\mathbf{H}$ mit den *Steepest Descent Images* $\mathbf{SD} = \nabla g_{N+1}\big(P\left(W\left(\mathbf{x}, \mathbf{p}\right)\right)\big)\left(\frac{\partial P(W)}{\partial \mathbf{p}}\right)$ bestimmt zu $\mathbf{H} = \sum_{\mathbf{u}\in g_0(\omega)} \mathbf{SD}^{\mathrm{T}}\mathbf{SD}$.

3.2 Bewegungsabhängige Regularisierung

Die durch das Aperturproblem entstehenden Ambiguitäten bei der Berechnung des optischen Flusses verursachen Ungenauigkeiten, die schließlich zu einer fehlerhaften Schätzung der Kopfpose führen [4]. Mathematisch drückt sich das Aperturproblem in einer schlecht konditionierten Hesse-Matrix $\mathbf{H}$ bei der Berechnung der inkrementellen Posenänderung $\Delta\mathbf{p}$ aus.

In der Literatur wurde zur Vermeidung von Singularitäten, welche zu Unstabilitäten bei der Invertierung des Hesse-Matrix führen, das Hinzufügen eines konstanten Regularisierungsparameters zur Hesse-Matrix $\mathbf{H} = \sum_{\mathbf{u}\in g_0(\omega)} \Gamma^2\mathbf{SD}^{\mathrm{T}}\mathbf{SD}$ mit $\Gamma = \sqrt{\lambda}\mathbf{I}$ vorgeschlagen.

Dadurch kann zwar die Konditionszahl gesenkt und die Stabilität erhöht werden [3], die Kopfposenschätzung ist allerdings weiterhin unzureichend und führt insbesondere bei Drehbewegungen in die Bildebene zu einem Abbruch des Trackings [4].

Um nun die aktuelle Bewegung in die Regularisierung zu integrieren, kann nach [4] durch Definition eines Regularisierungsparametervektors $\boldsymbol{\lambda} = \left(\lambda_{r_{\mathrm{x}}}, \lambda_{r_{\mathrm{y}}}, \lambda_{r_{\mathrm{z}}}, \lambda_{t_{\mathrm{x}}}, \lambda_{t_{\mathrm{y}}}, \lambda_{t_{\mathrm{z}}}\right)^{\mathrm{T}}$ die regularisierte Hesse-Matrix wie folgt definiert werden:

$$\mathbf{H}_{\mathrm{Reg}} = \mathbf{H} + \sum_{\mathbf{u}\in g_0(\mathbf{x})} \boldsymbol{\lambda}\left(\frac{\partial P\left(W\right)}{\partial \mathbf{p}}\right)\left(\frac{\partial P\left(W\right)}{\partial \mathbf{p}}\right)^{\mathrm{T}}\boldsymbol{\lambda}^{\mathrm{T}}.$$

Durch geschickte Wahl von $\boldsymbol{\lambda}$ ist es somit möglich, die aktuelle Bewegung direkt in die Regularisierung zu integrieren. Es ist wichtig zu bemerken, dass sich durch die 6×6-Matrix $\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}$ auf der Hauptdiagonale direkt die isolierten Bewegungen beeinflussen lassen, während auf den Nebendiagonalen Kreuzterme auftreten, mit denen sich kombinierte Bewegungen manipulieren lassen.

4 Bewegungsadaptive Bestimmung der Regularisierungsparameter

Um eine unabhängige Schätzung der Bewegung zu erhalten, wird im verfolgten Ansatz die Homographie zwischen zwei aufeinanderfolgenden Frames für den Bildausschnitt, der den Kopf repräsentiert, bestimmt. Dazu wird im jedem neuen Frame N basierend auf der letzten Schätzung der Kopfpose und der daraus resultierenden Position des Kopfmodells ein Suchbereich bestimmt. In diesem wird nun nach stabilen Punkten $\tilde{\mathbf{u}}_i = [u, v, w] = [u, v, 1]$ mit dem *SURF*-Verfahren [8] gesucht, welche mit ihrem 128-dimensionalen Merkmalsvektor und ihrer Position sowie dem aktuellen Frame abgespeichert werden. Die Punkte $\tilde{\mathbf{u}}_i^{N-1}$ und $\tilde{\mathbf{u}}_i^{N}$ können nun genutzt werden, um daraus die Homographie zwischen den Frames $N-1$ und N zu bestimmen.

4.1 Homographieberechnung

Eine Homographie lässt sich als 3×3-Matrix $\mathbf{M}_\mathrm{H}$ ausdrücken, wobei $\mathbf{M}_\mathrm{H}$ acht Freiheitsgrade besitzt und das neunte Element ein Skalierungsfaktor ist, der aus der Berechnung von $\mathbf{M}_\mathrm{H}$ mit homogenen Koordinaten resultiert. Mit einer Homographie lassen sich neben affinen Transformationen auch perspektivische Transformationen, die bei Drehungen von 2D-Körpern in die Bildebene hinein entstehen, beschreiben. Für die Bestimmung der Homographie sind somit $n \geq 4$ Punkte notwendig. Für die korrespondierenden Punktepaare $\tilde{\mathbf{u}}_i^{N}$ und $\tilde{\mathbf{u}}_i^{N-1}$ kann dann jeweils zur Berechnung der Homographiematrix $\mathbf{M}_\mathrm{H}$ mit den Zeilenvektoren $\mathbf{h_1}^\mathrm{T}, \mathbf{h_2}^\mathrm{T}, \mathbf{h_3}^\mathrm{T}$ der Zusammenhang

$$\tilde{\mathbf{u}}_i^{N} \sim \mathbf{M}_\mathrm{H}\tilde{\mathbf{u}}_i^{N-1} = \begin{pmatrix} \mathbf{h_1}^\mathrm{T} \\ \mathbf{h_2}^\mathrm{T} \\ \mathbf{h_3}^\mathrm{T} \end{pmatrix} \tilde{\mathbf{u}}_i^{N-1}$$

aufgestellt werden. Damit erhält man

$$\begin{pmatrix} w'\tilde{\mathbf{u}}_i^{N-1\mathrm{T}} & \mathbf{0} & -u'\tilde{\mathbf{u}}_i^{N-1\mathrm{T}} \\ \mathbf{0} & w'\tilde{\mathbf{u}}_i^{N-1\mathrm{T}} & -v'\tilde{\mathbf{u}}_i^{N-1\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{h_1} \\ \mathbf{h_2} \\ \mathbf{h_3} \end{pmatrix} = \mathbf{0}\,,$$

womit sich dann schließlich für vier Punkte ein 8×9-Gleichungssystem ergibt, das mit Hilfe eines *Least-Squares*-Schätzers gelöst werden kann.

4.2 Auswahl und Zuordnung der Merkmale

Das Verfolgen der Punkte beginnt ab dem ersten Frame eines Videos und wird kontinuierlich weiter geführt. Für aufeinanderfolgende Frames können dann auf Basis des 128-dimensionalen Merkmalsvektors unter Berücksichtigung des mittleren quadratischen Fehlers zwischen Merkmalsvektoren zwei Punkte $\tilde{\mathbf{u}}_i^{N-1}$ und $\tilde{\mathbf{u}}_j^{N}$ in unterschiedlichen Frames einander zugeordnet werden, wobei $j = i$ einer erfolgreichen Zuordnung entspricht. Abbildung 1(a) illustriert anhand des Videos *jim9* aus dem *Boston University head pose dataset* die Zuordnung von Merkmalen.

(a) Zugeordnete Merkmale. (b) Merkmalsauswahl Alter. (c) Merkmalsauswahl Metrik. (d) Größte Distanz.

Abbildung 1: Frame 120 des Videos *jim9*. Zu sehen sind zugeordnete Merkmale von Frame 119 auf 120 (a), das Alter der Merkmale (b), die Güte der Metrik (c) sowie die größte Distanz der Merkmale (d). Bei dem Alter und der Metrik entsprechen große Rauten einem hohen Gewicht.

Alter Bei der Featureauswahl nach dem Alter steigt das Gewicht mit zunehmender Anzahl an Frames, an denen ein erfolgreiches Zuordnen von Merkmalen stattfindet. Hierbei geschieht die Zuordnung nur im binären Sinne der Erfüllung eines Schwellwertes, während die Güte der Zuordnung ungeachtet bleibt. Dies ist gut erkennbar in Abb. 1(b), wo die Länge der Trajektorien (Alter) mit der Größe der Marker korreliert. Die Charakteristik Alter gibt somit eine Information über die Stabilität des Merkmals.

Metrik Die Metrik (Abb. 1(c)) ergibt sich direkt aus dem mittleren quadratischen Fehler zweier zugeordneter Merkmalsvektoren. Dabei wird das Gewicht stets nur von der Zuordnung von Frame $N-1$ auf N beeinflusst, da eine erfolgreiche Zuordnung in der Vergangenheit keine Auswirkungen auf die Güte der Zuordnung zum aktuellen Frame und damit auch keine sinnvolle Auswirkung auf die Bestimmung der Homographie hat.

Distanz Bei der Gewichtung nach der Distanz werden vier Merkmale aus einer Aufteilung in vier Quadranten des Suchbereichs möglichst so gewählt, dass ihre Abstände zueinander maximal sind. Abb. 1(d) zeigt beispielhaft die Verteilung der gewählten Merkmale für Frame 120 des Videos *jim9*.

RANSAC Als weitere Möglichkeit zur Auswahl der Merkmale für die Berechnung der Homographie wurde der *Random Sample Consensus*-Algorithmus (RANSAC) getestet, welcher als Parameterschätzverfahren für Stichproben mit vielen Ausreißern verwendet werden kann. Nachteil des Verfahrens ist ein erhöhter Rechenaufwand, der sich aus der iterativen Suche des *Consensus Sets* ergibt.

4.3 Bestimmung der Regularisierungsparameter

Es hat sich gezeigt, dass die direkte Verwendung der Posenschätzung aus der Homographieberechnung zur Aktualisierung der Kopfpose nicht zielführend ist. Daher soll nun die aus der Schätzung der Homographie folgende, unabhängige Posenänderung $\Delta\mathbf{p}^{\mathrm{H}} = \left(\Delta r_{\mathrm{x}}^{\mathrm{H}}, \Delta r_{\mathrm{y}}^{\mathrm{H}}, \Delta r_{\mathrm{z}}^{\mathrm{H}}, \Delta t_{\mathrm{x}}^{\mathrm{H}}, \Delta t_{\mathrm{y}}^{\mathrm{H}}, \Delta t_{\mathrm{z}}^{\mathrm{H}}\right)^{\mathrm{T}}$ indirekt verwendet werden, um Parameter zur Regularisierung des optischen Flusses zu bestimmen.

Hierzu wird ein funktionaler Zusammenhang $\boldsymbol{\lambda} = f(\Delta\mathbf{p}^{\mathrm{H}})$ bestimmt. Zur Bestimmung von $f(\,\cdot\,)$ wurden für alle Frames in allen Videos die auftretenden Bewegegungen nach deren Auftrittshäufigkeit und maximalen Bewegungen untersucht. Aus den minimal zulässigen und maximal auftretenden Bewegungen ergeben sich dann zwei Randbedingungen zur Bestimmung von f. Es wurde dann empirisch ein exponentielles Modell gewählt, welches die unabhängig geschätzte Bewegungsänderung durch Verfolgung der stabilen Punkte auf die Regularisierungsparameter abbildet.

5 Ergebnisse

Um die Methoden zur Auswahl der Merkmale zu evaluieren, wurden insgesamt 45 Videos aus dem *Boston University head pose dataset* ausgewertet. Die Videos zeigen jeweils 200 Frames einer Einzelperson unter verschiedenen Kopfbewegungen unter konstanten Beleuchtungsbedingungen. Die Ground-Truth-Daten wurden mit einem *Flock of Birds*-Tracker aufgenommen, der eine Genauigkeit von $0{,}5°$ aufweist.

5.1 Quantitative Auswertung

Die Ergebnisse des *Mean Absolute Error* (MAE) der *roll*-, *yaw*- und *pitch*-Bewegungen (Rollen, Gieren, Nicken) sind in Tabelle 1 zu sehen. Bei der Auswertung wurden nur die Videos, bei denen ein erfolgreiches Tracking durchgeführt werden konnte, berücksichtigt. Dabei gilt ein Tracking als nicht erfolgreich, wenn ein mittlerer Fehler von mehr als $10°$ bezüglich einer Rotation erreicht wurde.

Tabelle 1: *Mean Absolute Error* gemittelt über 38 Videos für unterschiedliche Wahlen der Merkmalspunkte zur Homographieberechnung.

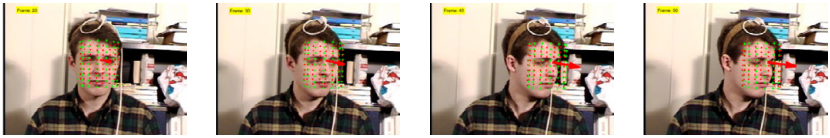
MAE in °	Roll	Yaw	Pitch	Gesamt
Alter	2,38	6,00	3,89	12,26
Metrik	2,18	6,05	3,70	11,84
Distanz	2,11	5,88	3,90	11,89
RANSAC	2,12	4,79	3,62	10,54

Man erkennt, dass die Metrik ähnlich wie die Distanz ein aussagekräftiges Charakteristikum zur Auswahl der Merkmale ist. Das Alter alleine gibt zwar Aufschluss über die Stabilität eines einzelnen Merkmalspunktes, ist allerdings ungeeignet als Kriterium bei der Auswahl der Merkmalspunkte zur Bestimmung der Homographie. Auf Kosten erhöhten Rechenaufwandes wird mit RANSAC das beste Ergebnis bezüglich der Genauigkeit der Kopfposenschätzung erzielt.

5.2 Beispielhafte Auswertung

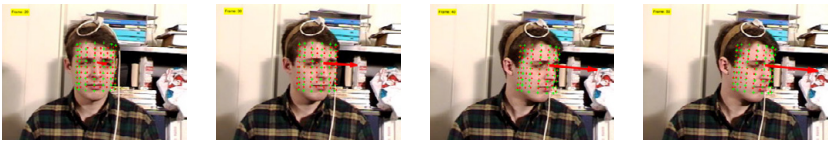
Zur Illustration der Unterschiede der Kopfposenschätzung mit dem Alter der Merkmals und RANSAC als Merkmalsauswahl sind in den Abb. 2 und 3 jeweils vier Frames des Videos *jim9* dargestellt. Der rote Pfeil zeigt den aktuellen Normalenvektor der Frontalen des den Kopf modellierenden Zylinders an. Man erkennt, dass in Abb. 2 sich der Zylinder zum einen zu weit rechts vom Kopf befindet, was aus einer Überschätzung von t_{x} resultiert. Zum anderen erkennt man bei Vergleich von Abb. 2 und Abb. 3, dass die *yaw*-Rotation r_{y} stark unterschätzt wird, wenn man das Alter zur Merkmalsauswahl heranzieht. Die Missdeutung des optischen Flusses zwischen r_y und t_x in Abb. 2 verdeutlicht sehr anschaulich das Aperturproblem, während in Abb. 3 das verbesserte Ergebnis mit der vorgestellten Methode demonstriert wird. Abbildung 4 zeigt diesen Zusammenhang anhand der Ground-Truth und Schätzdaten für die beiden Experimente.

Die Verbesserung der Kopfposenschätzung lässt sich anhand von Abb. 5 am Verlauf der Regularisierungparameter für r_{y} und t_{x} nachvollziehen. In Abb. 5(a) lässt sich kein klar unterschiedliches Verhalten von



(a) Frame 20. (b) Frame 30. (c) Frame 40. (d) Frame 50.

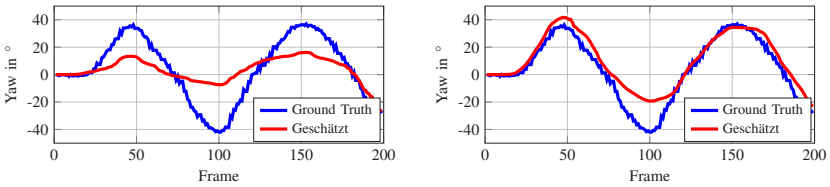
Abbildung 2: Kopfposenschätzung für *jim9* mit Alter als Merksmalsauswahl.



(a) Frame 20. (b) Frame 30. (c) Frame 40. (d) Frame 50.

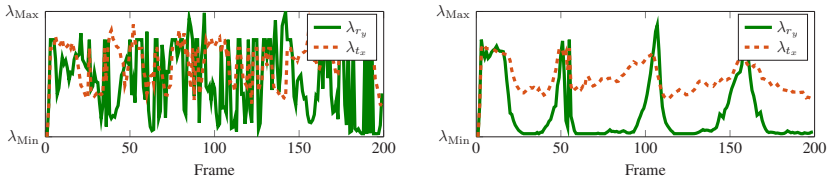
Abbildung 3: Kopfposenschätzung für *jim9* mit RANSAC als Merksmalsauswahl.

λ_{r_y} und λ_{t_x} erkennen. Bei Verwendung von RANSAC reagieren die Regularisierungsparameter wie gewünscht: Man beobachtet eine geringe Regularisierung (kleines λ_{r_y}) des optischen Flusses in Bereichen starker *yaw*-Bewegung. Eine starke Bewegungsänderung kann man insbesondere an den Null-Durchgängen in Abb. 4 erkennen (Frame 70, 120 und 180). Insbesondere zeigt Abb. 5(b) ein dominantes Verhalten von λ_{r_y}, welches im Vergleich zu λ_{t_x} durch niedrige Werte viel optischen Fluss zulässt.



(a) Merkmalsauswahl nach Alter. (b) Merkmalsauswahl mit RANSAC.

Abbildung 4: Schätzung und Ground-Truth der *yaw*-Bewegung (Gieren) für das Video *jim9*.



(a) Merkmalsauswahl nach Alter. (b) Merkmalsauswahl mit RANSAC.

Abbildung 5: Regularisierungsparameter λ_{r_y} und λ_{t_x} für *jim9*.

6 Zusammenfassung

In diesem Beitrag wurde eine Methode zur Online-Berechnung von Regularisierungsparametern zur besseren Konditionierung der Hesse-Matrix bei der Berechnung des optischen Flusses basierend auf der Registrierung von stabilen Merkmalen gezeigt. Dabei wurden unter-

schiedliche Kriterien zur Auswahl geeigneter Merkmale für eine unabhängige Schätzung der Bewegung durch Berechnung einer Homographie geprüft und die Genauigkeit der Schätzung verbessert.

Literatur

1. B. D. Lucas und T. Kanade, „An iterative image registration technique with an application to stereo vision“, in *Int. Joint Conf. Artificial Intelligence*, Vol. 81, 1981, S. 674–679.
2. G. D. Hager und P. N. Belhumeur, „Efficient region tracking with parametric models of geometry and illumination“, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, Nr. 10, S. 1025–1039, 1998.
3. J. Xiao, T. Moriyama, T. Kanade und J. F. Cohn, „Robust full-motion recovery of head by dynamic templates and re-registration techniques“, *Int. J. Imaging Systems and Technology*, Vol. 13, Nr. 1, S. 85–94, 2003.
4. S. Vater, G. Mann und F. Puente León, „A novel regularization method for optical flow based head pose estimation“, in *Automated Visual Inspection and Machine Vision*, F. Puente León und J. Beyerer, Hrsg., Vol. 9530 of Proceedings of SPIE. Bellingham, 2015.
5. M. La Cascia, S. Sclaroff und V. Athitsos, „Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models“, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 22, Nr. 4, S. 322–336, 2000.
6. Z.-W. Chen, C.-C. Chiang und Z.-T. Hsieh, „Extending 3D Lucas-Kanade tracking with adaptive templates for head pose estimation“, *Machine Vision and Applications*, Vol. 21, Nr. 6, S. 889–903, 2010.
7. J.-S. Jang und T. Kanade, „Robust 3D head tracking by online feature registration“, in *8th IEEE Int. Conf. Automatic Face and Gesture Recognition*. Citeseer, 2008.
8. H. Bay, T. Tuytelaars und L. Van Gool, „SURF: Speeded up robust features“, in *European Conference on Computer Vision*. Springer, 2006, S. 404–417.

Automatic corneal tissue classification using bag-of-visual-words approaches

Andreas Bartschat[1,2], Lorenzo Toso[1], Johannes Stegmaier[1], Arjan Kuijper[2], Ralf Mikut[1], Bernd Köhler[1] and Stephan Allgeier[1]

[1] Karlsruhe Institute of Technology (KIT),
Institute for Applied Computer Science (IAI),
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen
[2] Technische Universität Darmstadt (TUD),
Mathematical and Applied Visual Computing & Fraunhofer IGD,
Fraunhoferstr. 5, 64283 Darmstadt

Abstract Corneal confocal microscopy is a promising diagnostic method for peripheral neuropathy. It allows the recording of the sub-basal nerve plexus (SNP) and enables the morphological analysis of peripheral nerves. This work evaluates classification models for real-time evaluation of cornea images in order to find suitable methods for an automatic focus adaptation to the SNP. The analyzed Bag-of-Visual-Words method leads to models with an accuracy of 0.9, even on a small training dataset, and a runtime of $18\,\mathrm{ms}$ per image. Furthermore, the analysis of the support vector machine real-valued output shows high potential for the implementation of real-time focus optimization methods.

