

# Multimodal Computational Attention for Scene Understanding

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

der Fakultät für Informatik  
des Karlsruher Instituts für Technologie (KIT)

**genehmigte**

**Dissertation**

von

**Boris Schauerte**

aus Witten, Deutschland

Tag der mündlichen Prüfung: **13.06.2014**

Erster Gutachter: **Prof. Dr.-Ing. R. Stiefelhagen**

Zweiter Gutachter: **Prof. Dr.-Ing. T. Asfour**



# Multimodal Computational Attention for Scene Understanding

Boris Schauerte

## Abstract

Attention is the cognitive process that identifies subsets within sensory inputs (*e.g.*, from the millions of human sensory receptors) that contain important information to focus subsequent complex and slow processing operations on the most relevant information. This is a key capability of humans and animals that allows us to rapidly understand what is going on in a scene despite the limited computational capacities of the brain. Consequently, since attention serves as a gateway to later cognitive processes, efficient, reliable, and rapid attentional allocation is key to predation, escape, and mating – in short, to survival.

Like their biological counterparts, robotic systems have limited computational capacities. Consequently, computational attention models are important to allow for complex cognitive processing. For this purpose, we develop highly-efficient auditory and visual attention models. For visual attention, we use hypercomplex image processing and decorrelation to calculate what is interesting in an image and are able to efficiently predict where people will look in an image. For auditory attention, we use Bayesian methods to determine what are unexpected and thus surprising sounds. Here, we are able to reliably detect arbitrary, salient acoustic events. We fuse the auditory and visual saliency in a crossmodal parametric proto-object model. Based on the detected salient proto-objects in a scene, we can use multiple criteria to plan which part of the room the robot should attend next. We have successfully implemented this approach on robotic platforms such as KIT's ARMAR robot head to efficiently explore and analyze scenes.

In many situations, people want to guide our attention to specific aspects. For example, photographers compose their images in such way that the most important object automatically grabs the viewer's attention. Furthermore, people use non-verbal signals (*e.g.*, pointing gestures and gaze) to control where a conversation partner looks to include a specific nearby object in the conversation. Interestingly, infants develop the ability to interpret such non-verbal signals very early and it is an essential ability, because it allows to associate spoken words with the visual appearances of nearby objects and thus learn language.

In the second part, we first try to identify the most prominent objects in web images. Then, we start to integrate verbal and non-verbal social signals into our saliency model. As non-verbal signals, we consider gaze and pointing gestures. Both signals direct our attention toward spatial areas to narrow the referential domain, which we model with a probabilistic corridor of attention. As verbal signals, we focus on specific spoken object descriptions that have been shown to being able to directly guide visual saliency and thus influence human gaze patterns. Interestingly, verbal and non-verbal signals complement each other, *i.e.* as one signal type becomes ambiguous it is compensated with the other. We achieve the best results with machine learning methods to integrate the available information. This way, we are able to efficiently highlight the intended target objects in human-robot interaction and web images.

---

## Summary

We derived several novel quaternion-based spectral visual saliency models (QDCT, ESR, ESW, and EPQFT), all of which perform state-of-the-art on three well-known eye tracking datasets. Furthermore, we proposed to decorrelate each image’s color information as a preprocessing step for a wide variety of visual saliency models. We have shown that color space decorrelation can improve the performance by about 4% (normalized) for eight visual saliency algorithms on three established datasets with respect to three complementing evaluation measures. Although an improvement of 4% is far from drastic, it is nevertheless a considerable achievement, because we are not aware of any other method or preprocessing step that is able to consistently and significantly improve the performance of such a wide range of algorithms. Furthermore, we improved the state-of-the-art in predicting where people look when human faces are visible in the image. Compared to Cerf *et al.*’s approach, we were able to improve the performance by 8% (*i.e.*, 25.2% normalized by the ideal AUC) with automatic face detections.

To realize auditory attention, we introduced a novel auditory saliency model that is based on the Bayesian surprise of each frequency. To allow for real-time computation on a robotic platform, we derived Gaussian surprise, which is efficient to calculate due to its simple closed form solution. Since we addressed a novel problem domain, we had to introduce a novel quantitative, application-oriented evaluation methodology and evaluated our model’s ability to detect arbitrary salient auditory events. Our results show that Bayesian surprise can efficiently and reliably detect salient acoustic events, which is shown by  $F_1$ ,  $F_2$ , and  $F_4$  scores of 0.767, 0.892, and 0.967.

We combined auditory and visual saliency in a biologically-plausible model based on crossmodal proto-objects to implement overt attention on a humanoid robot’s head. We performed a series of behavioral experiments, which showed that our model exhibits the desired behaviors. Based on a formalization as multiobjective optimization problem, we introduced ego motion as a further criterion to plan which proto-object to attend next. This way, we were able to substantially reduce the amount of head ego motion while still preferring to attend the most salient proto-objects first. Our solution exhibits a low normalized cumulated joint angle distance (NCJAD) of 15.0%, which represents that the chosen exploration order requires a low amount of ego motion to attend all proto-objects, and a high normalized cumulated saliency (NCS) of 83.3%, which indicates that highly salient proto-objects are attended early.

We investigated how the spatial distribution of objects in images influences salient object detection. Here, we provided the first empirical justification for a Gaussian center bias. This is shown by a probability plot correlation coefficient (PPCC) of 0.9988 between a uniform distribution and the angular distribution of salient objects around the image center, and a PPCC of 0.9987 between a half-Gaussian distribution and the distribution of distances of salient objects to the image center. Then, we demonstrated that the performance of salient object



---

detection algorithms can be substantially influenced by undocumented spatial biases. We debiased the region contrast algorithm and subsequently integrated a well-modeled Gaussian bias. This way, we achieved two goals: First, through integration of our explicit Gaussian bias, we improved the state-of-the-art in salient object detection for web images and at the same time quantified the influence of the center bias. Second, we derived the currently best unbiased salient object detection algorithm, which is advantageous for other application domains such as, *e.g.*, surveillance and robotics.

We presented saliency models that are able to integrate multimodal signals such as pointing and spoken object descriptions to guide the attention in human-robot interaction. We started with an initial heuristic model that combines our spectral saliency detection with a probabilistic corridor of attention, *i.e.* the “probabilistic pointing cone”, to reflect the spatial information given by pointing references. Additionally, we discussed a biologically-inspired neuron-based saliency model that is able to integrate knowledge about the target object’s appearance into visual search. We outperform both models by training conditional random fields that integrate features such as, most importantly, our locally debiased region contrast, multi-scale spectral visual saliency with decorrelated color space, the probabilistic pointing cone, and target color models. This way, we are able to focus the correct target object in the initial focus of attention for 92.45 % of the images in the PointAT dataset, which does not provide spoken target descriptions, and 75.21 % for the ReferAT dataset, which includes spoken target references. This translates to an improvement of +10.37 % and +25.21 % compared to the heuristic and neuron-based saliency models, respectively.

Finally, we learn to determine objects or object parts that are being looked-at by persons in web images. This can be interpreted as a form of gaze following in web images. For this purpose, we integrated our work on salient object detection in web images and the interpretation of attentional signals in human-robot interaction. Consequently, we transferred our methods and train conditional random fields to integrate features such as, most importantly, spectral visual saliency, region contrast saliency, and a probabilistic corridor of interest that represents the observed gaze direction. This way, the looked-at target object is focused in the initial focus of attention for 66.17 % of images in a dataset that we collected from Flickr.

