Monokulare Kopfposenschätzung
Monocular head pose estimation

Benjamin Jäschke und Fernando Puente León

Karlsruher Institut für Technologie,
Institut für Industrielle Informationstechnik,
Hertzstr. 16, 76187 Karlsruhe

Zusammenfassung Es wird ein Verfahren zur Verfolgung der Kopfpose vorgestellt, welches die aus der photometrischen Stereometrie gewonnenen Oberflächennormalen für den Gesichtsbereich nutzt. Insbesondere wird hierbei auf die Wahl eines geeigneten Trackinggebiets sowie die Ableitung der Kopfposenänderung aus Normalenvektoren eingegangen. Der Vorteil dieses Verfahrens, welches mit mehr Aufwand verbunden ist als 2D-Verfahren, ist die leistungsfähigere Verfolgung der Pose bei starken Rotationsbewegungen.

Schlagwörter Kopfposenschätzung, Tracking, photometrisches Stereo.

Abstract We present an approach to head pose tracking by using surface normals which are gathered from photometric stereo. In particular, we focus on choosing an appropriate area for tracking and deriving the postural change from normal vectors. Despite causing more computational cost, the method presented is advantageous for tracking the head pose in case of large rotational movements.

Keywords Head pose estimation, tracking, photometric stereo.

1 Einleitung

Die Schätzung der Kopfpose ist für eine Vielzahl von Anwendungen von Interesse, bei denen eine Interaktion zwischen Mensch und

Maschine erfolgt. Oftmals wird die Kopfposenschätzung auch als Teillösung für das Problem der Blickrichtungsschätzung benötigt, bei dem zusätzlich die Augenbewegungen bestimmt werden müssen. Im hier beschriebenen Ansatz zur Kopfposenschätzung sollen lediglich Bildaufnahmen des Kopfes herangezogen werden, ohne die beobachtete Umgebung mit Markern oder anderen Referenzpunkten zu beeinflussen. Aufgrund von Mehrdeutigkeitsproblemen bei 2D-Verfahren wird in diesem Artikel ein Ansatz verwendet, der 3D-Information erfassen und berücksichtigen kann, sodass Mehrdeutigkeiten reduziert werden können.

Die nachfolgenden Abschnitte behandeln verwandte Methoden aus der Literatur, den Ansatz der photometrischen Stereometrie, die Ableitung der Kopfpose aus den gewonnenen Normalen, eine Vorstellung erster Ergebnisse, sowie die Planung für künftige Arbeiten.

2 Stand der Technik

In der Literatur wird eine Vielzahl verschiedener Ansätze zur Kopfposenschätzung beschrieben. Die Mehrheit dieser Ansätze nutzt lediglich 2D-Information. Es kann eine grobe Einteilung dieser Verfahren in pixelbasierte Methoden und merkmalsbasierte Methoden vorgenommen werden. Für einen umfassenden Überblick zu bekannten Verfahren der Kopfposenschätzung sei auf [1] verwiesen.

Pixelbasierte Methoden zeichnen sich dadurch aus, die als Bild aufgenommenen Pixel direkt heranzuziehen und zu vergleichen. Beispielsweise sei der optische Fluss [2] genannt, aus dem eine Ableitung der Kopfpose möglich ist. Es ergeben sich jedoch Mehrdeutigkeitsprobleme, die sich in einer schlechten Unterscheidbarkeit von Rotations- und Translationsbewegungen äußern. Diese Mehrdeutigkeiten lassen sich zwar reduzieren [3], sodass das Verfahren robuster bei starken Rotationen wird, jedoch lässt sich das Problem der Mehrdeutigkeiten nicht vollständig lösen.

Merkmalsbasierte Methoden beruhen auf der Erkennung signifikanter Punkte des Gesichts und nutzen diese zur Berechnung einer Kopfpose. Ein typischer Vertreter eines solchen Ansatzes sind die Active-Appearance-Modelle [4]. Auch hier zeichnet sich jedoch ab, dass die Auswertung von Abständen zwischen Merkmalen unter Mehrdeutig-

keiten leiden. Zudem müssen sämtliche Merkmale zunächst erkannt werden, was speziell bei starken Rotationen mit Problemen verbunden ist, da stets Merkmale durch Verdeckung das Bild verlassen oder neu hinzukommen.