Keywords Bag-of-visual-words, corneal tissue, image classification, real-time image processing, feature evaluation.

1 Introduction

Corneal confocal microscopy (CCM) is a spreading technique for investigations of cellular structures in the human cornea. It is non-invasive and allows *in vivo* imaging of the different tissue layers in the cornea with high resolution.

The early and accurate detection, characterization, and quantification of human peripheral neuropathy is important in medical diagnostics

for diseases like diabetes. High expectations are currently placed on CCM to allow rapid and detailed analysis of pathological alterations affecting the peripheral nerves that innervate the cornea, resulting not only in early diagnosis, but also providing insights into the progress and severity of the disease [1–3].

The sub-basal nerve plexus (SNP) of the cornea, located between the corneal epithelium and the stroma (see Figure 1), has the highest density of nerves. The nerve fiber bundles in this layer run in a radial pattern parallel to the corneal surface. However, the distribution of these nerve structures is inhomogeneous, which leads to the problem that a single image is not enough to acquire reliable and representative information. To improve the reliability, the analysis of multiple images from an extended central region of the SNP is possible [4, 5]. Even more insight is provided by the creation and analysis of mosaic images with an increased field of view [6, 7].

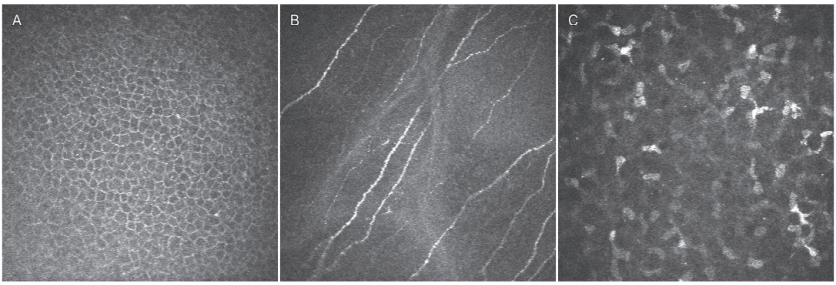


Figure 1: Example images of the tissues in the investigated cornea layers. (A) shows a typical epithelium tissue, (B) a typical SNP image with visible nerve fibers and (C) shows the stroma.

Due to anatomical layer irregularities and reversible folds caused by the contact of the CCM objective, the thickness of the corneal layers and consequently the depth below the surface of the SNP can vary. These phenomena can be countered by manual focus adjustment [6] or the recording of focus stacks and by selecting of the correct SNP images in a post-processing step. Both options are time-consuming and result in long examination sessions.

The automatic online adaptation of the focus to the SNP layer has been proposed as a potential alternative [6]. Such an approach would be able to reduce the examination time and also improve the comparability of results by making the process independent from human intervention, but it requires tracking of the corneal layer of interest by automatic focus adaptation. This automatic focus adaptation can be formulated as a classification problem. By using *a priori* knowledge of the corneal anatomy, the classification of the currently focused tissue can be used to implement an online focus adaptation.

The pipeline for the classification process follows the Bag-of-Visual-Words (BoVW) approach [8], consisting of the feature extraction, the creation of a visual dictionary and the classification process, described in the method section. Since the model should be able to work in real-time, not only the accuracy but also the runtime of the different methods are evaluated, in order to find well-suited parameters for the generic methods as well as the newly developed specialized feature extractor.

2 Data

The analyzed images are *in vivo* recordings of eleven healthy human subjects. The used microscope was the Heidelberg Retina Tomograph with the Rostock Corneal Module[1]. The focus of the imaging session is initialized to acquire depth scans of the SNP and the surrounding tissue layers, the epithelium and the stroma. However, through the entire procedure the focus is not readjusted for different locations, but the depth of the anatomical layer can vary. Thus, the recorded intervals are not guaranteed to cover SNP.

The recorded images cover an area of $0.16\,\mathrm{mm}^2$ of the cornea and are encoded in $8\,$bits with a resolution of 384×384 pixels. Due to non-uniform illumination, caused by the imaging technique, the image intensity decreases at the borders. Furthermore, anatomical irregularities and folds, caused by CCM, can lead to the appearance of multiple tissue types in a single image.

For the training and validation of classification methods, over $10{,}000$ images were labeled in five classes. Three classes are used for images of the three tissue layers: epithelium, SNP, and stroma. These three

[1] HRT II/RCM, Heidelberg Engineering, GmbH, Dossenheim, Germany

classes cover over $2{,}500$ images. The remaining images are labeled as the transition from epithelium to SNP or from SNP to stroma, due to the visibility of more than one tissue type.

Based on this division, a small training dataset is created from four subjects, selecting twelve images of each single-tissue class. The images of the other subjects are used for the evaluation as shown in Table 1. Since images from a single cornea are not independent, remaining images of subjects in S_{Train} were excluded from S_{Eval} to avoid an overestimation of classifier accuracy.

Table 1: Data distribution in S_{Train} and S_{Eval}.

S_{Train}				S_{Eval}			
Subjects	Epithelium	SNP	Stroma	Subjects	Epithelium	SNP	Stroma
1	3	3	1	3	60	0	6
2	7	2	0	4	0	29	109
5	0	4	6	6	4	0	0
10	2	3	5	7	303	673	609
				8	130	305	154
				9	23	43	0
				11	207	265	0
Total	12	12	12	Total	727	1315	878

3 Methods

3.1 Feature extraction

The goal of the feature extraction is to represent image statistics and characteristics as feature vectors. For this task, multiple methods were proposed. This includes well-known representatives such as SIFT [9] and SURF [10], but also ORB [11], BRISK [12], FREAK [13] and the recently proposed LATCH [14] descriptor are analyzed, to determine their suitability for the presented task. Additionally, a histogram-based descriptor is evaluated and, inspired by the visual appearance of the biological tissues, a descriptor based on box filters is designed. This leads to a total of eight compared methods.

The histogram-based feature descriptor BOWHIST represents an image patch by the histogram of the gray scale values with a quantization

to 64 bins. For the BOWBOX descriptor, two filter kernels are designed to highlight blob- and line-like structures in the images, visualized in Figure 2. The kernels consist of a single value for the background and the foreground. In order to compare the results of the filter outputs, the kernels $\mathbf{K}$ are normalized such that:

$$\sum_{i,j} K(i,j) = 0 \tag{1}$$

$$\sum_{i,j} |K(i,j)| = 1 \tag{2}$$

This improves the comparability and enables the decision, if a structure has either one or the other shape. The size of the kernels is chosen according to the size of the anatomical structures. In this case kernels with a side length of 15 pixels and a line width as well as a blob radius of 5 pixels are used.

Figure 2: The proposed box filter $\mathbf{K}^{Blob}$ on the left and $\mathbf{K}^{Line}$ on the right for the feature extraction in the CCM images.

Since lines appear not only in horizontal directions, the line kernel is applied to the original image $\mathbf{I}$ as well as rotations r of the image $\mathbf{I}_r$ with 45, 90 and 135 degrees. This results in the convolutions:

$$\mathbf{I}^{Hor} = \mathbf{I} * \mathbf{K}^{Line} \tag{3}$$

$$\mathbf{I}^{Diag1} = \mathbf{I}_{45} * \mathbf{K}^{Line} \tag{4}$$

$$\mathbf{I}^{Ver} = \mathbf{I}_{90} * \mathbf{K}^{Line} \tag{5}$$

$$\mathbf{I}^{Diag2} = \mathbf{I}_{135} * \mathbf{K}^{Line} \tag{6}$$

$$\mathbf{I}^{Line} = \text{MIP}(\mathbf{I}^{Hor},\ \mathbf{I}^{Diag1}_{-45,},\ \mathbf{I}^{Ver}_{-90},\ \mathbf{I}^{Diag2}_{-135}) \tag{7}$$

$$\mathbf{I}^{Blob} = \mathbf{I} * \mathbf{K}^{Blob} \tag{8}$$

By concatenating the histogram of the maximum intensity projection (MIP) $\mathbf{I}^{Line}$ of the line filtering results and the histogram of the blob detection $\mathbf{I}^{Blob}$, a feature vector of length 64 is computed and used as the patch description.

3.2 Sampling

The points for the feature extraction are selected through a dense sampling pattern. Here, a grid with a certain size is used, to specify sampling points. E. g., a size of 20 means, a sampling point is placed every 20 pixels in horizontal and vertical direction. At each of these sampling points, the feature description method is applied at a certain scale, describing the patch size and the weighting of the patch pixels. For the proposed methods, the scale describes the side length of the squared patch in pixels.

The complete process of feature extraction is therefore defined by three parameters: the feature descriptor, the scale of the feature descriptor and the size of the grid. By specifying a size smaller than the scale, also redundancy in the sampling process is possible through overlapping patches.

3.3 Visual dictionary

Using the described feature extraction, a feature vector is extracted for each sampling point. The representation of the whole image is acquired through the quantization of these vectors with an unsupervised learning method. A common choice for this task is the clustering with the k-means algorithm. In addition, the fuzzy c-means algorithm is implemented and tested. The main parameter for the clustering is the number of clusters, representing the codebook of visual words. For the distance measurement, the Manhattan distance and the Euclidean distance are evaluated.

The quantization of the feature vectors is computed by finding their nearest cluster center, corresponding to the most similar visual word. For each cluster center, the number of assigned feature vectors is summed up. With these sums, a histogram of visual words of an image is created and used for the classification task.

3.4 Classification

The classification of the images is realized with support vector machines (SVM), based on kernels with radial basis functions [15]. The multiclass problem consists of three classes for the three tissue types. For the ap-

plication of the SVMs, the one-against-all approach is chosen as shown in Figure 3. The three SVMs classify a single tissue against the other tissues and, by evaluating the real-valued output, the predicted class is selected.

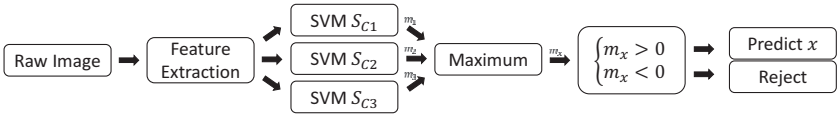


Figure 3: Pipeline of the one-against-all classification process of the histograms using SVMs.

The output of the SVMs is treated as a membership m for a certain class, representing the likeliness of the class. Since the SVMs are optimized for different problems, there is no guarantee of the membership to have an appropriate scale (*cf.* [15] Chapter 7.1.3). However, the analysis of the normalized membership shows reasonable results and enables a more detailed analysis of the classification than the accuracy. The normalization m_{norm} is computed separately for each class. Which means, all membership values of each class are divided through the maximum of the vector over all training memberships $\mathbf{m}_{train}$ of that class. Furthermore, the result is limited to the interval $[-1,\ 1]$:

$$m_{norm} = \min\left(1, \max\left(-1, \frac{m}{\max(\mathbf{m}_{train})}\right)\right) \tag{9}$$

The prediction of the method is the class with the highest membership if it is positive. If no membership fulfills this criterion, the image is rejected.

4 Results

The results of the performance evaluation are visualized in Figure 4. Based on more than 2000 parameter combinations, the Pareto optimal results of every method are selected to show the best possible solutions in terms of accuracy as well as time. The evaluation shows that SURF is able to acquire reasonable results and, if a runtime of more than $50\,\text{ms}$ is tolerated, BOWBOX provides the best accuracy.

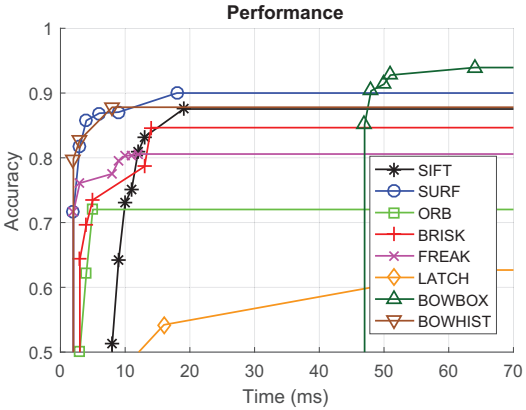


Figure 4: Pareto optimal results of the performance evaluation of the feature extraction methods in terms of accuracy and runtime on the Dataset S_{Eval}.

Table 2: Sensitivity of the runtime for single parameters as well as pairs of parameters for the method BOWBOX. Small values indicate a high influence of the selected parameters.

	Std. deviation for the runtime of BOWBOX				
Parameter	Size	Scale	Cl. Method	Cl. Center	Cl. Distance
Size	1.5601	0.6776	1.5623	1.5498	1.5269
Scale		5.6468	5.7284	5.8001	5.7094
Cl. Method			5.8814	5.9144	5.8884
Cl. Center				5.8816	5.8991
Cl. Distance					5.8657

Since a possible application of the method is the real-time classification of images, the influence of each parameter is investigated by fixing that parameter and analyzing the standard deviation on the runtime when varying all other parameters. The result for the BOWBOX method is shown in Table 2. On the diagonal, the standard deviation for the single parameters is shown. A small value indicates that if this parameter is not changed, the other parameters have only a small impact on the runtime. The results show that fixing the size reduces the standard deviation significantly. Even more insight is provided by the analysis of

fixing two parameters, leading to the result that the parameter size and scale have a large impact on the runtime.

The detailed analysis of these two parameters on the method BOWBOX in terms of accuracy as well as runtime is shown in Figure 5. The conclusion of these graphs is that the size parameter is a compromise between accuracy and time, since a reduction leads to a better accuracy but also to a higher runtime and vice versa. However, the scale hits an upper bound and here, a scale of 40 is a good compromise, since any further increase in scale increases the runtime but the change of accuracy is negligible.

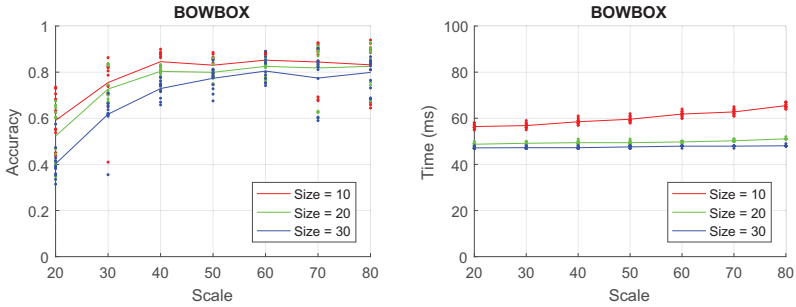


Figure 5: Influence of the scale and size parameters to the accuracy (left) and runtime (right) for the classification, based on the BOWBOX feature descriptor. The lines represent the mean of the results on the Dataset S_{Eval}.

The same conclusions can be drawn from the evaluation of the SURF descriptor shown in Figure 6. Optimal scale values are 20 for a size of 10 or 25 for a size of 20, if a faster runtime is preferred.

The evaluation of the membership m_{norm} on a depth scan is shown in Figure 7. Since the scale of the membership is not comparable, it is normalized by Equation (9). The plot shows a reasonable result for the membership within the epithelium and SNP sections. Unfortunately, the membership for the epithelium is also high in the stroma section, but the predictions for the remaining transition sections look reasonable. Thus, it might be possible to formulate a gradient-based optimization, to keep the microscope focus in the center of the SNP.

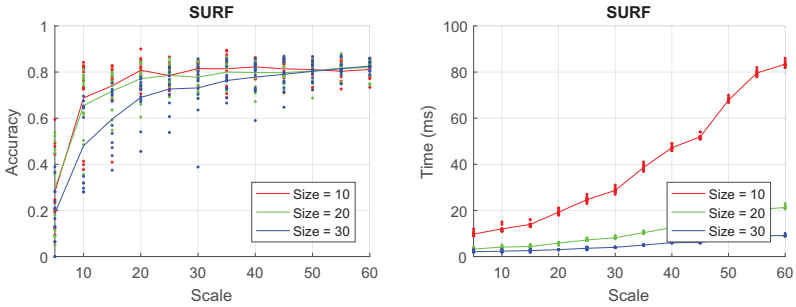


Figure 6: Influence of the scale and size parameters to the accuracy (left) and runtime (right) for the classification, based on the SURF feature descriptor. The lines represent the mean of the results on the Dataset S_{Eval}.

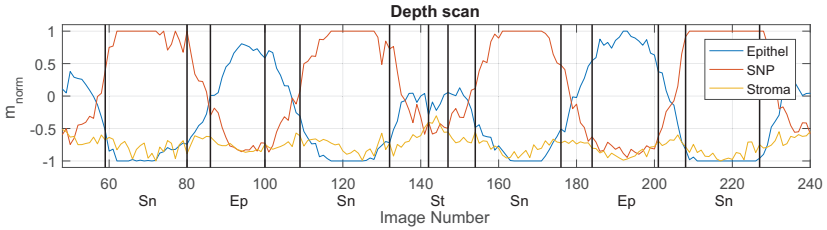


Figure 7: Visualization of the normalized membership of a depth scan from all SVMs in the one-against-all configuration of Subject 2 from S_{Train}. The x-axis shows the image numbers and is additionally labeled with the visible tissues (Ep: epithelium, Sn: SNP, St: Stroma, transition sections are unlabeled). The highest membership is the most likely class of an image and is predicted, if it is positive.

5 Conclusion

This work evaluates feature descriptors and parameters for the classification of CCM images, using the BoVW approach. The goal was to find methods, suitable for real-time focus adaptations during the imaging of the cornea. This is necessary, since anatomical irregularities and folds cause a varying depth of the tissue layers and a reliable imaging of the

SNP can currently only be achieved by manual focus adaptations or the imaging of depth scans, resulting in long examination sessions.

The methods are evaluated on more than $2{,}500$ labeled CCM images of eleven subjects. The results showed that the methods are suitable for the task, leading to maximum accuracies of 0.9 and a runtime of 18 ms per image for well-known representatives such as SURF. The introduced specialized feature descriptor, based on box filter, was able to achieve a higher accuracy of almost 0.94. However, the achieved runtime of 64 ms allows only applications with 15 images per second or less. These results show that the BoVW method is suitable for the classification task, even with small training datasets. Furthermore an evaluation with larger datasets, as presented in [16], showed comparable results, even though the accuracy can be further improved.

The detailed analysis of the SVMs output used for the classification showed high potential for its use in an optimization strategy of the focus adaptation. A gradient-based approach might be able to keep the focus at the center of the SNP, by trying to optimize the real-valued outputs of the SVMs.

References

1. M. Tavakoli, P. Hossain, and R. A. Malik, "Clinical applications of corneal confocal microscopy," *Clinical Ophthalmology*, vol. 2, no. 2, pp. 435–45, 2008.
2. A. Cruzat, D. Pavan-Langston, and P. Hamrah, "In vivo confocal microscopy of corneal nerves: Analysis and clinical correlation," in *Proc. Seminars in Ophthalmology*, vol. 25, no. 5-6. Taylor & Francis, 2010, pp. 171–177.
3. D. Ziegler, N. Papanas, A. Zhivov, S. Allgeier, K. Winter, I. Ziegler, J. Brüggemann, A. Strom, S. Peschel, B. Köhler *et al.*, "Early detection of nerve fiber loss by corneal confocal microscopy and skin biopsy in recently diagnosed type 2 diabetes," *Diabetes*, p. DB_131819, 2014.
4. D. Vagenas, N. Pritchard, K. Edwards, A. M. Shahidi, G. P. Sampson, A. W. Russell, R. A. Malik, and N. Efron, "Optimal image sample size for corneal nerve morphometry," *Optometry & Vision Science*, vol. 89, no. 5, pp. 812–817, 2012.

5. M. Parissi, G. Karanis, S. Randjelovic, J. Germundsson, E. Poletti, A. Ruggeri, T. P. Utheim, and N. Lagali, "Standardized baseline human corneal subbasal nerve density for clinical investigations with laser-scanning in vivo confocal microscopy," *Investigative Ophthalmology & Visual Science*, vol. 54, no. 10, pp. 7091–7102, 2013.
6. S. Allgeier, S. Maier, R. Mikut, S. Peschel, K.-M. Reichert, O. Stachs, and B. Köhler, "Mosaicking the subbasal nerve plexus by guided eye movements," *Investigative Ophthalmology & Visual Science*, vol. 55, no. 9, pp. 6082–6089, 2014.
7. K. Winter, P. Scheibe, B. Köhler, S. Allgeier, R. F. Guthoff, and O. Stachs, "Local variability of parameters for characterization of the corneal subbasal nerve plexus," *Current Eye Research*, vol. 41, no. 2, pp. 186–198, 2016.
8. J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proc. 9th International Conference on Computer Vision*, vol. 2. IEEE, 2003, pp. 1470–1477.
9. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
10. H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.
11. E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 2564–2571.
12. S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 2548–2555.
13. A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 510–517.
14. G. Levi and T. Hassner, "LATCH: Learned arrangements of three patch codes," in *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.
15. C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
16. A. Bartschat, "Automatic classification of cornea tissues for autofocus algorithm," Master's thesis, Technische Universität Darmstadt, 2016.