To quantify the performance of our approaches, we had to collect several datasets and even propose novel evaluation procedures, because we often addressed novel tasks, problems, and domains. We derived novel evaluation procedures for these tasks: First, we quantified the ability of our auditory saliency model to determine arbitrary salient acoustic events. Therefore, we relied on measures that are commonly used to evaluate salient object detection algorithms. Second, we introduced several novel evaluation measures to evaluate tradeoffs made by our multiobjective exploration path strategies. Furthermore, we proposed several measures to quantify the ability of saliency models to highlight target objects and focus the objects after a minimum amount of focus of

---

attention shifts. We collected novel datasets for the following tasks: First, we created a dataset that consists of 60 videos to evaluate multiobjective exploration strategies. Second and third, we recorded two new datasets to evaluate how well we are able to guide our saliency model in human-robot interaction; in the absence (PointAT) and presence (ReferAT) of spoken target object information. For this purpose, fourth, we also gathered the Google-512 dataset to train our color term models. Fifth, to evaluate the identification and segmentation of gazed-at objects in web images, we collected the Gaze@Flickr dataset that we selected out of one million Flickr images.

# Multimodal Computational Attention for Scene Understanding

Boris Schauerte

## **Kurzzusammenfassung**

Aufmerksamkeit ist der kognitive Prozess, der dafür verantwortlich ist, die beschränkten Bewusstseinsressourcen auf die sensorischen Reize (beispielsweise Tonfrequenzen und Bildinhalte) zu konzentrieren, die wahrscheinlich wichtige Informationen für spätere kognitive Prozesse (beispielsweise Aktivitäts- und Objekterkennung) enthalten. Entsprechend ist Aufmerksamkeit eine bedeutende Fähigkeit von Menschen und Tieren, die es ermöglicht trotz der eingeschränkten kognitiven Kapazitäten die wichtigsten Inhalte von Szenen schnell zu erfassen und zu verarbeiten – eine Schlüsselfähigkeit für das Überleben, denn es ermöglicht schnelle Reaktionen auf plötzliche, unerwartete und möglicherweise lebensbedrohende Ereignisse. Somit lässt sich Aufmerksamkeit als Filter oder Tor für spätere kognitive Prozesse interpretieren. Dies bedeutet allerdings auch, dass alle nachfolgenden Prozesse auf eine schnelle, effiziente, und zuverlässige Zuweisung der kognitiven Kapazitäten durch den Aufmerksamkeitsprozess angewiesen sind.

Wie ihre menschlichen und tierischen Vorbilder besitzen auch Roboter nur eingeschränkte Rechenkapazitäten. Demzufolge stellen Aufmerksamkeitsmodelle ein wichtiges Hilfsmittel dar, um auf Robotern die beschränkten Ressourcen zu verteilen und komplexe kognitive Prozesse zu ermöglichen. Zu diesem Zweck haben wir hocheffiziente visuelle und akustische Aufmerksamkeitsmodelle entwickelt, die hervorstechende – sprich „saliente“ – Reize detektieren und für spätere Prozesse kennzeichnen. Zur Berechnung der visuellen Salienz nutzen wir holistische spektrale Bildverarbeitungsverfahren und Farbkorrelation. Auf diese Weise sind wir in der Lage, effizient vorherzusagen, auf welche Bereiche ein menschlicher Betrachter seine Aufmerksamkeit in Bildern richten wird. Zur Berechnung der auditorischen Salienz verlassen wir uns auf den bayesschen Wahrscheinlichkeitsbegriff, um zu berechnen wie unerwartet und somit überraschend ein akustisches Signal ist. Auf diese Weise sind wir in der Lage effizient beliebige, akustisch interessante Reize zu detektieren, auf die die Aufmerksamkeit gerichtet werden sollte. Wir fusionieren die auditorische und visuelle Salienzinformation in einem modalitätsübergreifenden parametrischen Protoobjektmodell. Aufbauend auf den detektierten salienten Protoobjekten in der Umgebung kann der Roboter anschließend planen auf welche Bereiche und Reize er seine sensorischen und kognitiven Kapazitäten optimalerweise richten sollte. Wir haben eine solche Szenenexplorationsstrategie erfolgreich auf unterschiedlichen Roboterplattformen implementiert, um effizient automatisch Szenen zu erkunden und die Objekte in der Umgebung des Roboters zu analysieren.

---

Es gibt viele Situationen in denen Menschen versuchen die Aufmerksamkeit anderer Personen auf bestimmte Aspekte zu lenken. Beispielsweise arrangieren Photographen ihre Bilder derart, dass wichtige Objekte automatisch den Blick des Betrachters auf sich ziehen. Des Weiteren nutzen wir oft sogar unbewusst Gesten und andere nonverbale Signale, um die Aufmerksamkeit unserer Gesprächspartner auf Objekte in der Umgebung zu lenken, die wir in das Gespräch einbeziehen wollen. Interessanterweise lernen Kinder bereits in einem sehr frühen Entwicklungsstadium solche nonverbalen Kommunikationssignale (insbesondere Blick- und Zeigerichtungen) zu interpretieren. Dies ist ein essentieller Aspekt der Kindesentwicklung, weil es die Interpretation nonverbaler Signale vereinfacht, gesprochene Worte mit sichtbaren Objekten in der Umgebung zu assoziieren – eine wichtige Fähigkeit zum Erlernen einer Sprache.

Im zweiten Teil der Dissertation konzentrieren wir uns darauf, die Objekte in Bildern zu identifizieren, auf die eine andere Person die Aufmerksamkeit lenken will. Wir beginnen mit der automatischen Identifikation des hervorstechendsten, zentralen Objektes in Photos und anderen Bildern aus dem Internet. Anschließend integrieren wir nonverbale und verbale Signale in ein Aufmerksamkeitsmodell zur Unterstützung der Mensch-Roboter-Interaktion. Hier konzentrieren wir uns auf Zeigegesten und bestimmte sprachliche Beschreibungen, von denen bekannt ist, dass sie die wahrgenommene Salienz und somit Aufmerksamkeit von Personen unbewusst beeinflussen können. Interessanterweise ergänzen sich nonverbale und verbale Signale derart, dass ein zusätzliches Signal eingesetzt wird, wenn nur ein Signaltyp das Zielobjekt nicht eindeutig beschreiben würde. Zeigegesten und Blickrichtung lenken die Aufmerksamkeit auf bestimmte Bereiche entlang der Blick- oder Zeigerichtung, in denen sich die relevanten Objekte befinden. Wir modellieren diese Information mithilfe eines probabilistischen Modells des aufgespannten Aufmerksamkeitskorridors. Zusätzlich kann vorhandene Information über das Aussehen eines gesuchten Objektes die wahrgenommene Salienz so beeinflussen, dass Bildbereiche hervorgehoben sind, die ähnliche Merkmale aufweisen wie das Zielobjekt. Wir präsentieren heuristische und biologisch motivierte Modelle die es uns ermöglichen die vorhandenen Informationen zu fusionieren. Allerdings erreichen wir die besten Ergebnisse mit Methoden des maschinellen Lernens. Mit unseren entwickelten Methoden sind wir in der Lage, effizient das Zielobjekt in zwei unterschiedlichen Domänen hervorzuheben: Erstens während der Mensch-Roboter-Interaktion mit Zeigegesten und Sprache. Zweitens für Bilder aus dem Internet; entweder mit oder ohne Einfluss der Blickrichtung von im Bild sichtbaren Personen.