Sämtliche dieser Verfahren bringen den Nachteil mit sich, die Kopfpose aus Daten abzuleiten, die einer Projektion unterzogen worden sind. Die fehlende Tiefeninformation führt bei pixelbasierten Ansätzen zu den im vorherigen Abschnitt beschriebenen Mehrdeutigkeiten. Bei merkmalsbasierten Verfahren führt die fehlende Tiefeninformation zur Verfälschung von Nachbarschaftsbeziehungen zwischen Merkmalen. Die veränderlichen Abstände zwischen einzelnen Merkmalen können zwar in beschränktem Rahmen zur Schätzung der Kopfpose genutzt werden, jedoch wird es zwangsläufig bei starken Rotationen zu Problemen kommen, die sich aus der Verdeckung und dem veränderten Blickwinkel auf die Merkmale ergeben. Besonders deutlich wird dies bei Active-Appearance-Modellen im Wangenbereich: Wird ein Merkmal am „Rand des Gesichts“ aus frontaler Ansicht festgelegt, so ist bei Drehung des Kopfes dieses Merkmal zunächst wiederzufinden und anschließend mit dem Merkmal vor der Kopfdrehung zu vergleichen. Bereits die Festlegung des Merkmalsorts im Wangenbereich ist in diesem Fall mit einer großen Unsicherheit behaftet und wird anschließend in die Posenberechnung eingehen.

3 Photometrisches Stereo

Als Vorbetrachtung sollen zunächst kurz die Grundlagen der photometrischen Stereometrie vorgestellt werden. Bei diesem Verfahren werden mehrere Aufnahmen unter unterschiedlichen Beleuchtungsbedingungen genutzt und hieraus die Normalenrichtungen der Oberfläche berechnet. Es werden folgende Annahmen getroffen:

- Es liegt eine Lambert'sche Oberfläche vor [5].
- Die Beleuchtungsrichtung für jede der Aufnahmen muss bekannt sein. Die Beleuchtungsrichtung wird als Vektor $\mathbf{l}_k$ bezeichnet und beschreibt die Richtung, unter der sich von der zu untersuchenden Oberfläche aus betrachtet die k-te punktförmige Lichtquelle befindet.

Sind diese Annahmen erfüllt, so ergibt sich für den hier betrachteten Fall mit vier Beleuchtungsrichtungen der folgende Zusammenhang zwischen den Intensitäten $i_k(\mathbf{x})$ an den Pixelorten $\mathbf{x} = (x, y)^\mathsf{T}$ und dem Oberflächennormalenvektor $\mathbf{n}(\mathbf{x})$:

$$\begin{pmatrix} i_1(\mathbf{x}) \\ i_2(\mathbf{x}) \\ i_3(\mathbf{x}) \\ i_4(\mathbf{x}) \end{pmatrix} = \rho(\mathbf{x}) \begin{pmatrix} \mathbf{l}_1^\mathsf{T} \\ \mathbf{l}_2^\mathsf{T} \\ \mathbf{l}_3^\mathsf{T} \\ \mathbf{l}_4^\mathsf{T} \end{pmatrix} \mathbf{n}(\mathbf{x}) \,. \tag{1}$$

Hierbei beschreibt die Albedo $\rho(\mathbf{x})$ das Rückstrahlvermögen der Oberfläche, welches neben der Oberflächenausrichtung die Intensitäten $i_k(\mathbf{x})$ beeinflusst. Dies ist jedoch unproblematisch, da die Normalenvektoren die Länge 1 aufweisen, sodass sich sowohl die skalaren Werte $\rho(\mathbf{x})$, als auch die Vektoren $\mathbf{n}(\mathbf{x})$ berechnen lassen. Dieses Verfahren lässt sich erfolgreich auf das Gesicht anwenden [6]. Im Anschluss an die Normalenberechnung erfolgt üblicherweise eine Rekonstruktion der Oberfläche. Dieser Schritt ist für das hier vorgestellte Verfahren nicht nötig, da die Normalen bereits die gewünschte 3D-Information über den Kopf liefern.