Robustes Gesichtstracking durch Fusion von Active-Appearance-Modellen und präziser Irislokalisation

Robust face-tracking by fusing active appearance models and precise iris localization

Sebastian Vater, Ralph Ivancevic und Fernando Puente León

Karlsruher Institut für Technologie,
Institut für Industrielle Informationstechnik,
Hertzstraße 16, 76187 Karlsruhe

Zusammenfassung Active-Appearance-Modelle (AAM) leiden unter Stabilitätsproblemen bei großen Kopfbewegungen, Beleuchtungsänderungen oder Verdeckungen. In dieser Arbeit wird gezeigt, wie durch Kombination eines AAM mit einer präzisen Irislokalisation die Robustheit beim Verfolgen eines Gesichtes verbessert werden kann. Die Integration einer separaten Irislokalisation in einen AAM-Tracking-Algorithmus wird anhand repräsentativer Ergebnisse validiert.

Schlagwörter Tracking, Active-Appearance-Modelle, Irislokalisation.

Abstract Active appearance models (AAM) suffer from stability problems when faced with large head movements, changing lighting conditions or occlusions. In this work, we combine an AAM with precise iris localization to enhance the robustness of face tracking. The integration of a separate iris localization into the AAM tracking algorithm is validated by means of representative examples.

Keywords Tracking, active appearance models, iris localization.

1 Einleitung

Gesichtstracking ist ein wichtiger Bestandteil der berührungslosen Mensch-Maschine-Interaktion, das in zahlreichen Anwendungen als Vorverarbeitungsschritt eine entscheidende Rolle spielt und in den vergangenen Jahren ein lebendiges Forschungsgebiet darstellt [1]. Die Kenntnis der Gesichtsposition ist notwendig, um eine Gesichtserkennung durchzuführen [2] oder weitere Merkmale, wie etwa die Augenposition, zu bestimmen [2, 3], um eine Blickrichtungsschätzung durchzuführen [4].

Methoden zur Verfolgung des Gesichtes kann man in Tracking durch Detektion sowie reine Trackingverfahren einteilen. Ein prominentes Beispiel für ein Verfahren nach Tracking durch Detektion ist der verbreitete Ansatz nach Viola und Jones, bei dem in jedem Eingangsbild einer Videosequenz das gesuchte Objekt mit Hilfe eines *sliding window*-Ansatzes durch Klassifikation jedes einzelnen Fensters detektiert wird [5]. Reine Trackingverfahren basieren auf der Berechnung des optischen Flusses [6,7], verwenden direkt Grauwerte von Bildausschnitten und Online-Lernen [8] oder nutzen modellbasierte Ansätze wie Active-Appearance-Modelle (AAM) [9, 10]. Weiterhin lassen sich in der Literatur Beispiele für Ansätze finden, bei denen einzelne Merkmale, wie etwa die Augenposition [11] oder *SURF*-Deskriptoren [12], mit einem Trackingalgorithmus kombiniert werden, um das Gesichtstracking zu stabilisieren.

Aufgrund des breiten Anwendungsgebietes und der komplexen Umgebungsbedingungen, die sich aus Situationen mit unterschiedlichen Beleuchtungsbedingungen, komplexen Hintergründen sowie veränderlichen Kopfposen ergeben, stellt das Gesichtstracking weiterhin eine Herausforderung dar [1].

Um diesem Problem zu begegnen, soll in diesem Beitrag die Information, die in unabhängig verfolgten Merkmalspunkten enthalten ist, in eine erscheinungsbasierte Gesichtsverfolgung mittels eines Active-Appearance-Modells einfließen. Dazu wird eine recheneffiziente, präzise Irislokalisation durchgeführt und in ein AAM integriert. Durch Verknüpfung der Irisinformation mit dem Modellgitter des AAM wird dabei die Initialisierung präzisiert und die Schätzung stabilisiert.

2 Stand der Wissenschaft

Active-Appearance-Modelle [9] haben in den vergangenen Jahren zahlreiche Anwendungen im Bereich des Gesichtstrackings [10, 13] und der Gesichtserkennung [14] gefunden.

Die ursprünglich vorgestellte Methode zur Bildsuche durch iterative Anpassung der gesuchten Modellparameter unter Minimierung einer Gütefunktion [9] wurde in [15] durch Verwendung des optischen Flusses bei der Bestimmung der optimalen Modellparameter mit Hilfe des *Inverse Compositional Alignment*-Algorithmus [16] bezüglich Robustheit und Konvergenz verbessert. Eine Erweiterung der ursprünglichen 2D-AAM wurde in [17] durch 2D+3D-AAM vorgeschlagen. Dabei werden aus 2D-Trainingsdaten unter Berücksichtigung von Randbedingungen, die sich aus Orthogonalitätsforderungen der berechneten Eigenvektoren ergeben, geometrisch erlaubte 3D-AAM erzeugt. Die in [18] vorgestellten Coupled-View-AAM (CVAAM), bei denen ein Modell für unterschiedliche Kopfposen durch entsprechende Trainingsdaten erstellt wird, werden in [19] für die simultane Anwendung mit von beliebigen Kameras aufgenommenen Bildern erweitert. Eine große Herausforderung besteht darin, dass AAM für eine robuste Bildsuche eine präzise Initialisierung [2, 13] benötigen.

Eine einfache Methode, eine geeignete Bildregion zur Initialisierung des AAM zu finden, bietet ein Gesichtsdetektor, der für frontale sowie Gesichter im Profil trainiert werden kann. Ein Nachteil insbesondere der zuletzt genannten sowie anderer auf Tracking basierender Reinitialisierungsmethoden ist, dass lediglich eine grobe Region vorgegeben wird, welche allerdings keine direkte Information über die Lage der Modellpunkte des AAM enthält.

Um die Initialisierung zu verbessern und damit die Bildsuche robuster zu gestalten, wurden in [13] lokale Merkmale [20] und in [21] alle Modellpunkte des AAM mittels optischen Flusses verfolgt.

Korrekturen zur Initialisierung eines AAM in Verbindung mit einem zylindrischen Kopfmodell zur Schätzung der Kopfpose wurden in [22] durchgeführt. Ein Problem bei der Initialisierung stellt, unabhängig von dem für das AAM verwendeten Trainingsdatensatz, die Bestimmung der aktuellen Pose dar. Während einige Ansätze Multi-View-AAM nutzen [13, 14], spielt auch hier die Initialisierung unter nicht frontaler Pose eine wichtige Rolle.

3 Active-Appearance-Modelle

Die Implementierung von AAM für das Gesichtstracking erfolgt nach [15] für unabhängige AAM. Dabei werden beim Aufbau des AAM für die Form und die Textur getrennte Modelle trainiert. Das AAM wird dabei durch Linearkombinationen orthonormaler Form- und Erscheinungsvektoren $\mathbf{s}$ und $\mathbf{A}$ beschrieben. Die beim Aufbau des AAM benötigten statistischen Modelle werden aus einem Trainingsdatensatz erstellt, welcher Bilder von Gesichtern und die dazugehörigen Stützstellen, die inhaltlich ähnliche Punkte im Gesicht kennzeichnen, enthält. Die Form eines unabhängigen AAM wird durch den Vektor $\mathbf{s} = \left[u_1, v_1, u_2, v_2, \ldots, u_\nu, v_\nu\right]^\mathrm{T}$ beschrieben, wobei ν der Anzahl der Stützstellen entspricht, welche zur Beschreibung der Form verwendet werden, und u, v die Koordinaten eines Pixels $\mathbf{u} = (u, v)^\mathrm{T}$ im Grauwertbild $g(\mathbf{u})$ darstellen. Das Formmodell eines unabhängigen AAM wird dann aus einer Linearkombination der Grundform $\mathbf{s}_0$ und n zueinander orthonormalen Formvektoren $\mathbf{s}_i$ gebildet: $\mathbf{s}(\mathbf{p}) = \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i$. Hierbei stellen p_i die Modellparameter dar, durch die eine Instanz des Formmodells beschrieben ist.

Das Modell wird aus den von Hand gesetzten Stützstellen durch eine Hauptkomponentenanalyse gewonnen. Im Formmodell wird nur die Variation der Form, nicht aber die Skalierung, Rotation und Translation aufgenommen. Daher werden im ersten Schritt die Formen mithilfe der generalisierten Prokrustes-Analyse optimal zueinander ausgerichtet und die Grundform $\mathbf{s}_0$ bestimmt. Mittels Singulärwertzerlegung werden dann aus den optimal zueinander ausgerichteten Formen die Formvektoren $\mathbf{s}_i$ für $i = 1, \ldots, n$ gewonnen. Die Formvektoren entsprechen dabei den $2, \ldots, n+1$ Links-Singulärvektoren $\mathbf{l}$ der Dekomposition. Der erste Links-Singulärvektor $\mathbf{l}_1$ entspricht der Grundform und kann nicht für das Formmodell verwendet werden. Eine Entscheidungsgrundlage zur Bestimmung der Anzahl n bieten die Singulärwerte zu den zugehörigen Links-Singulärvektoren. Diese können als Maß verwendet werden, um zu bestimmen, welche Varianz des Trainingsdatensatzes das Modell abdecken soll, wobei die Summe aller Singulärwerte 100 % der im Trainingsdatensatz vorkommenden Formvariation entspricht.

Das Modell der Textur wird durch eine Linearkombination orthonormaler Texturvektoren $\mathbf{A}_i$ beschrieben. Das Texturmodell ist für die

Menge der Pixel $\mathbf{u} = (u, v)^{\mathrm{T}} \in \mathbf{s}_0$ in der Grundform wie folgt definiert: $\mathbf{A}(\mathbf{u}, \boldsymbol{\lambda}) = \mathbf{A}_0(\mathbf{u}) + \sum_{i=1}^{m} \lambda_i \mathbf{A}_i(\mathbf{u}) \quad \forall \mathbf{u} \in \mathbf{s}_0$. Der Pixelwert der Gesamttextur $\mathbf{A}(\mathbf{u}, \boldsymbol{\lambda})$ an der Stelle $\mathbf{u}$ berechnet sich durch eine lineare Überlagerung der Grundtextur $\mathbf{A}_0(\mathbf{u})$ mit den m Texturvektoren $\mathbf{A}_i(\mathbf{u})$, welche mit den Texturparametern λ_i multipliziert werden. Wie auch bei den Formvektoren $\mathbf{s}_i$ gilt, dass die Texturvektoren $\mathbf{A}_i$ orthonormal zueinander sein müssen.

Für die Modellbildung des Texturmodells werden die Texturen der Trainingsbilder per Backward-Warping auf die Grundform $\mathbf{s}_0$ transformiert. Zur Bestimmung der Durchschnittstextur $\mathbf{A}_0(\mathbf{u})$ wird der Mittelwert jedes Pixels $\mathbf{u}$ aller zuvor transformierten Trainingsbilder in $\mathbf{s}_0$ berechnet. Die Texturvektoren $\mathbf{A}_i(\mathbf{u})$ werden auf die gleiche Weise mittels Singulärwertzerlegung bestimmt wie die Formvektoren $\mathbf{s}_i$. Für das Texturmodell können die ersten $1, \ldots, m$ Links-Singulärvektoren der Texturmatrix als Texturvektoren verwendet werden.

Die Bildsuche erfolgt nach dem Verfahren des *Inverse Compositional*-Algorithmus [16]. Der vollständige Ausdruck, welcher schrittweise minimiert werden soll, lautet:

$$\sum_{\mathbf{u} \in \mathbf{s}_0} \left[g\left(W\left(\mathbf{u}, \mathbf{s}(\mathbf{p})\right)\right) - \mathbf{A}_0(\mathbf{u}) - \nabla \mathbf{A}_0(\mathbf{u}) \left. \frac{\partial W\left(\mathbf{u}, \mathbf{s}\left(\mathbf{p}\right)\right)}{\partial \mathbf{p}} \right|_{\mathbf{p}=\mathbf{0}} \Delta \mathbf{p} \right]^2 .$$

Die Warpingfunktion $W(\mathbf{u}, \mathbf{s}(\mathbf{p}))$ beinhaltet Translation, Rotation und Skalierung und entspricht den konkreten Punkten im Suchbild. Sie beschreibt die Transformation der Modellpunkte des AAM auf die Grundform $\mathbf{s}_0$ unter Berücksichtigung der aktuellen Parameterschätzung $\mathbf{p}$. Durch iteratives Lösen für $\Delta \mathbf{p}$ lässt sich dann durch Anpassen der Modellparameter die im *Least-Squares*-Sinne beste Position des Gesichtes finden.

4 Fusion

Bei der Bildsuche für einen Frame N durch iteratives Lösen für $\Delta \mathbf{p}$ im Bild $g^N(\mathbf{x})$ werden die Gitterpunkte von $\mathbf{s}^{N-1}(\mathbf{p})$ auf $\mathbf{s}^N(\mathbf{p})$ von Frame $N-1$ auf Frame N transformiert (zur besseren Übersichtlichkeit wird $W(\mathbf{u}, \mathbf{s}(\mathbf{p}))$ durch $\mathbf{s}(\mathbf{p})$ abgekürzt). Im hier verfolgten Ansatz sollen dabei insbesondere die Gitterpunkte $\mathbf{s}_{\mathrm{il}}^{\mathrm{AAM}}(\mathbf{p}) = (u_{\mathrm{il}}, v_{\mathrm{il}})^{\mathrm{T}}$ und

$\mathbf{s}_{\mathrm{ir}}^{\mathrm{AAM}}(\mathbf{p}) = (u_{\mathrm{ir}}, v_{\mathrm{ir}})^{\mathrm{T}}$ des AAM genauer betrachtet werden, wobei die Indizes il, ir für *Iris links, Iris rechts* stehen. Hierbei soll die Information aus einer unabhängigen Irislokalisation, welche die Punkte $\mathbf{s}_{\mathrm{il}}^{\mathrm{ILok}}$ und $\mathbf{s}_{\mathrm{ir}}^{\mathrm{ILok}}$ liefert, in die Schätzung des AAM integriert werden.

4.1 Präzise Irislokalisation

Der Ansatz zur Irislokalisation geht davon aus, dass jedes Pixel eines Bildes Teil einer Isophoten $\gamma(\mathbf{u})$ ist, welche als Kurve konstanter Intensität $g(\mathbf{u}) = \mathrm{const.}$ verstanden werden kann. Es kann dann gezeigt werden, dass für jedes Pixel der lokale Radius $r(\mathbf{u})$ berechnet werden kann, sodass durch Multiplikation mit dem lokalen Gradienten ein Verschiebungsvektor $\mathbf{d}(\mathbf{u}) = r(\mathbf{u}) \cdot (g_u(\mathbf{u}), g_y(\mathbf{u}))^{\mathrm{T}}$ bestimmt werden kann. Die Schätzung des Irismittelpunktes erfolgt dann durch Gewichtung der Start- und Zielpunkte $\mathbf{u}$ und $\mathbf{u}'$ dieser Verschiebungsvektoren mit der Rundheit $\mathrm{w}_{\mathrm{Rund}}(\mathbf{u}) = \sqrt{g_{uu}^2(\mathbf{u}) + 2g_{uv}^2(\mathbf{u}) + g_{vv}^2(\mathbf{u})}$ sowie einem grauwertbasierten Gewicht $\mathrm{w}_{\mathrm{Iris}}(\mathbf{u}) = \max(g(\mathbf{u})) - g(\mathbf{u})$ und anschließendem Aufsummieren der Gewichte. Der beschriebene Algorithmus wird in einem Suchbereich um jedes Auge ausgeführt, welcher sich aus der vorherigen Schätzung des AAM ergibt. Das Maximum der zweidimensionalen Gewichtungsmatrix für das jeweilige Auge liefert schließlich $\mathbf{s}_{\mathrm{il}}^{\mathrm{ILok}}$ und $\mathbf{s}_{\mathrm{ir}}^{\mathrm{ILok}}$.

4.2 Korrektur durch Fusion mit Irislokalisation

Zu Beginn des aktuellen Frames N wird die Verschiebung zwischen AAM und Irislokalisation für beide Augen bestimmt:

$$\Delta\mathbf{s}_{\mathrm{il}}^{\mathrm{init}} = \begin{pmatrix} u_{\mathrm{il}}^{\mathrm{AAM}} - u_{\mathrm{il}}^{\mathrm{ILok}} \\ v_{\mathrm{il}}^{\mathrm{AAM}} - v_{\mathrm{il}}^{\mathrm{ILok}} \end{pmatrix}, \quad \Delta\mathbf{s}_{\mathrm{ir}}^{\mathrm{init}} = \begin{pmatrix} u_{\mathrm{ir}}^{\mathrm{AAM}} - u_{\mathrm{ir}}^{\mathrm{ILok}} \\ v_{\mathrm{ir}}^{\mathrm{AAM}} - v_{\mathrm{ir}}^{\mathrm{ILok}} \end{pmatrix}.$$

Anschließend wird die aktuelle Schätzung des Gitters auf die aktuellen Punkte $\mathbf{s}_{\mathrm{il}}^{\mathrm{AAM}}(\mathbf{p}), \mathbf{s}_{\mathrm{ir}}^{\mathrm{AAM}}(\mathbf{p})$ zentriert. Die Vektoren

$$\mathbf{s}_{\mathrm{il,zentr.}}^{\mathrm{AAM}}(\mathbf{p}) = \mathbf{s}(\mathbf{p}) - \mathbf{s}_{\mathrm{il}}^{\mathrm{AAM}}(\mathbf{p}), \quad \mathbf{s}_{\mathrm{ir,zentr.}}^{\mathrm{AAM}}(\mathbf{p}) = \mathbf{s}(\mathbf{p}) - \mathbf{s}_{\mathrm{ir}}^{\mathrm{AAM}}(\mathbf{p})$$

geben dann die Abstände aller Gitterpunkte zu den aktuellen Irisschätzungen des AAM an. Nach Zentrierung werden $\mathbf{s}_{\mathrm{il,zentr.}}^{\mathrm{AAM}}(\mathbf{p})$ und

$\mathbf{s}^{\mathrm{AAM}}_{\mathrm{ir,zentr.}}(\mathbf{p})$ entsprechend $\Delta\mathbf{s}^{\mathrm{init}}_{\mathrm{il}}$ und $\Delta\mathbf{s}^{\mathrm{init}}_{\mathrm{ir}}$ bezüglich Rotation und Skalierung angepasst und anschließend mit $\mathbf{s}^{\mathrm{ILok}}_{\mathrm{il}}, \mathbf{s}^{\mathrm{ILok}}_{\mathrm{ir}}$ an die tatsächlichen Irispositionen zurücktransformiert.

5 Ergebnisse

5.1 Datensätze für das 2D-Training

Für das Training und die Auswertungen stehen insgesamt fünf Datensätze bestehend aus jeweils etwa 1.000 Bildern zur Verfügung, die alle am Institut für Industrielle Informationstechnik aufgenommen wurden. Für das Training wurden je nach Proband und Modell n Bilder aus einem Trainingsdatensatz gewählt und dann von Hand mit Stützstellen versehen. Dabei wurde für das Modell *Seb1* beispielsweise $n=4$ von insgesamt 813 Bildern gewählt. Weiterhin wurden noch die Modelle *Seb2*, *Ralph*, *ProbA* und *ProbB*, sowie ein generisches Modell *Gen*, welches mit Hilfe von Trainingsbildern aus Datensätzen ohne *Seb2* erstellt wurde, trainiert.

5.2 Tracking mit AAM

Zunächst wurde das AAM ohne Zuhilfenahme weiterer Information ausgewertet. Hierzu wurde das Modell *Seb1* auf die Daten *Seb1* angewandt, um Referenzdaten für diesen Satz zu generieren. Anschließend wurden die Modelle anhand des Datensatzes *Seb1* ausgewertet, wobei der Versuch *Seb2-1* die Anwendung des Modells *Seb2* auf *Seb1* benennt. Für den mittleren quadratischen Fehler aller 62 Stützstellen ergeben sich neben dem Referenzwert von 0 Pixeln Fehler für *Seb1-1*: $30{,}06$ Pixel für *Seb2-1*, $56{,}86$ Pixel für *Ralph-1* und $14{,}43$ Pixel für *Gen-1*. Abbildung 1 zeigt die Frames 284, 285, 295 und 300 für *Seb1-1*. Man kann gut erkennen, dass das Tracking ab Frame 285 trotz Reinitialisierung mit Hilfe des Gesichtsdetektors fehlschlägt, bis schließlich bei Frame 300 wieder eine erfolgreiche Initialisierung stattfindet.

Trotz des nicht vollständigen Trackings lassen sich anhand der quantitativen Ergebnisse Rückschlüsse ziehen. *Seb2-1* wurde mit Daten des gleichen Probanden trainiert, allerdings ohne Trainingsdaten aus *Seb1*. Es erzielt deutlich bessere Ergebnisse als die Anwendung des Modells *Ralph*. Die Ergebnisse des Modells *Gen*, welches auch Trainings-

(a) Frame 284. (b) Frame 285. (c) Frame 295. (d) Frame 300.

Abbildung 1: Abbruch des Trackings (a) und Reinitialisierung mit Detektor (b). Das AAM konvergiert nicht aufgrund schlechter Initialisierung und kann sich auch in weiteren Frames (c) nicht stabilisieren. Erfolgreiche erneute Reinitialisierung (d). Rot kennzeichnet die Gesichtsdetektion.

daten aus *Seb1* enthält, erzielt die besten Ergebnisse und zeigt die Abhängigkeit von Trainings- und Testdaten.