# Acknowledgments

Behind every dissertation lies a long journey with many ups and downs, rewarding challenges and frustrating obstacles, success and disappointment. I would not have finished my dissertation without the support from my family, help from friends, and advice of my supervisors. Therefore I would like to thank:

- My family for always being a safe haven and always believing in me. My mother for her deep interest in people and psychology that still influences me. My father for being an example that you can and should love your work. My sister, the first person that got me interested in the design of everyday things.
- My friends for being there when I need distance between me and my work. Knowing that I can always count on you is an invaluable gift.
- My co-workers with whom I spend countless hours discussing, laughing, motivating, and inspiring each other.
- All the students I supervised on their path. It was fun.
- Gernot A. Fink for convincing me to work on attention despite my initial reluctance.
- Thomas Ploetz for many important advices in the early stages of my journey.
- Rainer Stiefelhagen for the most important advice: just do good work and don't think about the rest.



# Contents

<b>1. Introduction</b>	<b>11</b>
1.1. Contributions . . . . .	14
1.2. Outline . . . . .	17
<b>2. Background</b>	<b>19</b>
2.1. Attention Models . . . . .	19
2.1.1. Visual Attention . . . . .	20
2.1.2. Auditory Attention . . . . .	26
2.1.3. Multimodal Attention . . . . .	30
2.2. Applications of Attention Models . . . . .	32
2.2.1. Image Processing and Computer Vision . . . . .	33
2.2.2. Audio Processing . . . . .	34
2.2.3. Robotics . . . . .	34
2.2.4. Computer Graphics . . . . .	36
2.2.5. Design, Marketing, and Advertisement . . . . .	36
<b>3. Bottom-up Audio-Visual Attention for Scene Exploration</b>	<b>39</b>
3.1. Related Work and Contributions . . . . .	42
3.1.1. Spectral Visual Saliency . . . . .	42
3.1.2. Visual Saliency and Color Spaces . . . . .	43
3.1.3. Visual Saliency and Faces . . . . .	44
3.1.4. Auditory Saliency . . . . .	45
3.1.5. Audio-Visual Saliency-based Exploration . . . . .	46
3.1.6. Scene Analysis . . . . .	48
3.2. Visual Attention . . . . .	50
3.2.1. Spectral Visual Saliency . . . . .	51
3.2.2. Color Space Decorrelation . . . . .	67
3.2.3. Modeling the Influence of Faces . . . . .	78
3.3. Auditory Attention . . . . .	85
3.3.1. Auditory Novelty Detection . . . . .	85
3.3.2. Evaluation . . . . .	89
3.4. Saliency-based Audio-Visual Exploration . . . . .	92
3.4.1. Gaussian Proto-Object Model . . . . .	92
3.4.2. Auditory Proto-Objects . . . . .	93
3.4.3. Visual Proto-Objects . . . . .	94
3.4.4. Audio-Visual Fusion and Inhibition . . . . .	97
3.4.5. Evaluation . . . . .	98
3.5. Multiobjective Exploration Path . . . . .	104
3.5.1. Exploration Path . . . . .	104
3.5.2. Exploration Strategies . . . . .	104
3.5.3. Evaluation . . . . .	106
3.6. Summary and Future Directions . . . . .	111

<b>4. Multimodal Attention with Top-Down Guidance</b>	<b>113</b>
4.1. Related Work and Contributions . . . . .	116
4.1.1. Joint Attention . . . . .	116
4.1.2. Visual Attention . . . . .	120
4.2. Debiased Salient Object Detection . . . . .	124
4.2.1. The MSRA Dataset . . . . .	126
4.2.2. MSRA’s Photographer Bias . . . . .	126
4.2.3. Salient Object Detection . . . . .	129
4.2.4. Debiased Salient Object Detection and Pointing . . . . .	133
4.3. Attentional Guidance in Human-Robot Interaction . . . . .	135
4.3.1. Pointing Gestures . . . . .	136
4.3.2. Language . . . . .	145
4.4. Gaze Following in Web Images . . . . .	159
4.4.1. Approach . . . . .	160
4.4.2. The Gaze@Flickr Dataset . . . . .	161
4.4.3. Evaluation . . . . .	163
4.5. Summary and Future Directions . . . . .	171
<b>5. Conclusion</b>	<b>173</b>
5.1. Summary . . . . .	173
5.2. Future Work . . . . .	175
<b>A. Applications</b>	<b>177</b>
A.1. Patient Agitation . . . . .	177
A.1.1. Method . . . . .	177
A.1.2. Qualitative Evaluation . . . . .	178
A.2. Activity Recognition . . . . .	178
A.2.1. Method . . . . .	180
A.2.2. Results . . . . .	181
<b>B. Dataset Overview</b>	<b>183</b>
<b>C. Full Color Space Decorrelation Evaluation</b>	<b>187</b>
<b>D. Center Bias Integration Methods</b>	<b>199</b>
<b>E. Who’s Waldo?</b>	<b>201</b>
. Publications	203
. Bibliography	205



# List of Figures

2.1.	Treisman’s feature integration theory model. . . . .	20
2.2.	Wolfe’s guided search model. . . . .	22
2.3.	The traditional structure of feature-based visual attention models. . . . .	23
2.4.	Psychologically motivated test patterns. . . . .	24
2.5.	Salient object examples. . . . .	26
2.6.	Illustration of the human ear’s structure and sound transmission. . . . .	28
2.7.	Kayser <i>et al.</i> ’s and Kalinli and Narayanan’s auditory saliency model. . . . .	30
2.8.	Attentive robot systems. . . . .	35
2.9.	Image retargeting examples. . . . .	36
2.10.	Advertisement and web design eye tracking examples. . . . .	37
3.1.	Hierarchical object analysis example. . . . .	48
3.2.	The opponent color model as input to the visual cortex. . . . .	50
3.3.	An example of the disadvantage of color channel separation. . . . .	52
3.4.	Visualization of the quaternion Fourier spectrum. . . . .	54
3.5.	Examples that show the difference between PQFT and EPQFT. . . . .	57
3.6.	Example images from the visual saliency evaluation datasets. . . . .	60
3.7.	Illustration of the shuffled, bias-corrected AUC calculation. . . . .	61
3.8.	Influence of quaternion color component weights. . . . .	64
3.9.	Example saliency maps based on different color spaces. . . . .	73
3.10.	Degree of color component correlation illustration. . . . .	76
3.11.	Saliency algorithm performance versus intra color correlation. . . . .	77
3.12.	Example images from the Cerf/FIFA dataset with their annotated face segments. . . . .	80
3.13.	Predictive performance depending on the face integration method. . . . .	83
3.14.	Illustration of Bayesian surprise’s relation to neurons. . . . .	86
3.15.	Auditory surprise example. . . . .	87
3.16.	Visual proto-object region extraction example. . . . .	94
3.17.	The ARMAR-III humanoid robot head and our PTU setup. . . . .	99
3.18.	How active foveation can improve visual perception quality. . . . .	99
3.19.	Evaluation of how a sensory focus improves audio perception quality. . . . .	100
3.20.	Cyclic focus of attention shift experiment illustration. . . . .	101
3.21.	An example of multimodal scene exploration. . . . .	102
3.22.	Sample images that show the two evaluation environments. . . . .	106
3.23.	An example of different focus of attention selection strategies. . . . .	108
3.24.	Exploration strategies evaluation results. . . . .	109
4.1.	How gaze can guide our attention. . . . .	113
4.2.	Example images from Yücel <i>et al.</i> ’s dataset . . . . .	119
4.3.	Example images to illustrate our different target domains. . . . .	124
4.4.	MSRA dataset example images. . . . .	125
4.5.	Salient object distribution Q-Q plots. . . . .	127