4 Ableitung der Kopfpose

Es sollen vorab die Bezeichnungen der Rotationen eingeführt werden, auf die in den folgenden Abschnitten zurückgegriffen wird. Bei Festlegung der Koordinatenrichtungen gemäß Abbildung 1 wird eine Rotation um die x-Achse als „Nicken“, eine Rotation um die y-Achse als „Gieren“ und eine Rotation um die z-Achse als „Rollen“ bezeichnet.

4.1 Vorauswahl Trackingregion

Die Anwendung der photometrischen Stereometrie liefert Normalenvektoren für jedes Pixel. Die Ableitung der Kopfpose hieraus erfordert folglich zunächst eine Auswahl derjenigen Normalenvektoren, die zum Tracking geeignet sind. Dies geschieht zunächst in Form einer Vorbetrachtung, die nur den Zweck hat, zum Tracking verwendbare Gesichtsbereiche zu finden. Das Verfahren selbst entscheidet im Rahmen des

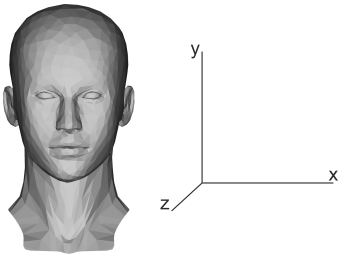

Abbildung 1: Festlegung der Koordinatenrichtungen.

Trackings, welche Gesichtsregion und welche Normalen tatsächlich verwendet werden. Beide Aspekte werden nachfolgend im Detail erläutert.

Für die Auswahl eines Gesichtsbereichs zum Tracking müssen vor allem die eventuell auftretenden Schwierigkeiten betrachtet werden, die bei der Anwendung der photometrischen Stereometrie oder dem Tracking entstehen können. Die photometrische Stereometrie nutzt Intensitäten zur Berechnung der Oberflächenausrichtung. Schattenwurf ist hierbei problematisch, da dieser gleichbedeutend damit ist, dass der betroffene Bildbereich nicht von der aktiven Beleuchtung erreicht wird. Die Folge ist eine Fehlinterpretation der Intensitäten und folglich eine falsche Normalenrichtung. Des Weiteren ist es möglich, dass die Forderung nach einer Lambert'schen Oberfläche verletzt wird. Dies ist insbesondere bei Spiegelungen der Fall, die im Bereich der Augen auftreten. Zudem sei im Sinne der Vorauswahl geeigneter Gesichtsregionen die Stirn als problematischer Bereich erwähnt, da hier Spiegelungen auftreten können, falls die beobachtete Person schwitzt.

Ferner kann eine Vorbetrachtung bezüglich der Geometrie und der Ausrichtung der zu trackenden Fläche erfolgen. Da Rotationen ebenso erfasst werden sollen wie Translationen, sind rotationssymmetrische Teilbereiche ungeeignet, was speziell die Nasenspitze betrifft. Für die Ausrichtung sind Flächen vorzuziehen, die in Richtung der Kamera – und folglich in Richtung der Lichtquellen – ausgerichtet sind. Dies bringt den Vorteil mit sich, dass beim Tracking von Rotationsbewegungen sichergestellt ist, dass auch nach der Rotation die Fläche noch von allen Lichtquellen beleuchtet werden kann. Während größerer Rotationsbewegungen ist dies nicht mit einer einzigen Gesichtsregion