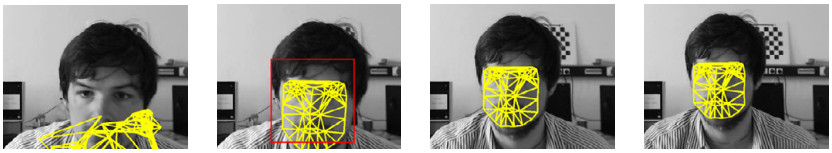
Während eine weitere vollständige, quantitative Auswertung nur durch aufwendiges Annotieren von Hand möglich ist, soll im folgenden Abschnitt qualitativ eine Stabilisierung des Algorithmus durch Intergration der Irislokalisation in das AAM gezeigt werden.

5.3 Reinitialisierung des AAM mit Hilfe der Irislokalisation

Im konventionellen Modell wird das AAM mit seiner Grundform $\mathbf{s}_0$ basierend auf der Gesichtsdetektion mit einem empirisch bestimmten Parameter skaliert, sodass $\mathbf{s}$ sicher im Bereich des Gesichtes liegt. Dies ist notwendig, um ein Konvergieren des Modells zu gewährleisten.

Eine Reinitialisierung wird durchgeführt, wenn mindestens fünf Modellpunkte außerhalb des Detektionsbereichs des Detektors liegen. Dies bringt folgende Probleme mit sich: Es kann erstens keine Reinitialisierung durchgeführt werden, wenn der Detektor fehlschlägt. Zweitens lässt sich durch den Detektionsbereich des Detektors lediglich eine grobe Position zur Initialisierung festlegen. Durch Hinzunahme der Irisinformation kann nun die Initialisierung besser angepasst und das Tracking stabilisiert werden. Insgesamt werden so 15 Frames erfolgreich getrackt, bei denen zuvor kein erfolgreiches Tracking stattfand. Dies ist in Abbildung 2 gezeigt.

Ein weiterer beobachtbarer Vorteil ist, dass sich bei gleichem Konvergenzkriterium die Anzahl der notwendigen Iterationen auf weniger als die Hälfte reduziert (im Beispiel von Frame 1 von 8 auf 3 Iterationen).



(a) Frame 284. (b) Frame 285. (c) Frame 295. (d) Frame 300.

Abbildung 2: Abbruch des Trackings (a) und Reinitialisierung mit Irislokalisation (b). Das AAM konvergiert durchgehend über Frame 295 (c) bis Frame 300 (d).

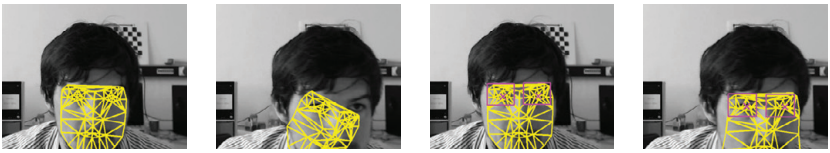
Trotz der Verbesserung nach Reinitialisierung bleibt zu bedenken, dass das Modell schon ab Frame 252 kein erfolgreiches Tracking mehr durchführte und erst bei Einsetzen der Reinitialisierungsbedingung und gleichzeitigem Erfolg des Detektors das Programm stabilisiert werden kann. Es soll nun untersucht werden, wie dies unabhängig von einer Reinitialisierung umgesetzt werden kann.

5.4 Stetige Korrektur durch Irislokalisation

Um eine kontinuierliche Korrektur der AAM-Schätzung wie in Abschnitt 4.2 beschrieben durchzuführen, wird die Irislokalisierung in einem Suchbereich, der um die letzte Schätzung des AAM für die Irispositionen gewählt wird, durchgeführt. Für jede Iteration wird dann das Gitter des AAM an die gefundenen Irispositionen angeglichen, indem basierend auf der aktuelle Form s die Initialisierung bezüglich Translation, Rotation und Skalierung korrigiert wird.

Stabilität bei Verdeckung Abbildung 3 zeigt jeweils die Frames 252 und 261 aus *Seb1-1* ohne (Abbildungen (a) und (b)) sowie mit stetiger Korrektur (Abbildungen (c) und (d)) des AAM. Es ist deutlich zu erkennen, dass bei Verdeckung des AAM die Korrektur das Modell stabilisiert.

Stabilität bei starken Verdrehungen Abbildung 4 zeigt Auswirkungen der Integration der Irislokalisation in das AAM bei Verdrehungen des Kopfes. Eine Stabilisierung des AAM ist deutlich erkennbar.



(a) Frame 252. (b) Frame 261. (c) Frame 252. (d) Frame 261.

Abbildung 3: Verdeckung des AAM. Abbruch des Tracking bei konventioneller Initialisierung ((a) und (b)) und Trackingergebnisse bei kontinuierlicher Korrektur durch die Irispositionen ((c) und (d)). In Magenta sind die Iriden sowie deren Suchbereiche eingezeichnet.

(a) Frame 59. (b) Frame 64. (c) Frame 84. (d) Frame 100.

Abbildung 4: Auswirkungen der Integration der Irisdetektion in das AAM auf die Stabilität bei starker Verdrehung des Kopfes für den Datensatz *ProbandA* mit Integration (oben) und ohne Integration (unten) der Irislokalisation.

6 Zusammenfassung

Der vorliegende Beitrag liefert ein Rahmenwerk zur Stabilisierung des Gesichtstrackings mit Hilfe von AAM durch Integration einer unabhängigen Irislokalisierung in das AAM. Es wurde gezeigt, wie die Information über die Irispositionen während der iterativen Bildsuche durch Anpassen der Modellparameter des AAM mit dem Modellgitter verknüpft werden kann. Ausgehend von der Diskussion konventioneller Trackingergebnisse konnte anhand qualitativer Beispiele eine Verbesserung der Robustheit gegenüber Verdeckungen und großen Kopf-

drehungen sowie eine stabilere Reinitialisierung mit vermindertem Rechenaufwand durch Fusion des AAM mit der Irislokalisation gezeigt werden.

Literatur

1. T. Zhang und H. M. Gomes, „Technology survey on video face tracking“, in *Proc. SPIE Imaging and Multimedia Analytics in a Web and Mobile World*. International Society for Optics and Photonics, 2014, S. 90 270F–90 270F–12.
2. B. Kroon, S. Maas, S. Boughorbel und A. Hanjalic, „Eye localization in low and standard definition content with application to face matching“, *Computer Vision and Image Understanding*, Vol. 113, Nr. 8, S. 921–933, 2009.
3. S. Vater und F. Puente León, „Combining isophote and cascade classifier information for precise pupil localization“, in *IEEE Int. Conf. Image Processing*, 2016, S. 589–593.
4. R. Valenti, N. Sebe und T. Gevers, „Combining head pose and eye location information for gaze estimation“, *Image Processing, IEEE Transactions on*, Vol. 21, Nr. 2, S. 802–815, 2012.
5. P. Viola und M. Jones, „Rapid object detection using a boosted cascade of simple features“, in *Proc. IEEE Comp. Soc. Conf. Computer Vision and Pattern Recognition*, Vol. 1, 2001, S. 511–518.
6. B. D. Lucas und T. Kanade, „An iterative image registration technique with an application to stereo vision“, in *Int. Joint Conf. Artificial Intelligence*, Vol. 81, 1981, S. 674–679.
7. J. Xiao, J.-x. Chai und T. Kanade, „A closed-form solution to non-rigid shape and motion recovery“, in *European Conf. Computer Vision*. Springer, 2004, S. 573–587.
8. Z. Kalal, K. Mikolajczyk und J. Matas, „Tracking-learning-detection“, *EEE Trans. Analysis and Machine Intelligence*, Vol. 34, Nr. 7, S. 1409–1422, 2012.
9. T. F. Cootes, G. J. Edwards und C. J. Taylor, „Active appearance models“, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, Nr. 6, S. 681–685, 2001, method of appearce.
10. J. Xiao, S. Baker, I. Matthews und T. Kanade, „Real-time combined 2D+3D active appearance models“, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, S. 535–542.
11. R. Valenti, Z. Yucel und T. Gevers, „Robustifying eye center localization by head pose cues“, in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009, S. 612–618.

12. J.-S. Jang und T. Kanade, „Robust 3D head tracking by online feature registration“, in *8th IEEE Int. Conf. Automatic Face and Gesture Recognition*. Citeseer, 2008.

13. M. Zhou, Y. Wang und X. Huang, „Real-time 3D face and facial action tracking using extended 2D+3D AAMs“, in *20th Int. Conf. Pattern Recognition*, 2010, S. 3963–3966.

14. J. Sung und D. Kim, „Pose-robust facial expression recognition using view-based 2D+3D AAM“, *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, Vol. 38, Nr. 4, S. 852–866, 2008.

15. I. Matthews und S. Baker, „Active appearance models revisited“, *Int. J. Computer Vision*, Vol. 60, Nr. 2, S. 135–164, 2004.

16. S. Baker, R. Gross, I. Matthews und T. Ishikawa, „Lucas-Kanade 20 years on: A unifying framework: Part 3“, *The Robotics Institute, Carnegie Mellon University*, 2003.

17. J. Xiao, S. Baker, I. Matthews und T. Kanade, „Real-time combined 2D+3D active appearance models“, in *Computer Vision and Pattern Recognition*, 2004, S. 535–542.

18. T. F. Cootes, G. V. Wheeler, K. N. Walker und C. J. Taylor, „Coupled-view active appearance models“, in *British Machine Vision Conference*, Vol. 1, 2000, S. 52–61.

19. C. Hu, J. Xiao, I. Matthews, S. Baker, J. F. Cohn und T. Kanade, „Fitting a single active appearance model simultaneously to multiple images“, in *British Machine Vision Conference*, 2004, S. 1–10.

20. S. Jianbo und C. Tomasi, „Good features to track“, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1994, S. 593–600.

21. R. C. Luo, C. Y. Huang und P. H. Lin, „Alignment and tracking of facial features with component-based active appearance models and optical flow“, in *Int. Conf. Advanced Intelligent Mechatronics*, 2011, S. 1058–1063.

22. J. Sung, T. Kanade und D. Kim, „Pose robust face tracking by combining active appearance models and cylinder head models“, *Int. J. Computer Vision*, Vol. 80, Nr. 2, S. 260–274, 2008.

Ein kamerabasierter Ansatz zur intuitiven Assistenz sehbehinderter Menschen

An intuitive camera-based system to assist visually impaired people

Tobias Schwarze, Martin Lauer[1], Manuel Schwaab[2], Michailas Romanovas[3], Sandra Böhm und Thomas Jürgensohn[4]

[1] Karlsruhe Institut für Technologie (KIT), Institut für Mess- und Regelungstechnik, 76131 Karlsruhe
[2] Hahn-Schickard-Gesellschaft (HSG), 78052 Villingen-Schwenningen
[3] Deutsches Zentrum für Luft und Raumfahrt (DLR), 17235 Neustrelitz
[4] Human-Factors-Consult GmbH (HFC), 12555 Berlin

Zusammenfassung Viele sehbehinderte Menschen sind in ihrer individuellen Mobilität stark beeinträchtigt. Wir stellen ein tragbares Assistenzsystem für blinde Menschen vor, welches die Umgebung mit einer binokularen Kamera erfasst und Hindernisse durch intuitives akustisches Feedback an den Benutzer übermittelt. In einer experimentellen Studie zeigen wir, wie die intelligente Umgebungswahrnehmung zur Sicherheit und Mobilität von Blinden beitragen kann.

Schlagwörter Mobilitätshilfe, Sehbehinderung, Stereoskopie, Sonifikation.

Abstract Many visually impaired persons are severely restricted in their individual mobility. We present a wearable technical aid which perceives the environment with a binocular camera and informs the user about obstacles with intuitive acoustic feedback. In an experimental study we show how the intelligent environment perception can increase safety and mobility of visually impaired people.

Keywords Wearable mobility aid, visual impairment, binocular vision, sonification.

1 Einleitung

Für sehbehinderte und blinde Menschen stellen unbekannte Umgebungen eine große Herausforderung dar. Unabhängige Fortbewegung ist für viele Blinde auf bekannte, gelernte Routen beschränkt. Der traditionelle Blindenstock erlaubt es, Hindernisse im direkt vor der Person liegenden Raum zu erfassen, liefert aber keine Information über weiter entfernte Gefahren. Überhängende Objekte, wie Fensterläden oder Laderampen, stellen eine große Gefahr dar, liegen aber außerhalb des Erfassungsbereiches. Blindenführhunde sind die hilfreichsten Assistenten, aber für die meisten Blinden nicht erschwinglich. Die Entwicklung eines intelligenten und kostengünstigen Hilfsmittels wäre ein wichtiger Beitrag zur Steigerung der Mobilität dieser Personen.

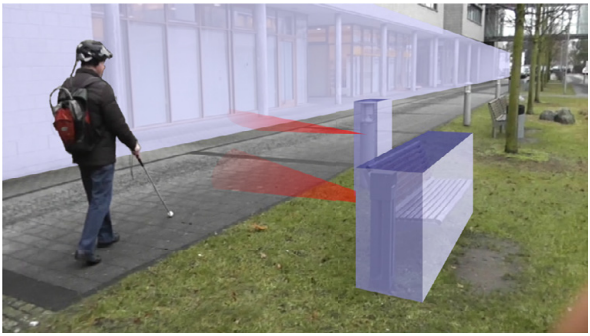


Abbildung 1: Das Assistenzsystem detektiert Hindernisse im Umfeld des Nutzers und übermittelt sie durch räumliche Töne.

Ansätze zu technischen Assistenzsystemen für Blinde gehen zurück bis in die 1960er Jahre, in denen Experimente mit tragbaren Ultraschallsensoren durchgeführt wurden [1, 2]. Bis heute wurden viele Ansätze basierend auf unterschiedlicher Sensorik vorgestellt [3]. Die Systeme vermitteln dem Nutzer durch haptisches oder akustisches Feedback entweder nicht passierbare Bewegungsrichtungen [4,5], oder sie führen den Nutzer in Richtung des freien Raums [6,7]. In beiden Fällen ist kein hohes Maß an Szenenverstehen nötig. Die korrekte Interpretation des Feedbacks ist dem Nutzer überlassen und kann zu einer hohen kognitiven Belastung führen.

Auf der anderen Seite zeigt sich die anhaltende Entwicklung im Bereich der Umfeldwahrnehmung in intelligenten Anwendungen in verschiedenen Bereichen z. B. der Robotik, der Fahrerassistenz oder der Überwachung. Dieser Fortschritt hat uns motiviert, ein Assistenzsystem zu entwickeln, welches die Umgebung interpretiert, um Feedback auf einem höheren Abstraktionslevel anzubieten. Die Nutzung kann hierdurch signifikant erleichtert werden, allerdings stellen sich große technische Herausforderungen. Das System muss tragbar, leicht und unauffällig sein. Die Umgebungsinformationen müssen robust erkannt und intuitiv an den Nutzer übermittelt werden, ohne die natürliche Wahrnehmung zu beeinträchtigen.

Dieser Beitrag beschreibt das Design eines solchen Assistenzsystems. Wir beschreiben die grundlegenden Algorithmen zum Szenenverstehen und die Art des akustischen Feedbacks, welche den Nutzer über die Umgebung informiert. Eine abschließende experimentelle Studie zeigt, wie Blinde von dem System profitieren können.

2 Anforderungen und Systemauslegung

Eine Mobilitätshilfe muss zuverlässige Informationen auf intuitive Weise an den Nutzer übermitteln. Akustisches Feedback bietet hierbei weite Möglichkeiten, relevante Informationen durch spezifische Töne erkennbar zu machen. Unser System erweitert die akustische Realität des Nutzers um eine künstliche akustische Welt, welche die lokale Umgebung des Nutzers beschreibt. Die Idee ist in Abbildung 1 skizziert.

Die Umgebungswahrnehmung des Systems basiert auf einem am Kopf getragenen binokularen Kamerasystem. Es ermöglicht, sowohl Texturmerkmale der Umgebung als auch Tiefenmessungen zu berücksichtigen, um das geometrische Szenelayout zu erfassen, Hindernisse zu erkennen, zu klassifizieren und zeitlich zu verfolgen. Der natürliche Blickwinkel ermöglicht es, die Wahrnehmung auf bestimmte Ziele oder Interaktionspartner zu richten, stellt aber gleichzeitig große Herausforderungen aufgrund der fast uneingeschränkten und unprädizierbaren Kopfbewegungen dar.

Objekte in der Umgebung werden durch virtuelle räumliche Töne ortsgerecht über einen Kopfhörer übermittelt. Die Latenz der kompletten Verarbeitungskette ist mit etwa 100 ms zu groß, um Wahrneh-

mung und Feedback direkt zu koppeln, da Kopfbewegungen während der Verarbeitung zu einer fehlerhaften Richtungswiedergabe führen würden. Eine nahezu latenzfreie Bestimmung der Eigenbewegung ist daher eine wichtige Voraussetzung, für die wir das System mit einer inertialen Messeinheit ergänzen. Die Kenntnis der Eigenbewegung erlaubt es ferner, alle erfassten Informationen in einem Umgebungsmodell zu akkumulieren, welches weit über den aktuellen Sichtbereich der Kameras hinausgeht.

3 Binokulare Umgebungswahrnehmung

Eine der technischen Herausforderungen ist die Entwicklung von Algorithmen zur Umfeldwahrnehmung, welche zuverlässig, robust, gleichzeitig aber effizient genug sind, um in Echtzeit auf einem tragbaren System mit eingeschränkten Kapazitäten eingesetzt zu werden.

Im Vergleich zu traditionellen technischen Mobilitätshilfen wie z. B. einem mit Abstandssensorik erweitertem Langstock [8] ist es für unsere Anwendung nicht ausreichend, den begehbaren Freibereich zu erfassen. Das System muss vielmehr erkennen können, welche Objekte den Freibereich einschränken. In städtischen Umgebungen mit Gebäuden, parkenden Autos, Fahrrädern, Fußgängern, Stühlen und Tischen, Treppen usw. ergibt sich eine große Menge an potentiell relevanter Information. Nur ein kleiner Teil davon kann an den Nutzer kommuniziert werden. Es ist daher nötig, die wahrgenommenen Informationen in eine kompakte Repräsentation zu überführen, in welcher irrelevante Details vernachlässigt werden. Die Repräsentation muss flexibel und ausdrucksstark genug sein, um die Vielzahl möglicher Objekte und deren Bewegung abzudecken, gleichzeitig kompakt genug, um eine effiziente Berechnung zu gewährleisten.

Große Teile innerstädtischer Szenen bestehen aus hohen Wänden, Gebäudefassaden, Zäunen oder Vegetation wie Hecken. Diese Art natürlicher oder menschengemachter Strukturen können als Szenenhintergrund aufgefasst werden, vor dem sich kleine, unabhängig positionierte Objekte befinden und den Szenenvordergrund bilden. Beide Teile unterscheiden sich stark im Ausmaß und darin, dass der Szenenhintergrund immer statische Struktur ist. Die Ausrichtung von Wänden und Hausfassaden bietet darüber hinaus für Blinde eine wertvolle globale

Orientierungshilfe. Dies motiviert uns, den Szenenhintergrund getrennt von möglicherweise beweglichen Objekten im Vordergrund zu modellieren.

Das unterliegende Umgebungsmodell ähnelt einer Blockwelt. Ebene Flächen repräsentieren die Geometrie des Szenenhintergrunds, Vordergrundobjekte werden als bewegliche Quader abstrahiert. In dieser Form stellt das Modell kompakt Informationen auf einem semantischen Level zur Verfügung, das über traditionelle Mobilitätshilfen hinausgeht.

Die Aufgabe des Wahrnehmungssystems ist es, dieses Modell zu füllen und es an die Umgebungssituation anzupassen, während der Nutzer sich durch die Szene bewegt. Ein dichter Disparitätsschätzer bildet die Grundlage für die Extraktion der geometrischen Hintergrundstruktur (Abschnitt 3.1). Vor dieser detektieren wir generische Hindernisse und verfolgen sie zeitlich (Abschnitt 3.2). Um das Modell außerhalb des aktuellen Erfassungsbereichs der Kamera fortzuführen, werden alle Beobachtungen in einem globalen Referenzsystem akkumuliert. Die Eigenposition in diesem System wird kontinuierlich mit einer Kombination aus visueller Odometrie und den Messungen der inertialen Messeinheit geschätzt (Abschnitt 4).

3.1 Geometrischer Szenenhintergrund

Das Modell des Szenenhintergrunds wird durch aneinandergefügte geometrische Ebenen beschrieben. Jegliche Bebauung wird durch Ebenen repräsentiert, welche orthogonal zu einer gemeinsamen Grundebene ausgerichtet sind.