4.6. Example images for RC'10, RC'10+CB, LDRC and LDRC+CB. . . . .	129
4.7. Cheng <i>et al.</i> 's implicit center bias. . . . .	131
4.8. Heuristic pointing integration example. . . . .	138
4.9. PointAT dataset example images. . . . .	141
4.10. Pointing gesture examples for the PointAT dataset. . . . .	144
4.11. Example of neuron-based top-down modulation. . . . .	147
4.12. Automatic target model acquisition. . . . .	149
4.13. ReferAT dataset example images. . . . .	151
4.14. ReferAT object database acquisition. . . . .	153
4.15. Pointing gesture examples for the ReferAT dataset. . . . .	154
4.16. Gaze@Flickr example images. . . . .	159
4.17. Gaze@Flickr dataset example images. . . . .	162
4.18. Gaze@Flickr example predictions. . . . .	164
4.19. Gaze@Flickr example predictions. . . . .	165
4.20. Gaze@Flickr object distribution Q-Q plots. . . . .	166
A.1. An illustration of Martinez's MRD. . . . .	178
A.2. Example sequence that was used to qualitatively evaluate surprise for agitation detection. . . . .	179
A.3. Example of Rybok's proto-object extraction. . . . .	180
E.1. Who and where is Waldo? . . . . .	201

# List of Tables

3.1.	Performance of selected visual saliency algorithms. . . . .	63
3.2.	Quaternion component weighting evaluation results. . . . .	66
3.3.	Statistical test visualization chart. . . . .	70
3.4.	Color space decorrelation results. . . . .	72
3.5.	Color space component correlations table. . . . .	75
3.6.	Evaluation results on the Cerf/FIFA dataset. . . . .	82
3.7.	Auditory surprise evaluation results. . . . .	91
3.8.	Composition of the exploration path evaluation data. . . . .	106
4.1.	MSRA salient object detection evaluation . . . . .	132
4.2.	LDRC vs RC'10 on pointing data. . . . .	133
4.3.	Target object detection on the PointAT dataset . . . . .	143
4.4.	Target object detection on the ReferAT dataset (no lang.). . . . .	155
4.5.	Target object detection on the ReferAT dataset (det. lang.). . . . .	156
4.6.	Target object detection on the ReferAT dataset (gt. lang.). . . . .	157
4.7.	Target object detection on the ReferAT dataset (color terms vs object models). . . . .	158
4.8.	Target object detection on the Gaze@Flickr dataset . . . . .	167
A.1.	Activity recognition results. . . . .	182
C.1.	AUC, CC, and NSS baseline algorithm performance. . . . .	188
C.2.	Color space decorrelation results (Bruce/Toronto, AUC). . . . .	189
C.3.	Color space decorrelation results (Bruce/Toronto, CC). . . . .	190
C.4.	Color space decorrelation results (Bruce/Toronto, NSS). . . . .	191
C.5.	Color space decorrelation results (Judd/MIT, AUC). . . . .	192
C.6.	Color space decorrelation results (Judd/MIT, CC). . . . .	193
C.7.	Color space decorrelation results (Judd/MIT, NSS). . . . .	194
C.8.	Color space decorrelation results (Kootstra, AUC). . . . .	195
C.9.	Color space decorrelation results (Kootstra, CC). . . . .	196
C.10.	Color space decorrelation results (Kootstra, NSS). . . . .	197
D.1.	Quantitative center bias integration comparison. . . . .	199



# List of Abbreviations

<b>2D</b>	2-dimensional
<b>3D</b>	3-dimensional
<b>4D</b>	4-dimensional
<b>AUC</b>	area under the curve
<b>AUROC</b>	area under the receiver operator characteristic curve
<b>CC</b>	correlation coefficient
<b>CJAD</b>	cumulated joint angle distance
<b>CRF</b>	conditional random field
<b>CS</b>	cumulated saliency
<b>DCT</b>	discrete cosine transform
<b>EMD</b>	earth mover's distance
<b>EPQFT</b>	Eigen pure quaternion Fourier transform
<b>ESR</b>	Eigen spectral residual
<b>ESW</b>	Eigen spectral whitening
<b>FFT</b>	fast Fourier transform
<b>FHR</b>	focus of attention hit rate
<b>FoA</b>	focus of attention
<b>FIT</b>	feature integration theory
<b>GBVS</b>	graph-based visual saliency
<b>GPU</b>	graphics processing unit
<b>GSM</b>	guided search model
<b>HOF</b>	histogram of optical flow
<b>HOG</b>	histogram of oriented gradients
<b>HRI</b>	human-robot interaction
<b>ICOPP</b>	intensity and color opponents
<b>ICS</b>	integrated cumulated saliency
<b>ICU</b>	intensive care unit
<b>iNVT</b>	iLab Neuromorphic Vision Toolkit
<b>IoR</b>	inhibition of return
<b>KIT</b>	Karlsruhe Institute of Technology
<b>KLD</b>	Kullback-Leibler divergence
<b>LBP</b>	local binary patterns
<b>MSER</b>	maximally stable extremal regions
<b>MCT</b>	modified census transform
<b>MDCT</b>	modified discrete cosine transform
<b>MRD</b>	medical recording device

<b>nAUROC</b>	normalized area under the receiver operator characteristic curve
<b>NCJAD</b>	normalized cumulated joint angle distance
<b>NCS</b>	normalized cumulated saliency
<b>NSS</b>	normalized scanpath saliency
<b>NTOS</b>	normalized target object saliency
<b>NP</b>	nondeterministic polynomial time
<b>PCA</b>	principal component analysis
<b>PHAT</b>	phase transform
<b>PHR</b>	pixel hit rate
<b>POS</b>	part-of-speech
<b>PGC</b>	probabilistic gaze cone
<b>PPC</b>	probabilistic pointing cone
<b>PPCC</b>	probability plot correlation coefficient
<b>PTU</b>	pan-tilt-unit
<b>PQFT</b>	pure quaternion Fourier transform
<b>QDCT</b>	quaternion discrete cosine transform
<b>Q-Q</b>	quantile-quantile
<b>RAM</b>	random access memory
<b>ROC</b>	receiver operating characteristic
<b>SNR</b>	signal-to-noise ratio
<b>SRP</b>	steered response power
<b>STCT</b>	short-time cosine transform
<b>STFT</b>	short-time Fourier transform
<b>STIP</b>	space time interest points
<b>SVM</b>	support vector machine
<b>TDOA</b>	time difference of arrival
<b>TRW</b>	tree-reweighted belief propagation
<b>TSP</b>	traveling salesman problem
<b>VOCUS</b>	visual object detection with computational attention system
<b>WTA</b>	winner-take-all
<b>ZCA</b>	zero-phase transform

This page intentionally left blank.