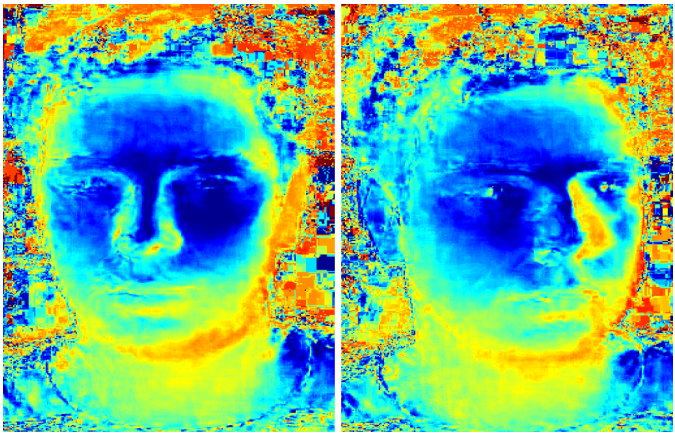

Abbildung 2: Exemplarische Darstellung zweier Nutzbarkeitskarten.

erfüllbar, weshalb das Trackinggebiet fortwährend angepasst werden muss, was in Abschnitt 4.3 näher erläutert wird.

Unter Betrachtung der beschriebenen Anforderungen lassen sich Nutzbarkeitskarten erstellen, wie in Abbildung 2 dargestellt. Hier sind diejenigen Bereiche blau dargestellt, die sämtliche Anforderungen erfüllen. Türkisfarbene bis grüne Regionen weisen auf eine Ausrichtung der Fläche hin, die nicht in Richtung der Kamera zeigt. Gelbe bis orangefarbene Bereiche kennzeichnen Regionen, in denen Schattenwurf vorliegt und die folglich keine Anwendung der photometrischen Stereometrie erlauben. Diese Karten sind für verschiedene Kopfhaltungen berechnet worden, von denen zwei hier exemplarisch dargestellt sind.

Die Karten zeigen direkt auf, dass der Schattenwurf durch die Nase die hintere Wange als Trackingregion nahezu unbrauchbar macht. Es bleiben letztlich die der Kamera zugewandte Wange und die Stirn als größere zusammenhängende Flächen, die sich zum Tracking eignen. Wie oben erwähnt, ist die Stirnregion jedoch mit dem Nachteil behaftet, dass eine schwitzende Person aufgrund von Spiegelungen nicht mehr sinnvoll zu interpretierende Oberflächennormalen liefern würde. Daher werden die Wangen als beste Option betrachtet, sofern sichergestellt ist, dass die der Kamera zugewandte Wange getrackt wird. Die Einzelheiten des hierfür notwendigen Vorgehens folgen im nächsten Abschnitt.

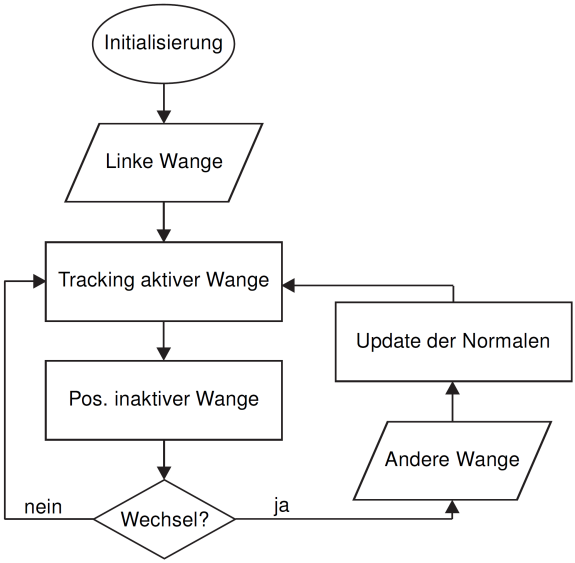

Abbildung 3: Ablaufplan des Wangenwechsel-Algorithmus.