Die Messung dieser Ebenen ist ein simultanes Schätzproblem mehrerer Modelle, das wir mit einer Kombination aus RANSAC und Least-Squares-Optimierung lösen. Bei der Rekonstruktion von Disparitätsmessungen in den euklidischen Raum entsteht ein nicht-linearer Fehler [9]. Messungen werden daher direkt im Disparitätsraum durchgeführt und die Modellparameter anschließend in den euklidischen Zielraum transformiert. Eine Ebene in Bildkoordinaten (u, v) mit Disparitätswerten $\delta(u, v)$ wird beschrieben durch $\alpha u + \beta v + \gamma + \delta(u, v) = 0$ und mittels

$$\mathbf{n}_x X + \mathbf{n}_y Y + \mathbf{n}_z Z + d = 0, \quad (\mathbf{n}, d) \propto (\alpha f, \beta f, \alpha c_u + \beta c_v + \gamma, bf)$$

in den euklidischen Raum transformiert. Die Brennweite f, der Bildhauptpunkt (c_u, c_v) und die Basisweite b sind aus der Kalibrierung des Kamerasystems bekannt. Drei zufällig gewählte Punkte ergeben eine Ebenenhypothese, welche gegen die Daten evaluiert wird, indem die Zahl der Punkte mit $|\alpha u+\beta v+\gamma+\delta(u,v)| < \epsilon$ bestimmt wird. Akzeptierte Hypothesen werden mittels iterativer, linearer Regression optimiert und in das Umgebungsmodell aufgenommen.

Für die zeitlich konsistente Modellierung führen wir für existierende Ebenen im Umgebungsmodell lediglich eine Regressionsoptimierung auf den prädizierten Parametern durch [10]. Für alle vertikalen Ebenen erzwingen wir hierbei als Nebenbedingung die Orthogonalität zum Normalenvektor $\mathbf{n}$ der gemeinsamen Grundebene:

$$\underset{\alpha,\beta,\gamma}{minimize} \sum_{i=1}^{N} (\alpha \cdot u_i + \beta \cdot v_i + \gamma + \delta_i)^2$$
$$subject\ to\ \mathbf{n}_x(\alpha f) + \mathbf{n}_y(\beta f) + \mathbf{n}_z(\alpha c_u + \beta c_v + \gamma) = 0$$

Eine wichtige Einrichtung in bebauten Umgebungen sind Treppen. Treppen können als spezielle Erweiterung der begehbaren Fläche angesehen werden und sind als solche Teil des Szenenhintergrunds. Da sie die Annahme einer lokal flachen Umgebung stark verletzen und für Blinde von deutlicher Relevanz sind, haben wir Treppen gesondert im Umgebungsmodell modelliert. Wir haben hierzu zwei Ansätze entwickelt, um die Parameter eines minimalen geometrischen Treppenmodells zu bestimmen und während der Passage mitzuführen. Sie basieren auf Messungen der Stufenflächen [11] bzw. der charakteristischen konkav/konvexen Kanten [12] aus der Disparitätskarte.

3.2 Hindernisdetektion

Objektdetektion basiert üblicherweise auf charakteristischen visuellen Merkmalen der gesuchten Objekte (z. B. [13]). Diese Ansätze eignen sich, solange die gesuchten Objektklassen bekannt sind. Um eine Eingrenzung zu vermeiden und generisch Hindernisse detektieren zu können, stützen wir uns auf Merkmale, die unabhängig von der visuellen Erscheinung sind. Ein oft eingesetztes Merkmal ist der optische Szenenfluss (z. B. [14,15]), welcher allerdings nur den bewegten Teil der Szene abdeckt und für statische Objekte keine Relevanz hat.

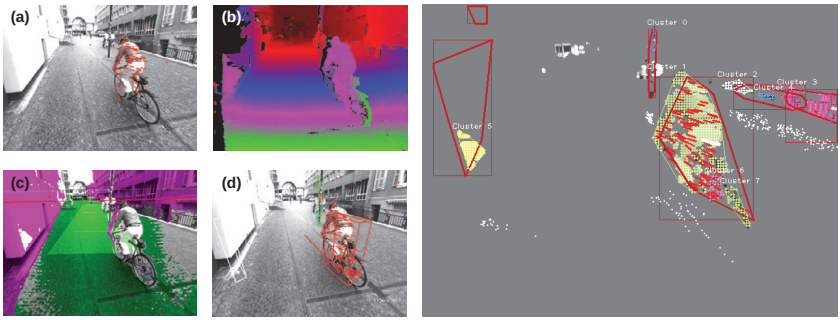


Abbildung 2: Links: Übersicht der Wahrnehmungsalgorithmen. (a) Szenenfluss aus Eigenbewegungsschätzung (b) dichte Disparitätsschätzung (c) Grundebene (grün) und vertikale Bebauung (pink) (d) zeitlich verfolgte Objekte abstrahiert zu ausgerichteten 3D-Quadern. Die geschätzte Bewegung des Fahrradfahrers ist als Linie visualisiert.
Rechts: Übersegmentierung des Disparitätsbildes für die Situation links. Segmente sind farblich getrennt, die Projektion der Hülle aller 3D-Punkte eines Objektes ist als rotes Polygon eingezeichnet.

Wir definieren ein Objekt als Gruppierung von benachbarten Punkten im euklidischen Raum, welche nicht Teil des Szenenhintergrunds sind. Die Hindernisdetektion wird dadurch in ein Segmentierungsproblem überführt, in dem jedes Segment eine Objektdetektion darstellt [16]. Eine Assoziation zu bestehenden Objekten erlaubt dann die zeitliche Verfolgung und Zustandsschätzung [17]. Während der Segmentierung können keine Annahmen über die Anzahl und Art der erwarteten Objekte oder deren typische Form oder Größe getroffen werden. Zusammen mit niedrig aufgelösten, teils stark fehlerbehafteten Disparitätsdaten ergibt sich ein schwieriges, generell ungelöstes Problem.

Zur robusten Detektion unter diesen Bedingungen kombinieren wir die Segmentierung der Disparitätsdaten mit der Objektzustandsschätzung. Um dicht benachbarte Objekte getrennt verfolgen zu können, erzeugen wir zunächst eine Übersegmentierung durch agglomeratives Clustering der Disparitätsdaten, wobei als Ähnlichkeitsmaß zwischen zwei Punkten allein der Disparitätswert betrachtet wird.

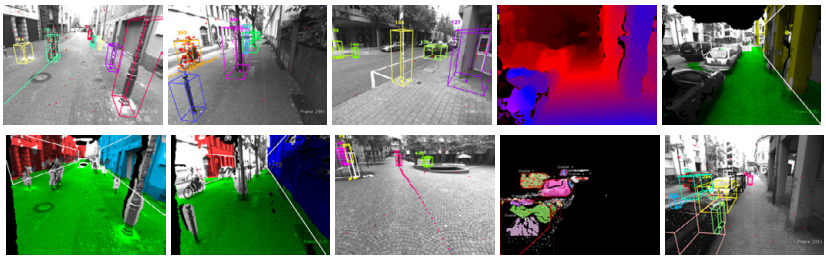
Die Cluster werden anschließend zu Objektdetektionen gruppiert, wobei wir die bestehenden Objekte des Umgebungsmodells heranzie-

hen. Jeder Cluster wird seinem nächsten Objekt zugewiesen, basierend auf (a) dem Überlapp des Clusters mit der in den Bildbereich projizierten Kontur des Objektes $\frac{A_{Cluster} \cap A_{Object}}{A_{Cluster}}$ (Abbildung 2) und (b) der Mahalanobis-Distanz zwischen Objekt und Clusterschwerpunkt. Die gruppierten Cluster dienen dann als Detektionen für die zugeordneten Objekte und werden im folgenden für die Zustandsschätzung genutzt.

Der Zustand jedes Objektes wird durch ein Modell konstanter Geschwindigkeit beschrieben, bestehend aus Position, gerichteter Geschwindigkeit und zusätzlich den Parametern eines umschließenden Quaders und dessen Ausrichtung. Jedes Objekt führt ferner eine Historie der 20 zuletzt rekonstruierten Detektionen mit, aus welchen die Objektkontur und die Größe und Ausrichtung der Box bestimmt wird.

Zur Initialisierung von Objekten im Umgebungsmodell werden Cluster genutzt, welche nicht zu bestehenden Objekten zugeordnet werden konnten. Objekte, die nicht verdeckt sind, sich aber im Sichtbereich der Kamera befinden, werden aus dem Umgebungsmodell entfernt, sobald sie wiederholt nicht durch Detektionen verifiziert werden konnten.

Zur Disparitätsschätzung nutzen wir die OpenCV SGBM Implementierung auf halber Auflösung der 640×480 px Eingangsbilder. Darauf aufbauend werden Szenengeometrie und Vordergrundobjekte parallel mit etwa 15 Hz aktualisiert (i7 2,4 GHz dual-core).



(a) (b) (c) (d)

Abbildung 3: Ergebnisse der Wahrnehmungsalgorithmen. (a,b) (unten) Szenenhintergrund mit überlagerter Grundebene und Bebauung. (oben) Hindernisse mit geschätzten Geschwindigkeiten. (c) Kleine Pfosten detektiert in 12 m Entfernung (oben) und vorbeifahrender Fahrradfahrer verfolgt bis 20 m Entfernung (unten). (d) Fehlerhafte Disparitätsmessungen auf geparkten Autos und deren Segmentierung (linke Spalte). Zwei fehlerhafte Objekte werden in der Nähe des Autos initialisiert (unten rechts, grün und blau).

4 Eigenbewegungsschätzung

Die Kenntnis der eigenen Position und Orientierung bezüglich des Umgebungsmodells ist elementar, um es einerseits über die Zeit konsistent fortzuführen, andererseits um konsistentes akustisches Feedback zu generieren. Die Eigenpose kann hierzu auf zwei Arten bestimmt werden: (a) mit Hilfe der inertialen Messeinheit (IMU), (b) mit Hilfe von kamerabasierter visueller Odometrie [18]. Die IMU hat den Vorteil, mit hoher Frequenz nahezu latenz- und driftfrei zu arbeiten. Die Orientierung des Kopfes kann auch unter sehr starken Bewegungen robust bestimmt werden. Die translatorische Bewegung des Kopfes hingegen ist nur stark driftbehaftet messbar. Visuelle Odometrie auf der anderen Seite kann alle 6 Freiheitsgrade bestimmen, allerdings mit einer Latenz von $50 - 100$ ms, deutlich kleinerer Messfrequenz von etwa $10 - 15$ Hz und unkompensierter Drift. Um die Nachteile beider Methoden zu kompensieren, kombinieren wir sie in einem integrierten Ansatz [19].

Ähnlich einem Gyrsokop misst auch die visuelle Odometrie eine Rotation zwischen zwei Zeitpunkten, welche – bis auf statistische Fehler – der Integration der Drehraten aus der Inertialsensorik zwischen den Zeitpunkten entspricht. Zwar liefert die visuelle Odometrie hier keine neuen Information, kann aber die Orientierungsschätzung statistisch verbessern. Außerdem ermöglicht sie einen Ausgleich von fehlerhaften Inertialmessungen, wie sie z. B. durch Magnetfeldstörungen auftreten können. Das Zusammenführen beider Messungen wird jedoch durch die unterschiedlichen Messfrequenzen beider Sensoren erschwert.

Die Basis unserer Modellierung ist ein übliches Orientierungsfilter basierend auf den Messungen der Drehraten und des magnetischen Feldes [20]. Der Filterzustand wird durch ein Quaternion sowie die Offsets der Gyroskope beschrieben. In unserem Setup erhalten wir Inertialmessungen für jeden Zeitpunkt, Messungen der visuellen Odometrie hingegen nur für $t_k, t_{k+n}, t_{k+2n}, \ldots$. Zur Berücksichtigung solch inkrementeller Messungen geringer Abtastrate setzen wir die Methode des stochastischen Klonens [21] ein. Zum Zeitpunkt t_k wird der Filterzustand mit einem Klon erweitert, welcher als Orientierungsschätzung für t_k festgehalten wird. Zum Zeitpunkt t_{k+n} liegt die Orientierungsschätzung der visuellen Odometrie vor. Diese entspricht der Differenz aus aktuellem und geklontem Filterzustand und geht als Innovationsschritt in das Filter ein (Abbildung 4).

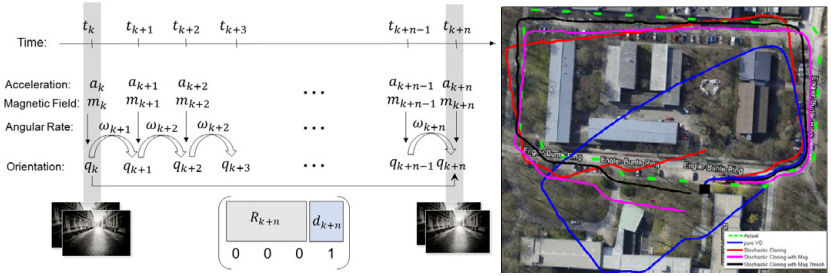


Abbildung 4: Links: Zeitliche Relation der Messungen aus IMU und visueller Odometrie. Zu den Zeitpunkten t_k and t_{k+n} erhalten wir Messungen der Odometrie, dazwischen nur Messungen der IMU. Rechts: Qualitativer Vergleich von visueller Odometrie (blau), dem finalen Filter (schwarz) und dem wahren Pfad (grün).

Die Position der Kameras wird außerhalb des Filters mit Hilfe der visuellen Odometrie bestimmt, wobei nun auf die referenzierte Orientierung des Filters zurückgegriffen werden kann. Das Filter liefert die Kopforientierung mit 400 Hz und die Kopfposition mit etwa 10 Hz. Das erlaubt die nahezu lantenzfreie Prädiktion des Umgebungsmodells, welche besonders wichtig für das akustische Feedback ist. Abbildung 4 zeigt den geschätzten Pfad für eine Strecke von etwa 500 m.

5 Akustisches Feedback

Aus dem Umgebungsmodell kann mit Kenntnis der aktuellen Kopfposition und Blickrichtung die relative Position aller umgebenden Objekten bestimmt werden. Jedes Objekt wird durch eine Tonquelle dargestellt, welche mit Hilfe von binauralen Rendertechniken örtlich korrekt wahrgenommen werden kann. Hierzu wird die kopfbezogene Übertragungsfunktion (engl. Head-Related Transfer Function (HRTF)) benötigt, um Töne richtungsabhängig zu erzeugen, als wären sie auf natürlichem Weg durch die Kopfform und das Außenohr verzerrt und entsprechend dem Zwischenohrabstand verzögert worden [22]. Ein akustisches Bild der Umgebung entsteht, mit dem der Nutzer durch Kopfdrehung frei interagieren kann. Um die kognitive Belastung gering

zu halten, sind hierbei einige Aspekte zu beachten. Der wichtigste ist die Auswahl von geeigneten Tönen, welche intuitiv zu verstehen sind.

Um Verwirrung zu vermeiden, müssen sich die künstlichen Töne klar von natürlichen Umgebungsgeräuschen unterscheiden lassen. Sie sollten angenehm klingen und semantische Information über das Hindernis transportieren, z. B. dessen Art, Bewegung oder Gefahrenpotential. Um geeignete Objektkategorien zu definieren, wurde eine Studie mit blinden Teilnehmern durchgeführt. Diese ergab eine Einteilung in breite Hindernisse (z. B. Bänke), pfostenartige Hindernisse, überhängende Hindernisse (z. B. Laderampen) und sich nähernde Objekte.

Eine wichtige Rolle spielt der technische Aspekt der Lokalisierbarkeit der gewählten Töne. Um die Richtung eines Tones hören zu können, sollte dieser ein möglichst breites Frequenzspektrum abdecken. Dies steht meistens im Widerspruch mit dem Wunsch nach angenehm klingenden Tönen. Insbesondere hohe Frequenzen sind hier problematisch, welche für die Lokalisierbarkeit besonders hilfreich sind. Gleichzeitig sollten die verschiedenen Töne im Zusammenspiel harmonieren, da im Betrieb mehrere Töne gleichzeitig zu hören sind. In einer zweiten, simulativen Studie wurden hierzu Experimente zur Lokalisierbarkeit, dem subjektiven Empfinden und der Zuordnung zu den in der ersten Studie gefundenen Objektkategorien durchgeführt.

Ein letzter wichtiger Aspekt ist die Auswahl von relevanten Objekten in der gegebenen Situation. In städtischen Umgebungen befinden sich üblicherweise deutlich mehr Objekte in Nähe des Nutzers als die maximale Anzahl Töne, die gleichzeitig unterscheidbar sind. Wir wählen hierfür die drei Objekte, die aufgrund ihrer Entfernung und der Abweichung von der aktuellen Bewegungsrichtung die höchste Relevanz haben. Virtuelle Tonquellen werden auf diesen Objekten platziert, mit der HRTF gefaltet und die Lautstärke entsprechend der Entfernung angepasst.

6 Untersuchung mit blinden Probanden

Die prototypische Umsetzung des Systems basiert auf einem Helm, in den die Kameras und die inertiale Messeinheit eingelassen sind (Abbildung 5). Für akustisches Feedback sind Kopfhörer angebracht, welche durch die freischwebende Position die natürliche akustische Umge-

Abbildung 5: Links: Kameras, Kopfhörer und inertiale Messeinheit (nicht sichtbar) sind prototypisch in einen Helm integriert. Rechts: Proband mit dem System im Testparcours.

bung des Nutzers nicht abschatten, sondern lediglich überlagern. Alle Berechnungen werden auf einem Notebook ausgeführt, das in einem Rucksack getragen wird.

Um das prinzipielle Systemkonzept und den potentiellen Nutzen zu bestätigen, wurde ein Feldtest mit blinden Testpersonen durchgeführt. Gerade weil das Verhalten des Nutzers durch das Feedback des Systems beeinflusst wird, ist es wichtig, den Zusammenschluss aus Wahrnehmungsalgorithmen, akustischem Feedback und Nutzerverhalten zu testen. 8 Probanden zwischen 20 und 50 Jahren evaluierten das Systemkonzept. Fünf der Teilnehmer sind unabhängig mobil, drei sind auf fremde Unterstützung angewiesen.

Der erste Teil des Tests diente zur Eingewöhnung und bestand aus zwei großen Hindernissen auf einem offenen Feld. Durch Annäherung oder Vorbeilaufen an den Objekten, wobei sich der Ton entsprechend um die Person bewegt, konnten sich die Teilnehmer schnell mit dem Prinzip des räumlichen akustischen Feedbacks vertraut machen.

Im zweiten Teil der Studie ging es um das gezielte Ausweichen vor Hindernissen. Entlang einer Rasenkante, welche zur Orientierung diente, wurden verschiedenartige Hindernisse mit wenigen Metern Entfernung aufgestellt (flache Kisten, pfostenartige und ein überhängendes Hindernis), teils direkt auf dem Pfad, teils lateral versetzt (siehe Abbildung 5). Zu Beginn tendierten die Teilnehmer dazu, kurz stehenzubleiben, sobald ein neues Hindernis vertont wurde, und durch Kopfdrehung die genaue Richtung zu verifizieren. Später reduzierten sie ihre Geschwindigkeit, bis das Hindernis

mit dem Blindenstock erreicht werden konnte. Als schwierig stellte sich die Einschätzung der Entfernung allein aufgrund der Lautstärke des Hindernisses heraus. Eine längere Phase ist hier nötig, um das Verhältnis zwischen Lautstärke und Distanz zu lernen. In einem zweiten und dritten Durchlauf hatten einige der Teilnehmer eine ausreichende Einschätzung entwickelt, um Objekten frühzeitig auszuweichen.

Das Prinzip des akustischen Feedbacks wurde in den Simulatorstudien vielfach skeptisch bewertet. Die Erfahrung mit dem realen System unter realistischen Bedingungen fiel hier für die Teilnehmer positiver aus als erwartet. Die akustische Überlagerung der Umgebung führte nicht zu einem Gefühl der Einschränkung des natürlichen Hörens. Das Konzept, den Nutzer über seine Umwelt zu informieren, statt Navigationshinweise zu geben, wurde positiv aufgenommen. Eine Entscheidung komplett dem Assistenzsystem zu überlassen stellt für viele Menschen eine hohe Hürde dar – sie bleiben gern selbst in Kontrolle. Alle Teilnehmer konnten sich vorstellen, ein derartiges System zur Assistenz einzusetzen.

7 Zusammenfassung

Mit dem Ziel, die individuelle Mobilität von sehbehinderten und blinden Menschen zu erhöhen, haben wir ein tragbares, kamerabasiertes Assistenzsystem entwickelt. Mit der Kombination aus Umfeldwahrnehmung auf einem erhöhten semantischen Niveau und intuitivem akustischem Feedback konnten viele Einschränkungen bestehender Hilfsmittel überwunden werden.

Die Wahrnehmung der Umgebung basierte auf einer abstrakten Modellierungen des Szenenhintergrunds durch robuste Schätzung geometrischer Modelle und wurde komplementiert durch generische, potentiell bewegliche Objekte im Szenenvordergrund, die zeitlich verfolgt werden. Das entstandene kompakte Modell diente als Basis für akustisches Feedback. Ein wichtiges Element war hierbei die hochfrequente, driftarme und latenzfreie Schätzung der Kopfposition und -orientierung, für die wir eine inertiale Messeinheit mit kamerabasierter visueller Odometrie kombiniert haben.

Die Umsetzung des akustischen Feedbacks basierte auf einer Reihe von Simulatorstudien, mit deren Hilfe geeignete Konzepte und Töne für

ein intuitives Verständnis gefunden wurden. Basierend auf den Ergebnissen wurde ein Konzept entwickelt, das den Nutzer über potentielle Gefahren informiert, ihm die Aktionsentscheidung aber nicht abnimmt. Dies trägt vermutlich positiv zur Akzeptanz eines solchen Systems bei.