# 1

## Introduction

We immediately spot a warning triangle on a street or a black sheep in a flock. Yet, although we know what we are looking for, it can take us minutes to find Waldo, who blends into a crowd of nondescript people. When it comes to hearing, we are able to selectively listen to different speakers in a crowded room that is filled with a multitude of ongoing conversations. And, an unexpected, unfamiliar sound at night can awaken and scare us, making our hearts race as a means to prepare us for fight or flight. These examples illustrate how our brain highlights some visual or auditory signals while suppressing others. Understanding what our brain will highlight is not just fundamental to understand and model the human brain but forms the basis for best practices in various application areas. For example, based on a set of basic cognitive rules and guidelines, movie directors compose the camera shots of scenes in such a way that the relevant information gets subconsciously highlighted. Furthermore, horror movies use harsh, non-linear, and unexpected sounds to trigger strong emotional responses.

This form of highlighting is better known as “selective attention” and describes mechanisms in the human brain that determine which parts of the incoming sensory signal streams are currently the most interesting and should be analyzed in detail. Attentional mechanisms select stimuli, memories, or thoughts that are behaviorally relevant among the many others that are behaviorally irrelevant. Such attentional mechanisms are an evolutionary response to the problem that the human brain – due to computational limitations – is not able to fully process all incoming sensory information and, as a consequence, has to select and focus on the potentially most relevant stimuli. Otherwise humans would not be able to rapidly understand what is going on in a scene, which however is key to predation and escape – in short, to human survival and evolution.

Thus, attention serves as a gateway to later cognitive processes (*e.g.*, object recognition) and visual attention is often compared with a “spotlight”. Following the spotlight metaphor, only scene elements that are illuminated by the spotlight are fully processed and analyzed. By moving the spotlight around the scene, we can iteratively build up an impression of the entire scene. For example, in the human visual system this is implemented in the form of rapid, subconscious eye movements, the so-called “saccades”. By moving the eye, fixating and analyzing one location at a time, the small fixated part of the scene is projected onto the

fovea. The fovea is the central part of the retina that is responsible for highly resolved, sharp, non-peripheral vision. As a consequence, attention does not just reduce the necessary computational resources, but it ensures the best possible sensory quality of the fixated sensory stimuli for subsequent stages – thus, it represents an evolutionary solution to manage perceptual sensory quality and computational limitations. Orienting the eyes, head, or even body to selectively attend a stimulus is called “overt attention”. In contrast, “covert attention” describes a mental focus (*e.g.*, to focus on a specific aspect of an overtly focused object) that is not accompanied by physical movements.

Since only a small part of the signal will be analyzed, the definition of what is potentially relevant – *i.e.*, “salient” – is absolutely critical. Here, we have to differentiate between two mechanisms. First, bottom-up, stimulus-driven saliency highlights signals as being salient that differ sufficiently from their surrounding in space and time. For example, due to its unnatural triangular shape and color, the advance warning triangle is highly salient; as is the black sheep that visually “pops out” of the flock of white sheep. Similarly, a sudden, unexpected sound attracts our auditory attention, because it differs substantially from what we have heard before. Bottom-up attention is also often referred to as being “automatic”, “reflexive”, or “stimulus-driven”. Second, top-down, user-driven factors can strongly modulate or, in some situations, even override bottom-up attention. Such top-down factors can be expectations or knowledge about the appearance of a target object that is being searched (*i.e.*, the basis for so-called “visual search”) that influences which distinctive features should attract our attention. For example, during a cocktail party we are able to focus our auditory attention on a location (*i.e.*, the location of the person we want to listen to) and specific frequencies to highlight the voice of our conversation partners and suppress background noise, which allows us to better understand what is being said. Similarly, for example, when visually searching for a red object, all red objects in the scene become more salient. However, top-down attention also faces limitations that can be experienced when looking for Waldo, which – due to the presence of distractors – is still a challenging problem even though we exactly know how Waldo looks like (see Appx. E). Furthermore, in many situations, bottom-up attention can not be suppressed entirely and highly salient stimuli can still attract the attention independent of conflicting top-down influences. Top-down attention is also commonly referred to as being “voluntary”, “centrally cued”, or “goal-driven”.

Naturally, auditory and visually salient stimuli are integrated into a crossmodal attention model and work together. For example, when we hear a strange, unexpected sound behind our back, then we will naturally turn our head to investigate what has caused this sound. Furthermore, information that we acquire from speech (*e.g.*, about the visual appearance of an object) can modulate the visual saliency. In fact, in recent years, it becomes more and more apparent that the sensory processing in the human brain is multisensory to such extent that,

---

for example, lipreading or the observation of piano playing without hearing the sound can activate areas in the auditory cortex.

Attention models try to model what the human brain considers as being salient or interesting. Traditionally, attention models have been used to model and predict the outcome of psychological experiments or tests with the goal to understand the underlying mechanisms in the human brain. However, attention models are not just interesting to achieve a better understanding of the human brain, because to know what humans find interesting is an important information for a wide range of practical applications. For example, we could optimize the visual layout of advertisement or user interfaces, reduce disturbing signal compression artifacts, or suppress annoying sounds in urban soundscapes. In general, knowing what is potentially relevant or important information opens further application scenarios. For example, we could focus machine learning algorithms on the most relevant training data. An application area that seems to be particularly in need of attention mechanisms is robotics, because robots that imitate aspects of human sensing and behavior face similar challenges as humans. Accordingly, attention models could be used to implement overt and covert attention to save computational resources, improve visual localization, or help to mimic aspects of human behavior in human-robot interaction.

In this thesis, we describe our work on two aspects of multimodal attention:

- 1. Bottom-up Audio-Visual Attention for Scene Exploration**

In the first part, we describe how we realized audio-visual overt attention on a humanoid robot head. First, we define which auditory and visual stimuli are salient. For this purpose, we use spectral visual saliency detection with a decorrelated color space for visual saliency and a probabilistic definition of surprise to implement auditory saliency detection. Then, we determine and localize auditory and/or visually salient stimuli in the robot’s environment and, for each salient stimulus, we represent the spatial location and extent as well as its saliency in the form of so-called proto-objects. This makes it possible to fuse the auditory and visual proto-objects to derive crossmodally salient regions in the environment. Based on these salient spatial regions, *i.e.* our salient proto-objects, we implement overt attention and plan where the robot should turn its head and look next. Here, we do not just incorporate each proto-object’s saliency, but use a multiobjective framework that allows us to integrate ego-motion as a criterion.

- 2. Multimodal Attention with Top-Down Guidance**

In the second part, we investigate attention models for situations in which a person tries to direct the attention toward a specific object, *i.e.* an intended target object. Here, we address two top-down influences and application domains: First, how photographers and other artists compose images to direct the viewer’s attention toward a specific salient object that forms the picture’s intended center of interest. Second, how interacting people

use verbal (*e.g.*, “that red cup”) and non-verbal (*e.g.*, pointing gestures) signals to direct the interaction partner’s visual attention toward an object in the environment to introduce or focus this object in the conversation and talk about it. For this purpose, we rely on machine learning methods to integrate the available information, where we also build on features that we derived in the first part of this thesis. Finally, we combine both tasks and address web images in which people are looking at things. Thus, we shift from the first part’s focus on the general interestingness of signals to being able to let top-down information guide visual saliency and highlight the specific image regions that depict intended target objects.