4.2 Wangenwechsel-Algorithmus

Das Verfahren zur Auswahl der zum Tracking genutzten Wange ist in Abbildung 3 illustriert. Nach der Initialisierung, bei der mittels Gesichts- und Augendetektion die Bildregionen der beiden Wangen detektiert werden, erfolgt erstmalig eine Berechnung der Oberflächennormalen. Für die Initialisierung dürfen keine großen Rotationen des Kopfes vorliegen. Nach der Festlegung der möglichen Trackingregionen erfolgt die Auswahl der „aktiven Wange“, die tatsächlich zum Tracking genutzt wird. Diese wird zunächst auf die linke Wange festgelegt.

Anschließend beginnt das Verfahren, die Trackingschleife zu durchlaufen. Diese umfasst in jeder Iteration die folgenden Schritte:

- Das eigentliche Tracking erfolgt gemäß der im nachfolgenden Abschnitt beschriebenen Methode.
- Es wird beurteilt, ob die inaktive Wange bei der vorliegenden Pose

sichtbar ist. Ist dies der Fall, wird die entsprechende Bildregion geschätzt.

- Es wird beurteilt, ob ein Wechsel der aktiven Wange sinnvoll ist. Dies geschieht aktuell durch Festlegung eines Schwellenwertes für die Gierbewegung des Kopfes. Eine Entscheidung anhand der Qualität der Trackingregion ist derzeit in Entwicklung und soll vor allem bei größeren translatorischen Bewegungen bessere Entscheidungen bei der Wahl der aktiven Wange ermöglichen.

- Wird ein Wechsel der aktiven Wange nicht für sinnvoll erachtet, ist die Iteration beendet und das Tracking wird in der nächsten Iteration mit derselben Wange vollzogen.

- Ist ein Wechsel der aktiven Wange nötig, so erfolgt zunächst ein Update der Normalenvektoren. Dies geschieht mit Hilfe der im zweiten Schritt geschätzten Bildregion. Für diese werden die Normalen bestimmt, die in der nächsten Iteration als Referenz für die Bestimmung der Posenänderung dienen werden.

4.3 Adaption der Trackingregion

Die adaptive Veränderung des Trackingbereichs betrifft nicht die Auswahl der zum Tracking genutzten Wange, sondern bewirkt lediglich eine leichte Verschiebung der Trackingregion. Dies soll sicherstellen, dass auch bei starken Drehungen des Kopfes Normalen genutzt werden, die in Richtung der Kamera zeigen. Somit wird einerseits verhindert, dass bei weiterer Drehung das Trackinggebiet nicht mehr von allen Lichtquellen beleuchtet wird, und andererseits unterliegt die „Normalendichte“ weniger starken Veränderungen. Die Normalendichte entsteht durch die Berechnung eines Normalenvektors für jedes Pixel der aufgenommenen Bilder. Eine Gesichtsregion, die von der Kamera frontal erfasst wird, weist ein anderes Verhältnis von Pixelanzahl zu Gesichtsfläche auf, als dies bei einer nach hinten geneigten Gesichtsregion der Fall ist. Diese Pixeldichte überträgt sich direkt auf die berechneten Normalen. Da zum Zweck des Trackings ein Optimierungsproblem formuliert wird, soll anhand der Adaption des Trackingbereichs ein negativer Einfluss durch die Veränderung der Normalendichte vermieden werden.

Das Verfahren durchläuft folgende Schritte:

- Aus dem vorhergehenden Iterationsschritt ist ein Referenzgebiet mit Referenznormalen bekannt. Dieses Gebiet wird zunächst erneut genutzt und es wird überprüft, ob eine Rotation stattgefunden hat. Das Ergebnis dieses Schritts sind Winkel um die drei Rotationsachsen.

- Es folgt ein Schritt zur Kompensation des bei Kopfdrehungen auftretenden translatorischen Anteils der Gesichtsbewegung. Zugrunde liegt hier das Problem, dass die Rotationsachse des Kopfes im Bereich des Atlas liegt, bei Fehlhaltungen der beobachteten Person liegt die Rotationsachse unter Umständen sogar hinter dem Kopf. Anhand der im vorangegangen Schritt bestimmten Winkel kann nun eine Schätzung erfolgen, wie weit sich die tatsächlich zum Tracking zu verwendende Region translatorisch bewegt hat. Das Ergebnis dieses Schritts ist eine neue Bildregion und folglich eine neue Normalenmenge.