Finale Tests unter realen Bedingungen konnten die Nützlichkeit des Sytems und dessen intuitive Nutzbarkeit aufzeigen. Der Wahrnehmungsbereich konnte von etwa einem Meter mit dem Blindenstock auf $10 - 20$ Meter erhöht werden, wodurch gefährliche Situationen frühzeitiger und zielgerichteter vermieden werden können. Zudem ermöglichte es die Warnung vor überhängenden Gefahren, welche mit dem Blindenstock nicht erfasst werden können.

Die Miniaturisierung und Integration in ein unauffälliges Brillengestell sind geplante weitere Schritte, um einen Alltagseinsatz zu realisieren. Das Kamerasystem kann hier weitere nützliche Anwendungen ermöglichen, wie z. B. das Lesen von Text oder die Erkennung spezieller Infrastruktur wie Haltestellen oder Zebrastreifen.

Literatur

1. L. Kay, „An ultrasonic sensing probe as a mobility aid for the blind“, *Ultrasonics*, Vol. 2, Nr. 2, 1964.
2. L. Russel, „Travel path sounder“, *Rotterdam Mobility Research Conference*, 1965.
3. D. Dakopoulos und N. Bourbakis, „Wearable obstacle avoidance electronic travel aids for blind: A survey“, *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 40, Nr. 1, Jan 2010.
4. G. P. Fajarnes, L. Dunai, V. S. Praderas und I. Dunai, „CASBLiP–a new cognitive object detection and orientation system for impaired people“, *trials*, Vol. 1, Nr. 2, 2010.
5. A. Rodríguez, J. J. Yebes, P. F. Alcantarilla, L. M. Bergasa, J. Almazán und A. Cela, „Assisting the visually impaired: Obstacle detection and warning system by acoustic feedback“, *Sensors*, Vol. 12, Nr. 12, 2012.
6. S. Shoval, I. Ulrich und J. Borenstein, „Navbelt and the guide-cane“, *IEEE Robotics & Automation Magazine*, Vol. 10, Nr. 1, 2003.
7. V. Pradeep, G. Medioni und J. Weiland, „Robot vision for the visually impaired“, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2010.

8. J. Benjamin und J. Malvern, „The new C-5 laser cane for the blind“, in *Carnahan Conference on Electronic Prosthetics*, 1973.
9. T. Schwarze und M. Lauer, „Geometry estimation of urban street canyons using stereo vision from egocentric view“, in *Informatics in Control, Automation and Robotics*. Springer International Publishing, 2015, Vol. 325.
10. ——, „Robust ground plane tracking in cluttered environments from egocentric stereo vision“, in *2015 IEEE International Conference on Robotics and Automation*, May 2015.
11. T. Schwarze und Z. Zhong, „Stair detection and tracking from egocentric stereo vision“, in *International Conference on Image Processing (ICIP)*, 2015.
12. H. Harms, E. Rehder, T. Schwarze und M. Lauer, „Detection of ascending stairs using stereo vision“, in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015.
13. A. Ess, B. Leibe, K. Schindler und L. Van Gool, „Moving obstacle detection in highly dynamic scenes“, in *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, May 2009.
14. H. Badino, U. Franke, C. Rabe und S. Gehrig, „Stereo Vision-Based Detection of Moving Objects under Strong Camera Motion“, in *Proceedings of the First International Conference on Computer Vision Theory and Applications*, Vol. 2, February 2006.
15. A. Colombari, A. Fusiello und V. Murino, „Segmentation and tracking of multiple video objects“, *Pattern Recognition*, Vol. 40, Nr. 4, 2007.
16. A. Yilmaz, O. Javed und M. Shah, „Object tracking: A survey“, 2006.
17. Y. Bar-Shalom, *Tracking and Data Association*. Academic Press Professional, Inc., 1987.
18. A. Geiger, J. Ziegler und C. Stiller, „Stereoscan: Dense 3d reconstruction in real-time“, in *IEEE Intelligent Vehicles Symposium*, 2011.
19. M. Romanovas, T. Schwarze, M. Schwaab, M. Traechtler und Y. Manoli, „Stochastic cloning Kalman filter for visual odometry and inertial/magnetic data fusion“, in *16th International Conference on Information Fusion*, 2013.
20. E. Kraft, „A quaternion-based unscented Kalman filter for orientation tracking“, in *Information Fusion, 2003. Proceedings of the Sixth International Conference of*, Vol. 1, July 2003.
21. S. Roumeliotis und J. Burdick, „Stochastic cloning: a generalized framework for processing relative state measurements“, in *ICRA '02 IEEE International Conference on Robotics and Automation*, Vol. 2, 2002.
22. J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. Cambridge, Mass. MIT Press, 1997.

Quantifizierung der geometrischen Eigenschaften von Schmelzzonen bei Laserschweißprozessen

Quantification of geometric properties of melting zones in laser welding processes

Danny Kowerko[1], Marc Ritter[2], Robert Manthey[1], Björn John[3] und Michael Grimm[3]

[1] Technische Universität Chemnitz,
Juniorprofessur Media Computing, 09107 Chemnitz
[2] Hochschule Mittweida – University of Applied Sciences,
Professur Medieninformatik, 09648 Mittweida
[3] 3D-Micromac AG, Technologie-Campus 8, 09126 Chemnitz

Zusammenfassung Laserschweißprozesse werden zwar mit Bildraten in Abhängigkeit der Auflösung von bis zu 2.400 fps in Echtzeit verfolgt, Aussagen zu den Prozessvorgängen sind jedoch nur bedingt exakt; besonders in Hinblick auf die Analyse neuer Parameter und deren Wirkung. Dieser Beitrag nutzt neue Kameratechnologie mit hohen Frameraten und bestimmt durch eine Kombination aus modellbasierter Bildverarbeitung und anschließender Datenverarbeitung kinetische und geometrische Abhängigkeiten der Schmelzzone zu den Geräteparametern.

Schlagwörter Laserschweißen, modellbasierte Bildverarbeitung, Schmelzzone.

Abstract Laser welding processes are tracked in real-time at high rates of up to 2,400 fps. Due to low resolution process surveillance is limited in accuracy especially w. r. t. the analysis of new parameters and their influence on the melting zone. This contribution exploits camera technology with extremely high frame rates and combines it with model based image and subsequent data processing to obtain kinetic and geometric dependencies of the melting zone w. r. t. underlying hardware parameters.

Keywords Laser beam welding, model-based image processing, melting zone.

1 Motivation

Die 3D-Micromac AG ist auf dem Gebiet der Lasermikrobearbeitung tätig und entwickelt dafür Verfahren, Maschinen und komplette Anlagen. Die Produkte und Dienstleistungen werden in zahlreichen Hightech-Industrien eingesetzt. Dazu gehören sowohl die Photovoltaik-, die Halbleiter- und Displayindustrie als auch die Mikrodiagnostik sowie die Medizintechnik. Besonders im letztgenannten Gebiet etablierten sich in den letzten Jahren zahlreiche Fertigungsprozesse, die das Schweißen metallischer Werkstoffe mittels Laserstrahl als festen Bestandteil beinhalten. Diese Prozesse bergen jedoch auf Grund ihrer jeweiligen Randbedingungen – explizit in Hinblick auf die Bauteildimensionen im Bereich der Lasermikrobearbeitung – völlig neue und vor allem spezifische Herausforderungen in sich. Häufig liegen die Werkstoffdicken der verwendeten Halbzeuge im Bereich von einigen $10\,\mu$m bis einigen $100\,\mu$m.

Eine erfolgreiche Umsetzung von Fügeprozessen ist dabei nur durch die Verwendung eines auf die Anforderungen abgestimmten Laserschweißsystems in Kombination mit einer genauen Prozessanalyse möglich. Da bei der Lasermaterialbearbeitung verschiedene Faktoren wie Laser-, Maschinen-, Werkstoff- und Werkzeugparameter unterschiedliche Einflüsse auf das Fügeergebnis ausüben [1], wird es durch eine geschickte Manipulation dieser Randbedingungen möglich, die Qualität der erzeugten Verbindung zu steuern. Die Analyse der sich einstellenden positiven oder negativen Effekte durch die Veränderung der jeweiligen Parameter erfolgt meist stichprobenartig durch nachgelagerte zerstörungsfreie und/oder zerstörende Prüfmethoden (z. B. Dauerschwingversuche, Metallographie, Röntgenuntersuchungen, Wirbelstromprüfung etc.).

Einen anderen Punkt in diesem Zusammenhang liefert die Analyse der Prozesse mittels Hochgeschwindigkeitsvideoaufnahmen mit über 2.000 Bildern pro Sekunde bei einer Auflösung von 1024×768 Pixel, wobei die Schmelzzone im Fokus der Betrachtungen steht. Infolgedessen sind ebenfalls Aussagen zum Prozessverhalten und somit Rückschlüsse auf die Qualität der Fügeverbindung möglich. Dies stellt sogar in Hinblick auf eine industrielle (Serien-)Bearbeitungen im Mikrometerbereich meist die einzige Möglichkeit der Qualitätseinschätzung dar, da andere Verfahren zu grobauflösend oder mit einem hohen zeitlichen

Aufwand verbunden sind. Die softwarebasierte Aufarbeitung und Analyse der auflaufenden Menge an Videodaten bildet dabei ein zentrales Element der Qualitätssicherung.

Dieser Beitrag verfolgt das Ziel, mit Methoden der Bildverarbeitung die Auswirkungen unterschiedlicher Randbedingungen auf die Schmelzzone und somit auf die Fügeverbindung zu erfassen und zu analysieren. Das Hauptaugenmerk liegt dabei auf der Charakterisierung des Prozesses als Funktion der Zeit, des Ortes und spezifischer Parameter wie z. B. Frequenz und Volumenstrom des Prozessgases. Der angestrebte Ansatz nutzt das optische Erscheinungsbild der Schmelzzone, das sich aufgrund der perspektivischen Verzerrung durch eine Ellipse mit zeitlich variabler räumlicher Ausdehnung approximieren lässt. Das ermöglicht, den Einfluss der unterschiedlichen Parameter auf die Schmelzzone zu charakterisieren und Prozessabhängigkeiten aufzuzeigen.

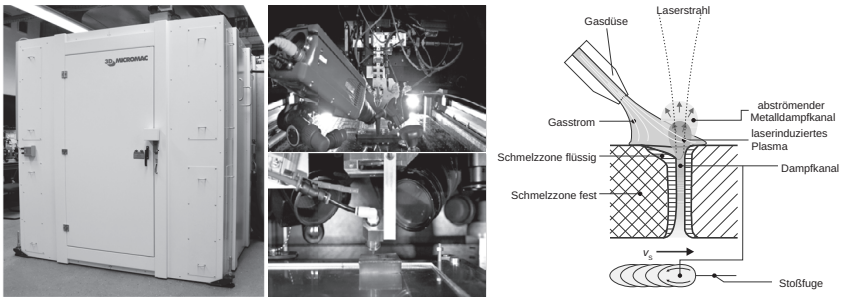


Abbildung 1: Links: Modulares Laserbearbeitungszentrum „microWELD“ mit Positionier- und Wiederholgenauigkeiten von unter $10\,\mu$m bei Geschwindigkeiten von größer $1\,\mathrm{m}\cdot\mathrm{s}^{-1}$. Mitte: Front- und Rückansicht des Versuchsaufbaus mit integrierter Hochgeschwindigkeitskamera und Beleuchtungssystem. Rechts: Schematische Darstellung des Tiefschweißprozesses mit lateral angeordneter, dem Prozess nachlaufender Schutzgasdüse.

2 Stand der Technik

Im Gegensatz zu nachgelagerten zerstörenden/zerstörungsfreien Werkstoffprüfungen bieten namhafte Lieferanten, wie z. B. die Trumpf GmbH & Co. KG als Hersteller von Laserkomplettsystemen oder die

Precitec GmbH & Co. KG (System- und Komponentenlieferant für laserspezifische Applikationen) unterschiedliche Systeme für eine Online-Überwachung von Laserfügeprozessen an. Mit deren Hilfe lassen sich ebenso quantitative sowie qualitative Aussagen zu jedem einzelnen Prozessschritt treffen. Über unterschiedliches Messequipment werden zahlreiche Daten erfasst, in der Auswertesoftware verarbeitet und protokolliert. Zudem besteht die Möglichkeit, entsprechende Regelgrößen aus den Messungen abzuleiten und direkt in den Prozess mit einfließen zu lassen [2]. Dies geschieht unter der Verwendung von in den Bearbeitungskopf der Anlage integrierten Highspeed-Kameras sowie entsprechende Triangulationsmesssysteme oder anderen Arten von Sensoren. Sie erlauben eine vollständige Überwachung der Prozesse sowohl hinsichtlich ihrer Fügestelle (Pre-Prozess), ihres Brennflecks (In-Prozess) als auch ihrer Schweißnaht (Post-Prozess) [3]. Der damit verbundene zeitliche/kostenintensive Aufwand entsprechender Nachprüfungen sinkt durch die Verwendung solcher Systeme deutlich und empfiehlt sie vor allem für Großserienanwendungen. Der Anwender

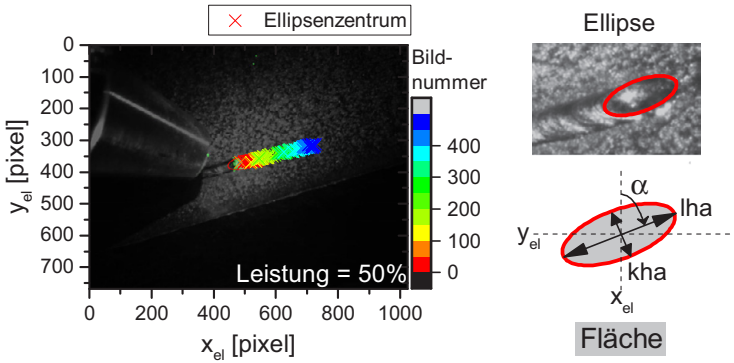


Abbildung 2: Die Nahaufnahme der Laserschweißzone zeigt die Schutzgasdüse, die Probe und einen Teil des entstandenen Schweißkanals. Die farbigen Kreuze schematisieren den im gezeigten Hochgeschwindigkeitsvideo nachfolgenden zeitlichen Verlauf des Schmelzzonenzentrums. Dieses wird, wie rechts im Bild schematisiert, durch den Mittelpunkt einer an die Schmelzzone approximierten Ellipse definiert. Die Doppelpfeile markieren die doppelte Länge der kleinen und großen Ellipsenhalbachsen (kha und gha). Der Abstand zwischen Düse und Schmelzzone bleibt dabei konstant.

wird somit in die Lage versetzt, hochpräzise qualitative Aussagen über nahezu 100 % seiner produzierten Fügeverbindungen treffen zu können.

Diese Systeme sind somit äußerst effektiv und bis zu einem gewissen Grade lernfähig, basieren jedoch auf dem „klassischen" Verhalten der Laserschweißprozesse. Dies spiegelt sich auch in der Auslegung der einzelnen Komponenten wider mit Bildraten von 333 bis 2.400 fps bei typischen Auflösungen weit unter 1 Mpixel. Für eine exakte Analyse erscheint dies zu gering.

Im Rahmen eines Forschungsprojekts, das in der „KMU-innovativ-Initiation" des BMBFs angesiedelt ist, wurde eine Technologie zum werkstoffunabhängigen Spannen von Folien durch Gasimpulse für das Lasermikrofügen durch die 3D-Micromac AG und die Professur Schweißtechnik der TU Chemnitz entwickelt. Infolgedessen konnten signifikante Effekte der Prozessbeeinflussung bei verschiedenen laserbasierten Materialbearbeitungsprozessen festgestellt werden. Unter Zuhilfenahme der Hochgeschwindigkeitsvideoanalyse des Prozesses sollen nachfolgend die Einflüsse der Parameter erfasst und dokumentiert werden.

3 Verfahren

Nachstehend werden die Versuchsanordnung mit Hinblick auf die Schmelzzonenentstehung und eine entsprechende Bildverarbeitungskette zur Quantifizierung der Schmelzzone genauer erläutert.

3.1 Aufbau und Spezifikation des Laserschweißmoduls

In der Fachliteratur ist die Lasermaterialbearbeitung durch die Eigenschaften der Laserstrahlung charakterisiert, welche ausgehend von der Strahlquelle über die Strahlführung und durch geeigneten Optiken auf dem Werkstück fokussiert wird, um dort unterschiedliche Prozesse zu initiieren [4, 5]. In Abhängigkeit der Intensität (I_L; Formel (1)) und deren Einwirkzeit auf das Werkstück kann sich eine Schmelzzone ausbilden. Vor diesem Hintergrund lassen sich Verbindungen zwischen unterschiedlichen Werkstücken realisieren.

$$I_\mathrm{L} = \frac{\mathrm{Laserleistung}}{\mathrm{Strahlquerschnittsflaeche}} \tag{1}$$

Überschreitet die Strahlungsintensität die Verdampfungtemperatur des Werkstückmaterials (z. B.: $I_{\mathrm{L;Stahl}} > 1 \cdot 10^6\,\mathrm{W \cdot cm^{-2}}$), entsteht im Zentrum der Strahlquerschnittsfläche in der Schmelzzone eine Dampfkapillare, auch Keyhole genannt. Ab diesem Punkt ist dies gemäß der Fachliteratur als Tiefschweißprozess bekannt (vgl. schematische Darstellung in Abb. 1 rechts). Durch die Bewegung des Laserstrahls und somit der Dampfkapillare relativ zum Bauteil wird die eigentliche Schweißverbindung realisiert. Das Material entlang der Vorderfront der Kapillare wird aufgeschmolzen, umströmt sie und erstarrt an der Rückseite zur Schweißnaht. Folglich ermöglicht diese Prozessvariante ein hohes Aspektverhältnis (1:10) von Nahtbreite zu Nahttiefe.

Die Umsetzung dieser Art von Materialbearbeitungsprozessen geschieht auf extra hierfür konzeptionierten Anlagen, wozu auch das modulare Laserbearbeitungszentrum „microWELD“ der Firma 3D-Micromac AG gehört. Dahinter verbirgt sich ein Maschinenkonzept mit obenliegenden Gantry-Achssystem, das in der Lage ist, Positionier- und Wiederholgenauigkeiten von unter $10\,\mu$m bei Geschwindigkeiten von größer $1\,\mathrm{m \cdot s^{-1}}$ zu realisieren. Das System kann dabei in Abhängigkeit der Applikation mit einer Festoptik oder einem Scannersystem ausgestattet werden. Ebenso ist auch die zum Einsatz kommende Strahlquelle nicht an das System gebunden, sondern frei wählbar. Im Rahmen der durchgeführten Untersuchungen zur Lokalisierung und geometrischen Approximation der Schmelzzone wurde auf eine Single-Mode-Faserlaserquelle ($\lambda = 1.064\,\mathrm{nm}$) der Firma IPG Photonics mit einer Laserleistung von bis zu 1 kW zurückgegriffen. Die Aufnahmen der Hochgeschwindigkeitsvideoanalyse wurden mit Hilfe einer frei in den Maschinenraum der Anlage integrierbaren Kamera des Typs Phantom v310 realisiert, die eine maximale Bildrate in Abhängigkeit der Auflösung von bis zu 500.000 fps erlaubt. In Kombination mit dem Belichtungssystem der Firma Cavitar Ltd, welches gesteuert über die Kamera mittels gepulstem Diodenlaser das Objekt beleuchtet, konnten hochgenaue Standbilder (vgl. Abb. 2) der Prozessvorgänge in einem Intervall von $500\,\mu$s ermittelt und der Analyse zur Verfügung gestellt werden. Sie bilden die Grundlage für die nachfolgenden geometrischen Untersuchungen, die durch eine Reihe von kaskadierten Methoden der industriellen Bildverarbeitung bestimmt werden. Hierzu werden insgesamt 95 Hochgeschwindigkeitsbildsequenzen mit bis zu 1.000 Bildern und einer Auflösung von 1024×768 Pixel mit einem Rohdatenvolumen von 127 GB

bzw. von 1,4 GB im Kompressionsformat H.264 herangezogen. Darunter sind anteilig zwei zeitlich versetzt aufgenommene Kameraperspektiven der gleichen Parametereinstellungen sowie vorwärts und rückwärts gerichtete Schweißprozesse enthalten.