## 1.1 Contributions

---

Among several other contributions, our major contributions to the state-of-the-art that we present and discuss in this thesis have been made in these areas:

1. **Visual saliency**

We focused on how we can represent color information in a way that supports bottom-up visual saliency detection. For this purpose, we investigated the use of quaternions for holistic color processing, which in combination with quaternion component weighting was able to improve the state-of-the-art in predicting where people look by a small margin. Based on our experiences with quaternion-based approaches, we investigated color decorrelation as a method to represent color information in a way that supports to independently process color channels. This way, we improved the predictive performance of eight visual saliency algorithms, again improving the state-of-the-art.

2. **Auditory saliency**

We proposed a novel auditory saliency model that is based on Bayesian surprise. Our model has a clear biological foundation and, in contrast to prior art, it is able to detect salient auditory events in real-time. The latter was an important requirement to implement auditory attention on a robotic platform. Since a similar approach has not been proposed and evaluated before, we also introduced a novel, application-oriented evaluation methodology and show that our approach is able to reliably detect arbitrary salient acoustic events.

3. **Audio-visual proto-objects and exploration**

Proto-objects are volatile units of information that can be bound into a coherent and stable object when accessed by focused attention, where the spatial location serves as index that binds together various low-level features into proto-objects across space and time. We introduce Gaussian proto-objects as novel, object-centred method to represent the 3-dimensional

spatial saliency distribution. In contrast to prior art, our representation is parametric and not grid-like such as, for example, elevation-azimuth maps or voxels. We implement proto-objects as being primitive, uncategorized object entities in our world model. This way proto-objects seamlessly form the foundation to realize biologically-plausible crossmodal saliency fusion, implement different overt attention strategies, realize object-based inhibition of return, and serve as starting point for the hierarchical, knowledge-driven object analysis. We are not aware of any prior system that integrates all these aspects in a similarly systematic, consistent, biologically-inspired way.

#### 4. **Salient object detection**

We investigated how the photographer bias influences salient object detection datasets and, as a consequence, algorithms. We provided the first empirical justification for the use of a Gaussian center-bias and have shown that algorithms may have implicit, undocumented biases that enable them to achieve better results on the most important datasets. Based on these observations, we adapted a state-of-the-art algorithm and removed its implicit center-bias. This way, we were able to achieve two goals: First, we could improve the state-of-the-art in salient object detection on web images through the integration of an explicit, well-modeled center-bias. Second, we derived the currently best performing unbiased algorithm, which can provide superior performance in application domains in which the image data is not subject to a center-bias.

#### 5. **Saliency with top-down guidance**

We were the first to create attention models that let the often complementary information contained in spoken descriptions of a target object’s visual appearance and non-verbal signals – *e.g.*, pointing gestures and gaze – guide the visual saliency and, as a consequence, the focus of attention. This way, we are often able to highlight the intended target object in human-robot interaction with the goal to facilitate to establish a joint focus of attention between interacting people. We started with biologically-oriented models, but achieved the best results with machine learning methods that learn how to integrate different features, ranging from our spectral visual saliency models to probabilistic color term models. After having demonstrated that this successfully works for human-robot interaction, we approach a more challenging domain and try to identify the objects of interest in web images that depict persons looking at things.

We provide a more detailed discussion of individual contributions in the related work part of chapter 3 (“Bottom-up Audio-Visual Attention for Scene Exploration”) and 4 (“Multimodal Attention with Top-Down Guidance”).

## Code and Impact

To support scholarly evaluation by other researchers as well as the integration of our methods into other applications, we made most algorithms that are described in this thesis open source. This includes, for example, the source code for auditory saliency detection – including Gaussian surprise – and our spectral visual saliency toolbox. The latter was downloaded several thousand times during the past years. Our code has also been successfully used in other research projects at the computer vision for human-computer interaction lab:

1. **Patient agitation**

Our Gaussian surprise model was used for patient agitation detection in intensive care units, see Sec. A.1.

2. **Activity recognition**

Our quaternion image signature saliency model and proto-objects were used to improve activity recognition, see Sec. A.2.

## Further Contributions

Related to this dissertation, but not thematically central enough to be described in detail in the main document, we contributed to further fields:

1. **Assistive technologies for visually impaired people**

Initially, we learned and applied color term models to integrate the top-down influence of spoken object descriptions on attention in human-robot interaction [SF10a]. However, these color models also became an essential element in our work on computer vision for blind people, because they allowed us to use sonification to guide a blind person’s attention toward certain spatial areas and help find lost things [SMCS12].

Furthermore, as part of this thesis, we use conditional random fields to learn to guide visual saliency in human-robot interaction, see Sec. 4.3. However, we also have applied this conditional random field structure, learning, and prediction methodology to identify the area in front of a walking person that is free of obstacles (see [KSS13]).

2. **Color term and attribute learning from web images**

We learned color term models (*i.e.*, representations of what people actually mean when they say “red” or “blue”) from images that were automatically gathered from the web. Here, we proposed to use image randomization in such a way that the color distributions of artificial or post-processed images, which are common in the domain of web images, better match the distribution of natural images [SF10b]. To further improve the results, we combined salient object detection as spatial prior and supervised latent Dirichlet allocation to treat color terms as linguistic topics [SS12a]. Inspired by our intended target application, we introduced an evaluation measure

that compares the classification results with human labels, which better reflects the oftentimes fuzzy boundaries between color terms.

## 1.2 Outline

---

The organization of this thesis is as follows:

### Main content

In chapter 2, we provide a broad overview of related work that forms the background to understand many ideas and concepts throughout this thesis. We first discuss aspects of auditory (Sec. 2.1.1), visual (Sec. 2.1.2), and multimodal attention (Sec. 2.1.3). Then, we overview attention model applications (Sec. 2.2).

In chapter 3, we present how we implemented audio-visual bottom-up attention for scene exploration and analysis. Here, we start with a discussion of the most relevant related work and how our approach deviates from prior art (Sec. 3.1). Then, we introduce how we define visual and auditory saliency (Sec. 3.2 and 3.3, respectively) before we explain how we fuse this information in an audio-visual proto-object model (Sec. 3.4). The proto-object model forms the basis to plan where the robot should look next, which is implemented solely saliency-driven (Sec. 3.4.4.A) and with the consideration of the necessary ego-motion (Sec. 3.5).

In chapter 4, we present how we learn to let top-down influences guide attention to highlight specific objects of interest. Again, we first discuss the most relevant related work and clarify our contributions (Sec. 4.1). Then, we analyze how the photographer bias in web images influences modern salient object detection algorithms and present a state-of-the-art method without such a bias (Sec. 4.2). Afterwards, we discuss how we integrate pointing gestures and spoken object references into an attention model that highlights the referred-to object (Sec. 4.3). Finally, we show how the methods that we first presented in an human-robot interaction context (Sec. 4.3) can be applied to web images to identify objects that are being looked at.

In chapter 5, we summarize the results of this thesis and discuss potential topics of future work.

### Appendices

In appendix A, we present two applications that rely on our saliency models and were developed at the computer vision for human-computer interaction lab. First, in Sec. A.1, we describe how Martinez uses our Gaussian surprise model to detect patient agitation in intensive care units. Second, in Sec. A.2, we describe how Rybok uses our quaternion image signature saliency model and proto-objects to improve the accuracy of an activity recognition system.

In appendix B, we overview and briefly describe all twelve datasets that are relevant to this thesis.

In appendix C, we provide further color space decorrelation results that supplement our evaluation in Sec. 3.2.2.B.