- Mit Hilfe des neuen Bereichs erfolgt eine erneute Winkelberechnung, die nun exaktere Ergebnisse für die Rotationswinkel um alle Achsen liefert.

- Eine rein translatorische Kopfbewegung wurde zu diesem Zeitpunkt noch nicht erfasst und erfolgt durch Verschiebung des Trackingbereichs. Hierbei soll die Kosinus-Ähnlichkeit der N im Trackingbereich befindlichen Normalen $\mathbf{n}_k$ zu den Referenznormalen $\mathbf{n}_{\mathrm{ref},k}$ maximiert werden: $\max\left(\sum_{k=1}^{N}\langle\mathbf{n}_k,\mathbf{n}_{\mathrm{ref},k}\rangle\right)$.

- Nach den bisherigen Schritten ist die Kopfposenänderung bekannt. Es folgt die eingangs motivierte Adaption des Trackingbereichs, der in der nächsten Iteration als Referenz dienen wird. Zu beachten ist, dass in den meisten Fällen nur geringe Bewegungen von Frame zu Frame vorliegen. Die oben formulierten Anforderungen an den Trackingbereich im Hinblick auf die Normalendichte werden deshalb in den meisten Fällen nicht verletzt. Es wird daher ein Schwellenwert festgelegt, ab dem eine Adaption erfolgt.

5 Diskussion erster Ergebnisse

Es sei vorab erwähnt, dass durch die Nutzung der photometrischen Stereometrie eine Verwendung von etablierten Datensätzen zur Kopfposenschätzung nicht möglich ist, da diese keine Aufnahmen unter definierten Beleuchtungsrichtungen beinhalten. Zum Zweck der ersten Tests werden Datenreihen verwendet, die durch Rotationsbewegungen des Kopfes unter Vorgabe von Referenzpunkten in der Umgebung des Probanden. Anhand dieser Referenzpunkte wird die Ground Truth festgelegt. Eine größer angelegte Datenaufnahme ist geplant.

5.1 Testergebnisse

Die Testergebnisse zeigen eine gute Erfassung starker Rotationen, wie Abbildung 4 zu entnehmen ist. Es wurden hierbei in erster Linie starke Gierbewegungen des Kopfes in Kombination mit leichten Nickbewegungen getestet, was einer Beobachtung des Umfelds nachempfunden ist. Sowohl der Wechsel zwischen den beiden Wangen als auch die Anpassung der Trackingregion bei starken Drehungen erzielen die gewünschte Wirkung, sodass das Tracking auch bei Rotationen bis $\pm 80°$ aufrecht erhalten werden kann. Dennoch zeigt der Ansatz Probleme, die in den nächsten Schritten gelöst werden müssen. Hier sei zunächst eine systematische Abweichung bei translatorischen Bewegungen genannt, die im nächsten Abschnitt näher erläutert wird. Zudem wird trotz der vielversprechenden Ergebnisse bei starken Gierbewegungen eine Verbesserung der Genauigkeit angestrebt. Ein Teil der Ungenauigkeit tritt durch die Kombination verschiedener Rotationen auf, als Beispiel sei hier der Fall einer starken Gierbewegung genannt, nach der zusätzlich eine Nickbewegung durchgeführt wird. Aus Sicht der Kamera wirkt diese Bewegung wie eine Rollbewegung, was an der Bindung der Bewegungsdefinition an die vom Kopf vorgegebenen Koordinatenachsen liegt.