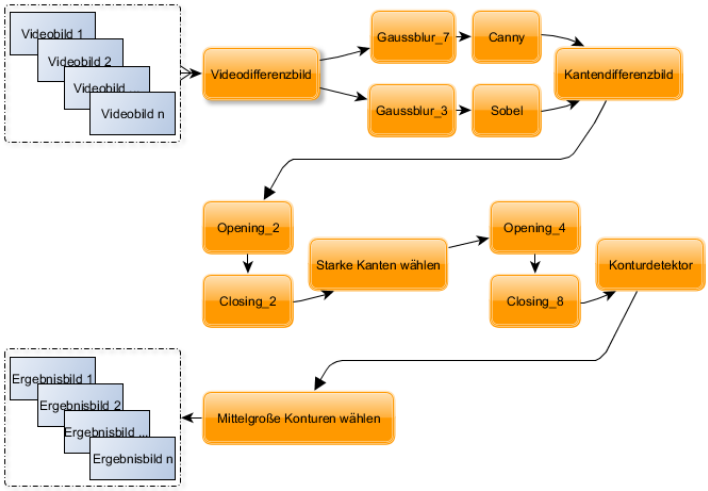


Abbildung 3: Schematische Darstellung des Verfahrens zur geometrischen Approximation der Schmelzzone im Eingangsvideo. Die Differenz zweier gegebener konsekutiver Einzelbilder bildet die Datenbasis für Weichzeichenoperatoren verschiedener Größen und ihrer zugehörigen Kantendetektoren, wodurch Rauschen unterdrückt und die signifikantesten Kanten detektiert werden. Durch Fusion dieser beiden Elementmengen bildet sich ein *Kantendifferenzbild*, das durch morphologische Operationen und die Auswahl der stärksten Kanten geglättet und bereinigt wird. Mittels des *Konturdetektors* erfolgt die Bestimmung der inhaltsbeschreibenden Ellipsen, die nach erneuter Filterung das Ergebnis darstellen.

3.2 Quantitative Analyse der Schmelzzone mittels Bildverarbeitung

Die Schmelzzonenanalyse unterteilt sich in einen bildverarbeitenden Teil, in welchem die Schmelzzone zunächst lokalisiert und durch eine Ellipse geometrisch approximiert wird. Darauf aufbauend werden die

Ellipsendaten im 2. Teil als Funktion der Zeit weiterverarbeitet und anschließend im Kontext der Geräteparameter dargestellt und diskutiert.

Die Lokalisierung der Schmelzzone erfolgt durch Methoden der Bildverarbeitung entsprechend dem in Abb. 3 dargestellten Verfahren und der Bildverarbeitungsbibliothek OpenCV [6]. Das bei der Versuchsdurchführung aufgenommene Video wird hierzu zunächst in Einzelbilder unterteilt und Differenzbilder berechnet, da der Bereich der Schmelzzone durch deutliche Veränderungen zwischen aufeinanderfolgenden Bildern hervortritt. Um das damit einhergehende kleinflächige und meist punktuell hervortretende Rauschen zu reduzieren, kommen gausssche Weichzeichner mit zwei unterschiedlichen Operatorradien von 3 und 7 Pixel und darauf abgestimmte Kantendetektoren zum Einsatz, deren Resultate in einem Kantendifferenzbild fusioniert werden.

Hierbei verbleiben nur die von beiden Detektoren erzeugten Kanten und Kantenbereiche für die anschließenden morphologischen Operatoren und eine stärkebedingte Kantenauswahl. Im Anschluss erfolgt die Entfernung von Objektkandidaten mit einer sehr kleinen Fläche oder einer geringen Helligkeit. Ebenso werden Bereiche mit großflächig unterbrochenen Kantenlinien eliminiert. Die nach diesen Filterungen verbleibenden Kandidaten stellen die Grundlage für eine geometrische Beschreibung mittels Ellipsen dar. Hier gelangt ein intellektuell über mehrere Testvideos parametrierter Selektionsprozess zum Einsatz, der kleinere oder zu große Flächen als großflächigeres Rauschen, partielle Spiegelungen und andere großflächige Bildänderungen betrachtet und entfernt. Die im Allgemeinen einzige verbleibende Ellipse bestimmt und beschreibt somit die Lage und die Eigenschaften der Schmelzzone durch ihre Geometrie. Diese Charakterisierung, visualisiert in Abb. 2, erlaubt den Verlauf der Ellipsenparameter wie die Lage des Zentrums (x_{el},y_{el}), die Länge der großen bzw. kleinen Halbachse (lha bzw. kha), den Lagewinkel α und den Flächeninhalt A als Funktion der Zeit bzw. Bildnummer darzustellen (siehe Abb. 4).

Der zeitliche Verlauf der berechneten Ellipsenparameter kann näherungsweise mittels linearer Regression mit der Funktion $y = m \cdot x + n$ angepasst werden, wobei y den jeweiligen Ellipsenparameter und x die Bildnummer repräsentiert. Aus den Fit-Parametern (m, n und Gütemaßen) lassen sich Informationen zur Geometrie und Dynamik der Schmelzzone ableiten. Eine perspektivisch bedingte Verkleinerung der Schmelzzone wird bspw. in der unteren rechten Abbildung (ebd.) sicht-

bar. Nachfolgend können diese als Funktion von Geräteparametern, die zwischen den Videos variieren, ausgewertet werden. Durch Zuhilfenahme von Gütemaßen der Regression wie die mittlere quadratische Abweichung lassen sich weiterhin Rauschen bzw. die Oszillationen der Schmelzzonen innerhalb eines Videos sowie innerhalb des gesamten Korpus untersuchen.

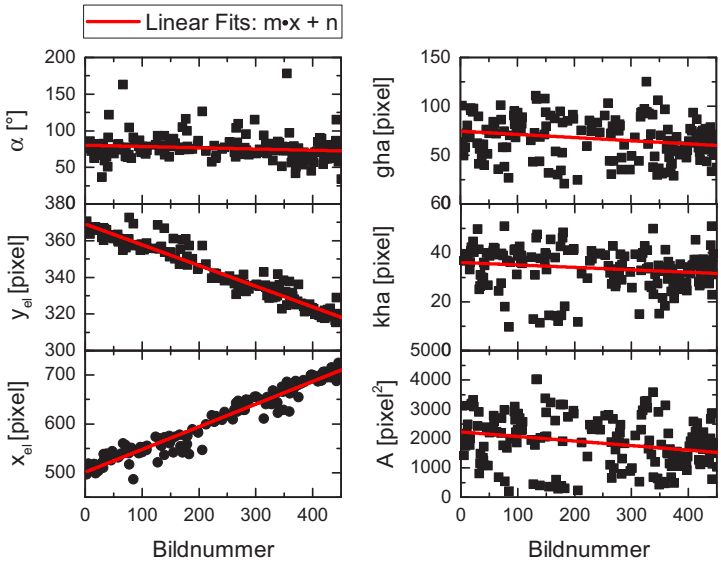


Abbildung 4: Zeitliche Verläufe der Ellipsenparameter: Winkel, Zentrum, Halbachsen und Fläche (von links oben nach rechts unten) und deren lineare Regression (rote durchgehende Linien).

3.3 Schmelzzonengeometrie als Funktion der Geräteparameter

Die zuvor dargestellte Regressionsanalyse liefert Anstiege (m) und Absolutglieder (n) für diverse Ellipsenparameter, aufgenommen unter 95 verschiedenen Geräteparametereinstellungen. Aus der in Abb. 4 (links) dargestellten Bewegung des Ellipsenzentrums lässt sich dessen mittlere Geschwindigkeit aus dem Anstieg der Regressionsgerade $m_{x_{el}}$ berechnen, was für die x-Koordinate in Abb. 5 beispielhaft dargestellt ist.

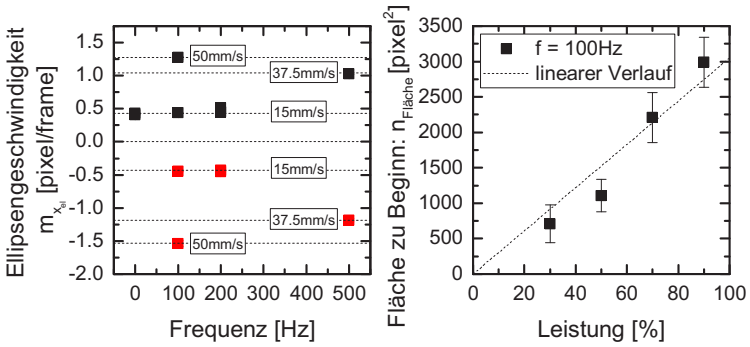


Abbildung 5: Ellipsenparameter als Funktion der Geräteparameter: Links: Mittlere Ellipsenzentrumsgeschwindigkeit bei verschiedenen Inertgaspuls-Frequenzen und Düsengeschwindigkeiten und unterschiedlicher Düsenbewegungsrichtung (rot/schwarz entspricht der Schmelzzone vor-/nachlaufend). Rechts: Die mittlere Ellipsenfläche zu Beginn des Experiments wächst linear mit zunehmender Laserleistung. Die Düsenbewegungsrichtung ist ausschließlich nachlaufend.

Diese Werte lassen sich als Schmelzzonen- bzw. Düsengeschwindigkeit auffassen und dienen der Kalibrierung zwischen realer Düsen-/Schmelzzonenbewegung und deren Repräsentation im jeweiligen Video. Die Richtungsinformation der Schmelzzone spiegelt sich im Vorzeichen wider und stellt eine Kontrolle für die manuell referenzierten Geräteparameter dar. Ein zusätzliches Tracking der Düse zur Geschwindigkeitsbestimmung des Schmelzprozesses ist daher nicht zwingend erforderlich. Absolute Schwankungen der Schmelzzonengeschwindigkeit bei gleichen Düsengeschwindigkeiten, aber unterschiedlichen Sekundärparametern können ein Hinweis auf tiefenperspektivische Effekte sein, beispielsweise wo nicht der exakt gleiche Verfahrweg gegangen wurde und sich der Abstand zur Kamera minimal verändert.

Das Absolutglied der Geradengleichung der Kurvenanpassung an den Ellipsenflächeninhalt $n_{Flaeche}$ erlaubt eine Modellierung der Schmelzzonengröße als Funktion der Laserleistung (siehe Abb. 5 rechts). Die vier Messpunkte stammen aus vier jeweils verschiedenen Videos, aufgenommen bei unterschiedlicher Leistung bei sonst konstanten Geräteparametern. Durch die Regression werden Störeffekte

wie Rauschen und perspektivisch bedingte Größenänderungen nahezu eliminiert. Es zeigt sich eine nahezu lineare Abhängigkeit zwischen Schmelzzonenfläche zu Beginn eines Experiments $n_{Flaeche}$ und der verwendeten Laserleistung. Die Unsicherheit der Regressionsparameter, dargestellt rechterhand als Fehlerbalken in Abb. 5 für $n_{Flaeche}$, erlaubt Rückschlüsse über die allgemeine Stärke des Rauschens in den Daten, das sich zudem über alle Parameterabhängigkeiten erstreckt und in Abb. 4 verdeutlichen lässt. Die Zunahme des absoluten Fehlers der Ellipsenflächenbestimmung, repräsentiert durch Fehlerbalken, korreliert hier jedoch mit einer Abnahme der relativen Fehler von $> 35\,\%$ auf etwa $10\,\%$ bei zunehmender Fläche. Wenngleich es eine physikalische Ursache dafür geben könnte, kann es auch ein Hinweis darauf sein, dass die räumliche Auflösung der Schmelzzone für die Genauigkeit dieser Bildverarbeitungskette eine wichtige Rolle spielt. Um physikalische und bildverarbeitungsbedingte Prozesse voneinander zu entfalten, sind perspektivische Geräteparameter und optischer Zoom aufeinander abzustimmen. Perspektivisch sind die hier gezeigten Abhängigkeiten verwendbar, um die Schmelzzonengröße im Bild auch bei unterschiedlichen Geräteparametern möglichst konstant zu halten, z. B. durch angepasste Bildvergrößerungen (Zoom) der sich im Aufbau befindlichen Kamera.

4 Zusammenfassung

Mittels eines mehrstufigen OpenCV-basierten Bildverarbeitungsverfahrens wurde die Schmelzzone in einem industriellen Laserschweißprozess geometrisch durch eine Ellipse modelliert und deren Parameter als Funktion der Zeit durch lineare Regression approxmiert. Einfache Beziehungen zu den Geräteparametern, wie z. B. der lineare Zusammenhang zwischen Schmelzzonenfläche und Laserleistung, konnten damit aufgezeigt werden. Perspektivisch sollen methodische Neuentwicklungen mit dem gewählten Ansatz untersucht werden können, wozu auch weitere Untersuchungen der Wirkung modulierter Prozessgasströme auf den Laserschweißprozess zählen. Ein mögliches Ziel besteht darin, die Charakterisierung der Schmelzzonendynamik und Spritzerbildung in Kombination mit den hier extrahierten oberflächigen Ergebnissen in Relation mit den aus Tiefenschnitten bekannten Tiefeninforma-

tionen zu setzen und somit eine computergestützte Optimierung der Geräteparameter in Bezug auf den gewünschten Tiefenschnitt zu erzeugen. Die interdisziplinäre Kooperation der beteiligten Institutionen erlaubt es, eine größere Produktionseffizienz auf Basis dieser Ergebnisse zu erzielen und versetzt das Unternehmen somit perspektivisch in die Lage, mit seinen Technologien internationale Standards und Innovationen zu schaffen, wodurch sich Kundenwünsche auch bei komplexen Projekten mit großer Perfektion erfüllen lassen.

Acknowledgements

Diese Arbeit wurde anteilig in der vom BMBF im Rahmen von Unternehmen Region geförderten InnoProfile-Transfer-Initiative *localizeIT* (Förderkennzeichen: 03IPT608X) durchgeführt.

Literatur

1. E. U. Beske, *Untersuchungen zum Schweißen mit kW Nd:YAG-Laserstrahlung*. VDI-Verlag GmbH Düsseldorf, 1992, Nr. 257.
2. Precitec GmbH & Co. KG (Hrsg.), „WeldMaster Plattform“, abgerufen am 14.09.2016. [Online]. Available: http://www.precitec.de/produkte/fuegetechnologie/prozessueberwachung/weldmaster-plattform/
3. TRUMPF GmbH & Co. KG (Hrsg.), „Triple sensor seamline pro – applications and potential“, abgerufen am 14.09.2016. [Online]. Available: http://www.trumpf-laser.com/fileadmin/DAM/trumpf-laser.com/Events/Techday/Automotive_Photonics/Triple_Sensor_Seamline_pro_applications_and_Potential.pdf
4. R. Poprawe, „Vorlesungsunterlagen Lasertechnik I + II“, 2003, Lehrstuhl für Lasertechnik LLT RWTH Aachen.
5. R. Poprawe, *Lasertechnik für die Fertigung – Grundlagen, Perspektiven und Beispiele für den innovativen Ingenieur*. Springer-Verlag Berlin Heidelberg, 2005.
6. G. Bradski, „The OpenCV Library“, *Dr. Dobb's Journal of Software Tools*, Vol. 25, Nr. 11, S. 120–126, Nov. 2000.

Bildbasierte Überwachung alternativer Brennstoffe eines Mehrstoffbrenners bei industriellen Verbrennungsprozessen

Image-based monitoring of alternative fuels of multi-fuel burner at industrial combustion processes

Markus Vogelbacher, Patrick Waibel, Jörg Matthes und Hubert B. Keller

Karlsruher Institut für Technologie (KIT), Institut für Angewandte Informatik, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen

Zusammenfassung Die Verwendung von Mehrstoffbrennern bei industriellen Verbrennungsprozessen ermöglicht die Substitution fossiler durch alternative Brennstoffe. Der Anteil an alternativem Brennstoff kann durch eine bildbasierte Überwachung mittels Infrarotkamera und einer Online-Bildauswertung gesteigert werden. In diesem Beitrag wird das Gesamtsystem zur Überwachung von Mehrstoffbrennern vorgestellt. Hierzu zählt die Vorauswahl geeigneter Bilder, die Bildvorverarbeitung bis hin zur Detektion der Streichlinie des Brennstoffes und der Ableitung geeigneter Kenngrößen zur Weiterverarbeitung im Prozessleitsystem. Damit liefert das System Informationen über den Verbrennungszustand in Echtzeit und ermöglicht dadurch einen hohen Anteil alternativer Brennstoffe.

Schlagwörter Mehrstoffbrenner, Bildvorverarbeitung, Brennstoffdetektion, Partikeldetektion, Streichlinie.

Abstract The use of multi-fuel burners in industrial combustion processes makes it possible to replace fractions of fossil fuels by alternative fuels. These fractions can be increased by a camera-based control of the burner using infrared cameras and online image analysis. In this article the complete system is introduced for monitoring a multi-fuel burner. This includes the pre-selection of images, the image preprocessing, the detection of the fuel streakline and the derivation of appropriate parameters for further processing in the process control system. Thereby, the system provides information about the combustion state in real time and allows an increased fraction of alternative fuels.

Keywords Multi-fuel burner, image pre-processing, fuel detection, particle detection, streakline.

1 Einleitung

Die Substitution fossiler Brennstoffe (Braun- oder Steinkohle) durch sogenannte alternative Brennstoffe (Kunststoffabfälle oder Reifenflusen) ist ein wichtiger Schritt zur Energiekostenminimierung und Reduktion der Umweltbelastung im Bereich industrieller Verbrennungsprozesse. Bei der Zementherstellung in Drehrohrverbrennungsanlagen werden aus diesem Grund zur Bereitstellung der notwendigen thermischen Energie vermehrt Mehrstoffbrenner (Abb. 1) eingesetzt. Diese ermöglichen es, verschiedene Arten an alternativen Brennstoffen in vorgegebenen Anteilen zu fossilen Brennstoffen mit zu verbrennen.

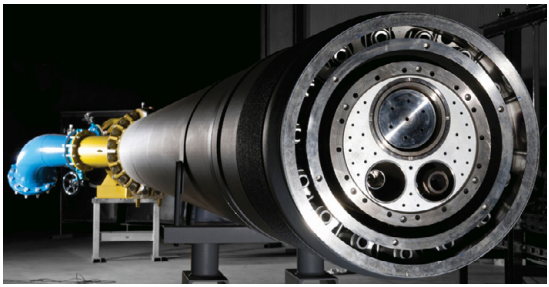


Abbildung 1: Mehrstoffbrenner der Firma *Unitherm Cemcon* [1].

Im Gegensatz zu fossilen Brennstoffen kann die Qualitat der alternativen Brennstoffe etwa durch unterschiedliche Feuchtigkeit oder Partikelgröße stark schwanken. Dies kann zu variierenden Verbrennungszeitpunkten, unterschiedlicher Streuung des Brennstoffes oder zur Einbringung von nicht verbranntem Brennstoff in das zu verarbeitende Material führen und damit direkt den Energieeintrag und auch die Produktqualität beeinflussen. Um einen konstanten Betrieb unter hohen Anteilen an alternativen Brennstoffen gewährleisten zu können, ist daher eine dauerhafte Überwachung des Verbrennungsverhaltens des alternativen Brennstoffes notwendig. Änderungen im Verbrennungsverhalten können dann durch Anpassungen der Brennereinstellungen ausgeglichen werden. Die schnelle Reaktion auf Änderungen kann durch momentan an Anlagen existierende Überwachungssysteme wie etwa die Auswertung von Laborwerten oder Kamerasystemen im visuellen oder nahinfraroten Wellenlängenbereich nicht gewährleistet werden. Zum einen ist die Auswertung von Laborwerten sehr langsam und nur zeitversetzt möglich und zum anderen sind die Brennstoffbestandteile in den vorhandenen Kamerasystemen nicht erkennbar. In diesem Beitrag wird daher ein Gesamtsystem bestehend aus Infrarotkamera und anschließender Bildauswertung vorgestellt, das eine Online-Überwachung des Brennstoffes und durch die Ausgabe von Kenngrößen eine frühzeitige Anpassung der Brennereinstellungen ermöglicht.

Der Einsatz von Infrarotkameras bei industriellen Verbrennungsprozessen ist aus anderen Aufgabenstellungen bekannt [2–5]. Bei der Überwachung von Mehrstoffbrennern erlaubt die Betrachtung eines speziellen infraroten Wellenlängenbereiches einen Blick durch störende Verbrennungsgase hindurch auf den interessierenden Brennstoff.

Vorhandene Verfahren zur Beurteilung einer Brennerflamme existieren bisher nur im visuellen Wellenlängenbereich. Dabei wird der Verbrennungsprozess auf Grundlage der Bildintensitäten beurteilt [6, 7] oder bestimmte geometrische oder grauwertbasierte Flammeneigenschaften detektiert [8–12]. Eine direkte Detektion des Brennstoffes ist durch diese Verfahren nicht möglich.

Im Folgenden wird zunächst in Abschnitt 2 die Aufnahmekonstellation zur Überwachung von Mehrstoffbrennern bei der Zementherstellung in Drehrohrverbrennungsanlagen vorgestellt. Abschnitt 3 beschreibt den Ablauf der neuen Bildauswertung basierend auf Infrarot-

aufnahmen, deren Ergebnisse in Abschnitt 4 vorgestellt werden. Eine Zusammenfassung und ein Ausblick folgen in Abschnitt 5.

2 Infrarotaufnahmen bei der Drehrohrverbrennung

In Drehrohröfen werden Materialien unter hohen Temperaturen und kontinuierlicher Drehbewegung in ein gewünschtes Endprodukt umgewandelt. Das Aufgabematerial durchläuft hierbei durch das bis zu $200\,m$ lange, in Längsrichtung leicht geneigte Drehrohr unter kontinuierlicher Durchmischung verschiedene Wärmezonen. Die notwendige thermische Energie wird durch Brenner am Beginn oder Ende des Drehrohres bereitgestellt.