Finally, in appendix E, we briefly introduce Waldo.



# 2

## Background

Although in principle all attention models serve the same purpose, *i.e.* to highlight potentially relevant and thus interesting – that is to say “salient” – data, attention models can differ substantially in which parts of the signal they mark as being of interest. This is to a great extent due to the varying research questions and interests in relevant fields such as, most importantly, neuroscience, psychophysics, psychology, and computer science. However, it is also caused by the vagueness as well as application- and task-dependence of the underlying problem description, *i.e.* what is interesting?

The purpose of this chapter is to provide an introduction to visual and auditory attention (Sec. 2.1) and its applications (Sec. 2.2) that serves as background information for the remainder of this thesis.

### 2.1 Attention Models

---

In general, it is possible to distinguish three types of attention models by the respective research field: First, neurobiological models try to understand and model in which part of the brain attentional mechanisms reside and how they operate and interact on a neurobiological level. Second, psychological models try to model, explain, and better understand aspects of human perception and not the brain’s neural system and layout. Third, computational models implement principles of neurobiological and psychological models, but they are also often subject to an engineering objective. Such an engineering objective is less to model the human brain or perception, but to be part of and improve artificial systems such as, *e.g.*, vision systems or complex robots.

For visual attention, the following text focuses on computational and to a lesser extent psychological models, because well-studied, elaborated psychological and computational models exist. Furthermore, a deep understanding of neurobiological aspects of the human brain’s neural visual system is of minor relevance for the remainder of this thesis. An interesting complementary lecture to this section is the excellent survey by Frintrop *et al.* [FRC10], which specifically tries to explain attention related concepts and ideas across the related fields of neurobiology, psychology, and computer science. For auditory attention, it is necessary to address neurobiological aspects of the human auditory system,

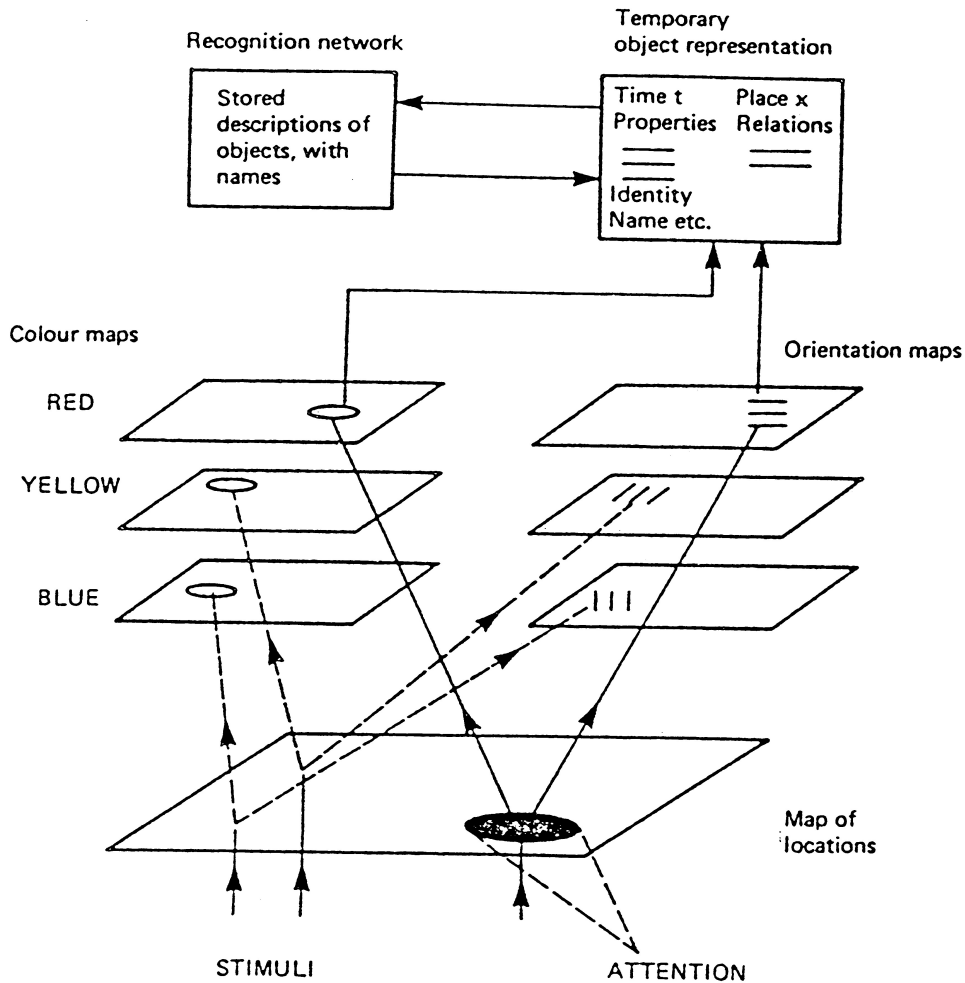


Figure 2.1.: Treisman's feature integration theory model. Image from [TG88].

because concise elaborated psychological and computational do not exist and a basic understanding of the human auditory system is important to understand the motivation of proposed computational models. Here, Fritz *et al.* and Hafter *et al.* provide very good neurobiological overviews of auditory attention [FEDS07, HSL07].

### 2.1.1. Visual Attention

**Psychological Models** The objective of psychological attention models is to explain and better understand human perception, not to model the brain's neural structure. Among the psychological models, the feature integration theory (FIT) by Treisman *et al.* [TG80] and Wolfe *et al.*'s guided search model (GSM) [Wol94] are probably by far the most influential models. Aspects of both models are still present in modern models and both models have constantly been adapted to

incorporate later research findings. A deeper discussion of psychological models can be found in the review by Bundesen and Habekost [BH05].

Treisman’s feature integration theory [TG80], see Fig. 2.1, assumes that “different features are registered early, automatically, and in parallel across the visual fields, while objects are identified separately and only at a later stage, which requires focused attention” [TG80]. This simple description includes various aspects that are still fundamental for psychological and computational attention models. Conspicuity in a feature channel are represented in topological “conspicuity” or “feature maps”. The information from the feature map is integrated in a “master map of location”. A concept that is nowadays most widely known as “saliency map” [KU85]. This master map of location encodes “where” things are in an image, but not “what” they are, which reflects the “where” and “what” pathways in the human brain [FRC10]. Attention is serially focused on the highlighted locations in the master map and the image data around the attended location is passed as data to higher perception tasks such as, most importantly, object recognition to answer “what” is shown at that location.

Although Treisman’s early model primarily focused on bottom-up perceptual saliency, Treisman also considered how attention is affected during visual search, *i.e.* when looking for specific target objects. A target is easier – *i.e.*, faster – to find during visual search the more distinctive features it exhibits that differentiate it from the distractors. To implement visual search mechanism in FIT, Treisman proposed to inhibit the feature maps that encode the features of distractors, *i.e.* non-target features.

Treisman *et al.* also introduced the concept of object files as “temporary episodic representations of objects”. An object file “collects the sensory information that has so far been received about the object. This information can be matched to stored descriptions to identify or classify the object” [KTG92].

Wolfe *et al.* [WCF89, Wol94] introduced the initial guided search model to address shortcomings of early versions of Treisman’s FIT model, see Fig. 2.2. As its name suggests, Wolfe’s GSM focuses on modeling and predicting the results of visual search experiments. Accordingly, it explicitly integrates the influence of top-down information to highlight potential target objects during visual search. For this purpose, it uses the top-down information to select the feature type that best distinguishes between target and distractors.