5.2 Wölbungsproblematik

Durch die Festlegung der Beleuchtungsrichtung für die Anwendung der photometrischen Stereometrie entsteht bei geringen Abständen zwischen Kopf und Lichtquellen ein systematischer Fehler. Während

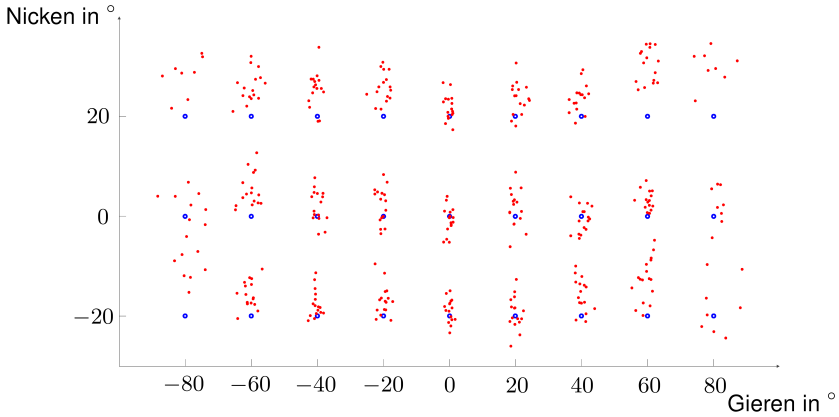

Abbildung 4: Testergebnisse (blau: Ground Truth, rot: Schätzungen).

in der Bildmitte die Beleuchtungsrichtungen korrekt sind, weichen sie mit zunehmendem Abstand zur Bildmitte mehr und mehr ab. Die Folge sind aufgrund des veränderten Einfallwinkels des Lichts geringere Intensitäten, die im Rahmen der Normalenberechnung fehlinterpretiert werden. Die sich ergebenden Normalen weisen einen Fehler auf, der sich in einem zu stark nach außen gerichteten Vektor zeigt. Die Folge hiervon wird deutlich, wenn zur Veranschaulichung eine Rekonstruktion der Oberfläche erfolgt: Sie weist eine Wölbung in der Bildmitte auf, die im Normalfall somit das Gesicht betrifft. Während mit dem in den vorherigen Abschnitten formulierten Trackingverfahren Rotationen des Kopfes von diesem Fehler nur geringfügig beeinflusst werden, verstärkt sich die Problematik bei translatorischen Bewegungen, die durch diesen Effekt von einer scheinbaren Rotation nach außen überlagert werden, selbst wenn ausschließlich translatorische Bewegungen vorliegen. Ein Verfahren zur Kompensation dieser Effekte ist derzeit in Entwicklung.

6 Zusammenfassung

Es wurde ein Ansatz zur Ableitung der Kopfpose mit Hilfe der photometrischen Stereometrie vorgestellt. Durch die Kombination mit der Auswahl der besser zum Tracking nutzbaren Wange und der Adapti-

on des Trackingbereichs können selbst starke Rotationen verfolgt werden. Die Vorteile gegenüber 2D-Verfahren sind somit deutlich erkennbar. In den nächsten Schritten wird an einer Anpassung der Normalenberechnung zur Kompensation der Lichtquellenrichtungen gearbeitet. Ferner ist die Aufnahme größerer Datensätze mit präziser Annotation der Kopfpose geplant, mit deren Hilfe auch eine vergleichende Einordnung zu anderen Verfahren erfolgen kann.

Literatur

1. E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.
2. B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," 1981.
3. S. Vater, G. Mann, and F. Puente León, "A novel regularization method for optical flow based head pose estimation," in *Automated Visual Inspection and Machine Vision*, ser. Proceedings of SPIE, J. Beyerer and F. Puente, Eds., vol. 9530. Bellingham, WA: SPIE, 2015.
4. T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.
5. J. Beyerer, F. Puente León, and C. Frese, *Automatische Sichtprüfung: Grundlagen, Methoden und Praxis der Bildgewinnung und Bildauswertung*. Berlin Heidelberg: Springer, 2012.
6. M. F. Hansen, G. A. Atkinson, L. N. Smith, and M. L. Smith, "3d face reconstructions from photometric stereo using near infrared and visible light," *Computer Vision and Image Understanding*, vol. 114, no. 8, pp. 942–951, 2010.