Für den hier betrachteten Prozess der Zementherstellung wird ein Mehrstoffbrenner am Ende des Drehrohres verwendet. Die Möglichkeiten zur kamerabasierten Überwachung dieses Mehrstoffbrenners sind aus baulichen und thermischen Gründen stark eingeschränkt. Eine geeignete Kameraposition befindet sich am Ende des Drehrohres, dem Ofenkopf. Die Kamera wird neben dem Mehrstoffbrenner mit Blick in den Ofen und mit seitlicher Sicht auf den Brenner angebracht (Abb. 2). Durch Verwendung einer Infrarotkamera im Wellenlängenbereich von $10{,}6\,\mu m$ werden die verschiedenen Brennstoffe am Austritt des Brennermundes sichtbar (Abb. 3(a)). Der alternative Brennstoff tritt hierbei aus

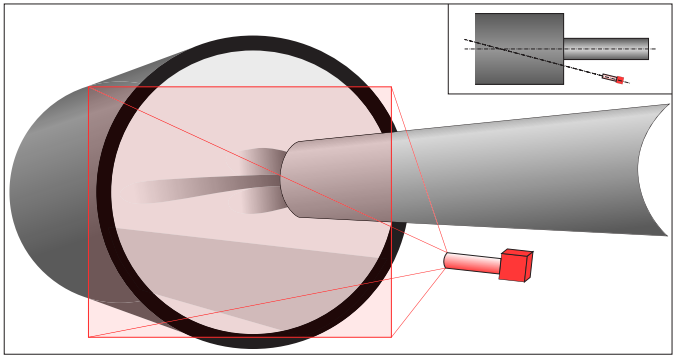
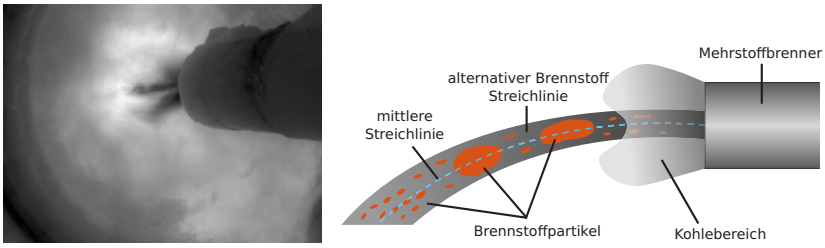


Abbildung 2: Position für den Kameraeinbau zur Brennerüberwachung (Seitenansicht und Aufsicht in der kleineren Abbildung).

dem Zentrum des Brennermundes aus und wird koaxial vom fossilen Brennstoff (Kohle) umgeben. Auf Grund des kürzeren Verbrennungszeitpunktes der Kohle tritt die Streichlinie des alternativen Brennstoffes, die den Aufenthaltsort des Brennstoffes beschreibt, aus dem Kohlebereich aus. In Abb. 3(b) kann der zu erwartende Aufenthaltsort der Brennstoffe anhand einer schematischen Seitenansicht nachvollzogen werden.



(a) (b)

Abbildung 3: (a) Infrarotaufnahme eines Mehrstoffbrenners im Drehrohrofen und (b) die schematische Seitenansicht zur Erläuterung des Bildinhaltes.

3 Bildauswertung zur Brennstoffüberwachung

Auf Basis der in Abschnitt 2 gewonnenen Infrarotbilder wird eine automatische Bildauswertung durchgeführt. Abb. 4 zeigt den Gesamtablauf des Bildverarbeitungsalgorithmus. Die einzelnen Schritte werden nachfolgend erläutert.

3.1 Überprüfung Kameraposition

Der Austrittspunkt der Brennstoffe wird als bekannt angenommen und dient in den weiteren Verarbeitungsschritten als Ausgangspunkt für die mittlere Streichlinie. Treten ungewollte Änderungen der Kamera- oder Brennerposition auf, würde dies zu einer fehlerhaften Messung führen. Aus diesem Grund wird im ersten Schritt die Brennerposition in Bildkoordinaten detektiert und auf Abweichung überprüft. Um lokale Störungen zu unterdrücken, wird hierzu zunächst ein zeitlicher Tiefpass auf eine Bildsequenz angewandt. Auf das gefilterte Bild wird

eine Kantendetektion und anschließend ein Template Matching mit einem geeigneten Template der Brennerspitze durchgeführt. Die Position mit der größten Übereinstimmung wird als Brennerposition ausgegeben. Solange diese innerhalb eines Toleranzfensters liegt, kann die nachfolgende Auswertung fortgesetzt werden. Bei zu großer Abweichung ist eine manuelle Neueingabe der aktualisierten Position für die weiteren Verarbeitungsschritte notwendig.

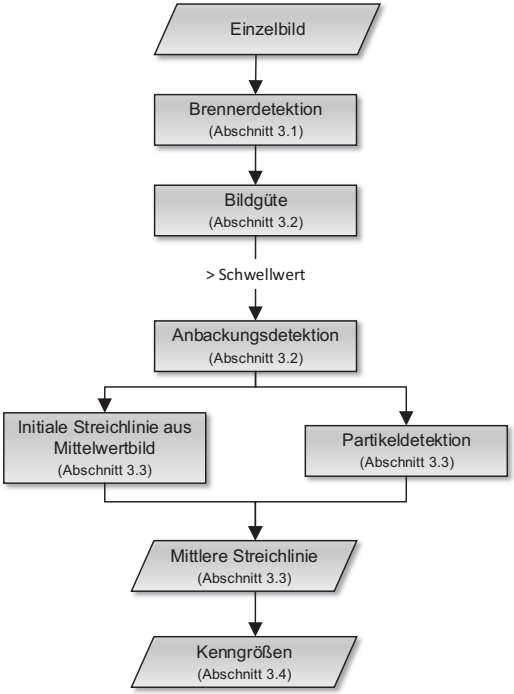


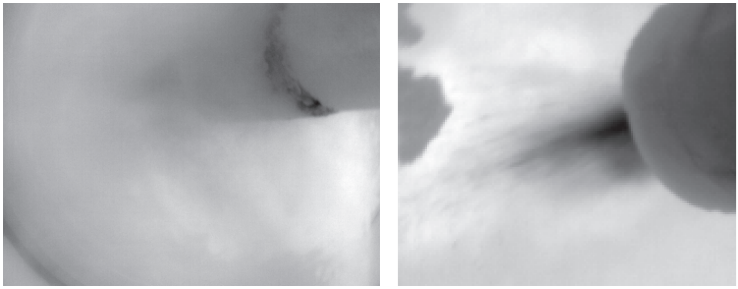
Abbildung 4: Ablauf der Bildverarbeitung zur Brennerüberwachung.

3.2 Bildvorauswahl und -vorverarbeitung

Zu starke Staubbeladung im Drehrohr kann zu Verdeckungen des Brennstoffbereiches am Brennermund führen (Abb. 5(a)). Für eine sinnvolle Auswertung müssen diese Bilder durch eine Vorauswahl aussor-

tiert werden. Durch die Abdeckung entstehen in den Bereichen mit schlechter Sicht Regionen mit homogenen Grauwerten. Zur Beurteilung der Sichtbarkeit wird daher der mittlere Gradient im Bereich, in dem Brennstoff zu erwarten ist, als Kontrastmaß eingeführt. Über einen Schwellwert können damit Bilder, die keine Information über den Brennstoffaufenthaltsort enthalten, aussortiert werden.

Neben der Verdeckung durch eine starke Staubbeladung können auch so genannte Anbackungen Teile des Brennstoffes überdecken und damit auch zu Fehldetektionen führen (Abb. 5(b)). Anbackungen entstehen durch sich anhäufendes anhaftendes Material an der Drehrohrwand und weisen eine ähnlich niedrige Temperatur wie Brennstoffpartikel auf. Für ein robusteres Ergebnis der Brennstoffdetektion werden diese Anbackungen segmentiert und für die weitere Auswertung ignoriert. Durch Extraktion des Grauwertverlaufes entlang des Bildrandes können Temperatureinbrüche durch eine lokale Minimasuche detektiert werden. Ausgehend von diesen Temperatureinbrüchen wird durch ein Region Growing Verfahren und eine anschließende morphologische Nachbearbeitung eine Segmentierung der Anbackungen erhalten (Abbildung 6).



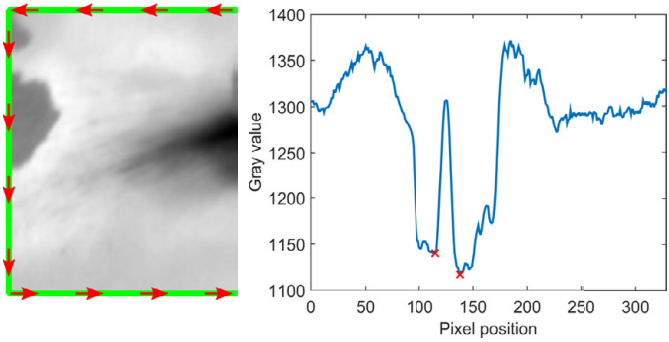
(a) (b)

Abbildung 5: (a) Infrarotaufnahme eines Mehrstoffbrenners bei starker Staubbeladung und (b) Verdeckung des Bildausschnittes durch Anbackungen.

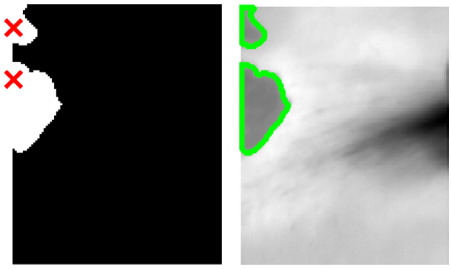
3.3 Streichlinie des Brennstoffes

Nach Vorauswahl der Bilder und Festlegung der Anbackungsbereiche, die für die weitere Auswertung nicht berücksichtigt werden, erfolgt die

eigentliche Bildauswertung zur Überwachung des alternativen Brennstoffes. Ziel ist es, die mittlere Streichlinie des Brennstoffes zu bestimmen. Eine Streichlinie beschreibt in der Strömungslehre den Pfad mehrerer Partikel, die dieselbe Startposition in einem Strömungsfeld besitzen. Die mittlere Streichlinie beschreibt damit die Hauptflugbahn des Brennstoffes und lässt somit eine Einschätzung für das aktuelle Verbrennungsverhalten des Brennstoffes zu. Zur Detektion der Streichlinie werden zwei Verfahren verwendet, die die unterschiedlichen Charakteristiken der Brennstoffflugbahn berücksichtigen.



(a)



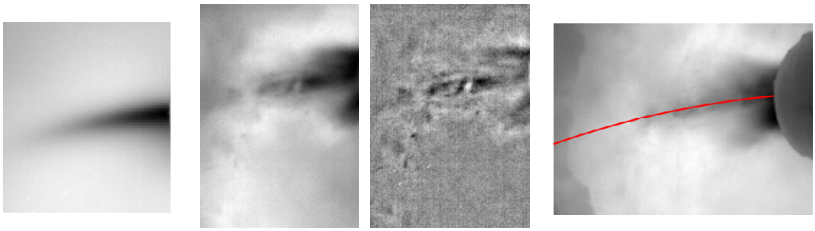
(b) (c)

Abbildung 6: Ablauf der Anbackungsdetektion: (a) Extraktion des Grauwertverlaufes entlang der Bildkante und die Detektion von Minima (rote Kreuze). (b) Ergebnis der Segmentierung nach dem Region Growing Verfahren und einer morphologischen Nachbearbeitung (rote Kreuze = Saatpunkte des Region Growing Verfahren aus der Minimadetektion). (c) Segmentierungsergebnis im Originalbild.

Nahe am Brennermund ist der Brennstoff noch stark gebündelt und es bildet sich ein dunkler Brennstoffschweif. Durch zeitliche Mittelwertfilterung kann dieser Effekt noch weiter verstärkt werden (Abb. 7(a)). Eine spaltenweise Minimumsuche auf diesem Mittelwertbild ermöglicht eine erste Schätzung des Aufenthaltortes des Brennstoffes in jeder Bildspalte und damit eine erste Schätzung der mittleren Streichlinie als Parabel zweiten Grades, die durch die Mitte des Brennermundes verlaufen muss.

Je weiter der Brennstoff sich vom Brennermund entfernt, desto stärker erfolgt eine Auffächerung des Brennstoffes, und einzelne Brennstoffpartikel treten zum Vorschein. Aus diesem Grund werden im zweiten Verfahren die Brennstoffpartikel detektiert. Zur Verstärkung kontrastreicher Bildbereiche wird zunächst eine Bildverbesserung mittels des Retinex Verfahrens [13] durchgeführt (Abb. 7(b)). Mit Hilfe von Laws Texturfilter (Level-Spot) können Bereiche, in denen Partikel zu erkennen sind, verstärkt werden. Eine Schwellwertsegmentierung des gefilterten Bildes liefert letztendlich die detektierten Partikelregionen.

Mit Hilfe der Partikeldetektionen wird dann im nächsten Schritt die initiale Streichlinie aktualisiert (Abb. 7(c)). Je mehr Partikeldetektionen vorliegen, desto stärker wird der Einfluss des Partikeldetektionsverfahrens. Dadurch erfolgt eine automatische Gewichtung beider Verfahren je nach aktuellen Eigenschaften des Brennstoffes (starke Streuung, d. h. viele Partikel sichtbar oder kompakter Brennstoffstrahl).



(a) (b) (c)

Abbildung 7: (a) Verdichtung des Brennstoffschweifes durch zeitliche Mittelwertfilterung. (b) Bildverbesserung mit Retinex Verfahren (Original links, Retinexbild rechts). (c) Mittlere Streichlinie einer Beispielsequenz.

3.4 Kenngrößenableitung

Aus der mittleren Streichlinie lassen sich unterschiedliche Kenngrößen zur Beschreibung des Verbrennungsverhaltens des alternativen Brennstoffes ableiten. Die Steigung der Streichlinie nahe am Brennermund liefert zum Beispiel eine Aussage über den Austrittswinkel des Brennstoffes. Außerdem kann mit Hilfe der Streichlinie auch der Auftreffpunkt von unverbranntem Brennstoff auf das Materialbett im Drehrohr berechnet werden. Damit kann verhindert werden, dass Brennstoff als zusätzlicher Reaktionspartner in Bereichen wichtiger chemischer Reaktionen eingebracht wird.

4 Ergebnisse der Bildauswertung

Zur Beurteilung der neu entwickelten Bildauswerteverfahren existieren keine Referenzdaten. Zur Bewertung der Verfahren wurden aus diesem Grund Referenzdaten manuell und mit der Hilfe von Expertenwissen erstellt.

Für die Überprüfung der Kameraposition wurden 4 unterschiedliche Brennerpositionen in 1.764 Sequenzen mit jeweils 3.000 Bilder ausgewertet. Tabelle 1 zeigt die Abweichung der detektierten Brennerposition. Das Ergebnis zeigt, dass das entwickelte Verfahren die Brennerposition robust ausgibt. Die Position wird mit einer Genauigkeit von etwa einem Pixel bestimmt.

Tabelle 1: Mittlere Abweichung der Brennerposition von den mittleren Werten über alle Sequenzen.

	Anzahl an Sequenzen	Abweichung der x Position	Abweichung der y Position
Pos. 1	1135	0.67 px	1.19 px
Pos. 2	563	0.53 px	0.86 px
Pos. 3	31	0.52 px	1.65 px
Pos. 4	35	0.51 px	0.79 px

Die Anbackungsdetektion wurde für 6 Sequenzen mit 3.000 Bildern überprüft und konnte dabei alle vorhandenen Anbackungen detektieren.

Die Vorauswahl der Bilder über das Kontrastmaß und die mittlere Streichlinie wurden mit Hilfe von Expertenwissen validiert. Die Experten zeichneten für 56 Sequenzen die mittlere Streichlinie und bewerteten die Sicht mit Schulnoten. Die Bewertung der Sicht kann mit dem skalierten Kehrwert des Kontrastmaßes verglichen werden und zeigt eine eindeutige Korrespondenz mit dem Expertenwissen. Für die mittlere Streichlinie ergibt sich die Abweichung zum Expertenwissen aus der Fläche zwischen der automatisch detektierten Streichlinie und der mittleren Streichlinie aus den Referenzdaten der Experten. Hierbei lagen 59 % der automatisch detektierten Streichlinien innerhalb der mittleren Abweichung der Experten. 89 % lagen innerhalb der maximalen Abweichung der Expertenbeurteilung. Auf Grund der schwierigen Aufnahmebedingungen zeigt sich daraus, dass die automatisch detektierte Streichlinie eine sehr gute Einschätzung für das Verbrennungsverhalten des alternativen Brennstoffes liefert.

5 Zusammenfassung und Ausblick

Das vorgestellte Verfahren zur Bestimmung der mittleren Streichlinie des alternativen Brennstoffes, bestehend aus Infrarotkamera und anschließender Bildauswertung, ermöglicht eine dauerhafte Überwachung des Verbrennungsverhaltens des Brennstoffes bei Mehrstoffbrennern. Durch die Ableitung geeigneter Kenngrößen ist eine frühzeitige Reaktion auf Veränderungen und somit ein konstanter Betrieb unter erhöhten Anteilen an alternativem Brennstoff möglich. Neben der Unterstützung des Anlagenführers durch Ausgabe der Kenngrößen auf einem Bildschirm ist es in Zukunft auch denkbar, die Kenngrößen im Prozessleitsystem für eine automatisierte Regelung zu nutzen.

Literatur

1. Unitherm Cemcon, „M.A.S. Kiln Burner – UNICAL Calciner Burner/M.A.S. GAS Burner“, Broschüre, 2015.
2. S. Zipser, A. Gommlich, J. Matthes, H. B. Keller und C. Fouda, „On the optimization of industrial combustion processes using infrared thermography“, in *23rd IASTED International Conference on Modelling, Identification and Control*, Vol. 412, Nr. 183, 2004, S. 386–391.

3. J. Matthes, P. Waibel und H. B. Keller, „A new infrared camera-based technology for the optimization of the waelz process for zinc recycling“, *Minerals Engineering*, Vol. 24, Nr. 8, S. 944–949, 2011.
4. P. Waibel, J. Matthes und H. B. Keller, „Segmentation of the solid bed in infrared image sequences of rotary kilns“, in *7th International Conference on Informatics in Control, Automation and Robotics*, 2010.
5. P. Waibel, M. Vogelbacher, J. Matthes und H. B. Keller, „Infrared camera-based detection and analysis of barrels in rotary kilns for waste incineration“, in *11th International Conference on Quantitative Infrared Thermography (QIRT)*, 2012.
6. W. B. Baek, S. J. Lee, S. Y. Baeg und C. H. Cho, „Flame image processing and analysis for optimal coal firing of thermal power plant“, in *International Symposium on Industrial Electronics (ISIE)*, Vol. 2, 2001, S. 928–931.
7. W.-B. Horng, J.-W. Peng und C.-Y. Chen, „A new image-based real-time flame detection method using color analysis“, in *Networking, Sensing and Control*, 2005, S. 100–105.
8. G. Lu, Y. Yan und M. Colechin, „A digital imaging based multifunctional flame monitoring system“, *Transactions on Instrumentation and Measurement*, Vol. 53, Nr. 4, S. 1152–1158, August 2004.
9. G. Lu, G. Gilabert und Y. Yan, „Vision based monitoring and characterisation of combustion flames“, *Journal of Physics: Conference Series, Sensors & their Applications XIII*, Vol. 15, S. 194–200, 2005.
10. G. Lu, Y. Yan, S. Cornwell, M. Whitehouse und G. Riley, „Impact of co-firing coal and biomass on flame characteristics and stability“, *Fuel*, Vol. 87, S. 1133–1140, 2008.
11. D. Sun, G. Lu, H. Zhou und Y. Yan, „Flame stability monitoring and characterization through digital imaging and spectral analysis“, *Measurement Science and Technology*, 2011.
12. ——, „Condition monitoring of combustion processes through flame imaging and kernel principal component analysis“, *Combustion Science and Technology*, Vol. 185, S. 1400–1413, 2013.
13. G. S. Rajput und Z. Rahman, „Hazard detection on runways using image processing techniques“, Enhanced and Synthetic Vision (SPIE), 2008.

Bildverarbeitung spielt in vielen Bereichen der Technik zur schnellen und berührungslosen Datenerfassung eine Schlüsselrolle. Beispielsweise in der Qualitätssicherung industrieller Produktionsprozesse, in der Robotik und zur Fahrerassistenz haben sich Bildverarbeitungssysteme einen unverzichtbaren Platz erobert. Diese Entwicklung wird unterstützt durch die Verfügbarkeit qualitativ hochwertiger und günstiger Sensorsysteme sowie durch die Zunahme der Leistungsfähigkeit von Rechnersystemen.

Der vorliegende Tagungsband des „Forums Bildverarbeitung", das am 1. und 2. Dezember 2016 in Karlsruhe als gemeinsame Veranstaltung des Karlsruher Instituts für Technologie und des Fraunhofer-Instituts für Optronik, Systemtechnik und Bildauswertung stattfand, enthält die schriftlichen Aufsätze der eingegangenen Beiträge. Darin wird über aktuelle Trends und Lösungen der Bildverarbeitung in den methodischen Schwerpunkten Bildgewinnung, Simulation, Klassifikation, Bewertung und 3D-Erfassung sowie für die Anwendungsschwerpunkte Mensch, Medizin und Industrie berichtet.

ISBN 978-3-7315-0587-7