**Computational Models – Traditional Structure** Most computational attention models follow a similar structure, see Fig. 2.3, which is adopted from Treisman’s feature integration theory [TG80] and Wolfe’s guided search model [WCF89, Wol94] (see Fig. 2.1 and 2.2, respectively). The first computational implementation of this model was proposed by Koch and Ullman [KU85], who also coined the term “saliency map” that is identical to the concept of Treisman’s “master map of location”. The general idea is to compute several features in parallel that are fused to form the final saliency map. This traditional struc-

feature  
integration  
theory

saliency map

visual search

object files

guided search

traditional  
structure

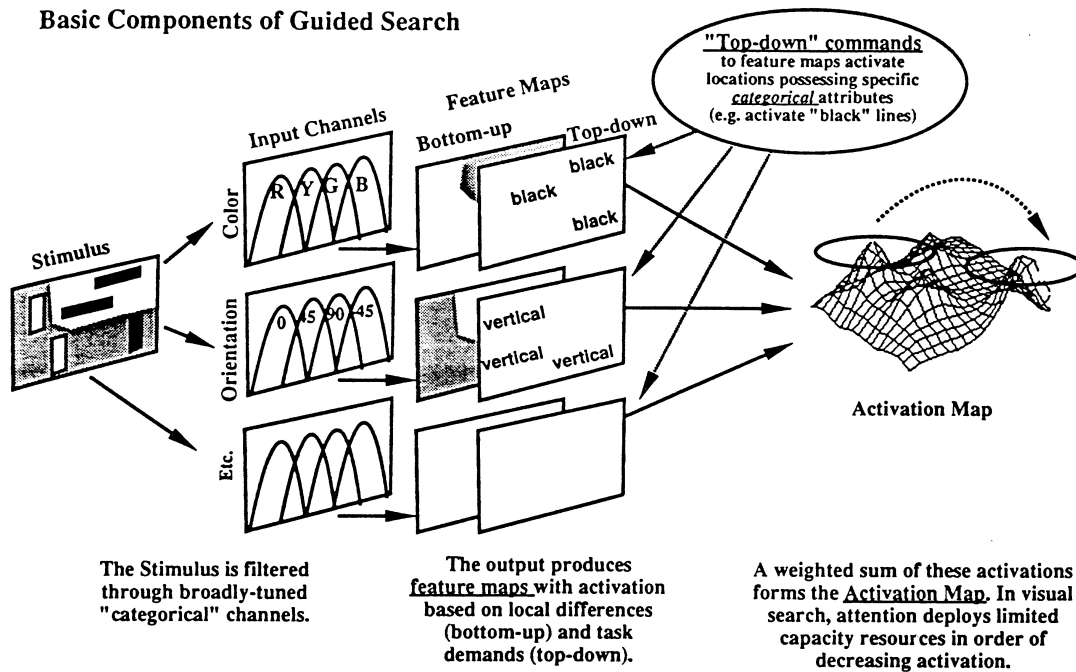


Figure 2.2.: Wolfe's guided search model. Image from [Wol94].

ture consists of several processing steps to calculate the saliency map and the different computational models differ in how they implement these steps. For example, Frintrop's visual object detection with computational attention system (VOCUS) uses integral images to calculate the center-surround differences [Fri06], Harel *et al.*'s graph-based visual saliency model [HKP07] implements Itti and Koch's model [IKN98], which is depicted in Fig. 2.3, in a consistent graph-based framework.

In this model, one or several image pyramids are computed to facilitate the subsequent computation features are computed on different scales. Then, image features are computed, which typically are based on local contrast operations such as, most importantly, "center-surround differences" that compare the average value of a center region with the average value in the surrounding region [Mar82]. The most common low-level feature channels are intensity, color, orientation, and motion. Each feature channel is subdivided into several feature types such as, for example, red, green, blue, and yellow feature maps for color. The features are commonly represented in so-called "feature maps", which are also known as "conspicuity maps". These feature maps are then normalized and fused to form a single "saliency map".

fusion How the conspicuity maps are fused is a very important aspect of attention models. It is important that image regions that stand out in one feature map are not suppressed by the other feature maps. Furthermore, the feature calculation can be non-linear, leading to strong variations in the value range across and

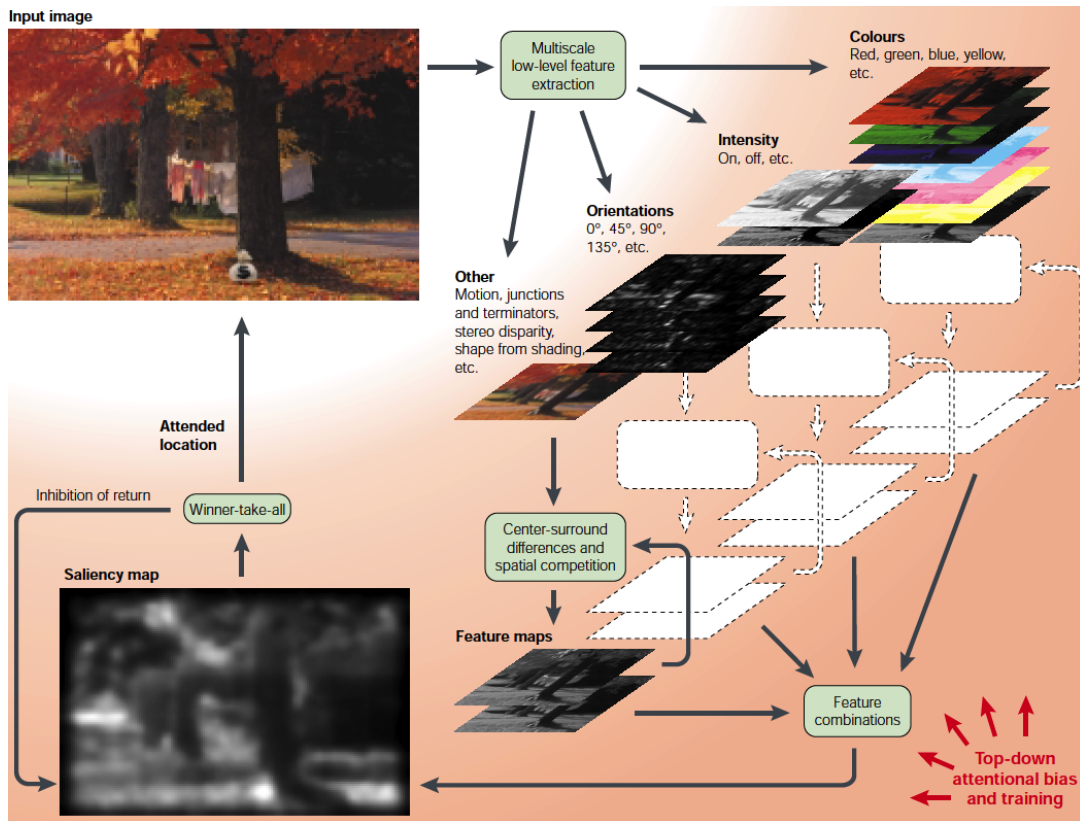


Figure 2.3.: The traditional structure of feature-based computational visual attention models, which is the basis of Itti and Koch’s [IKN98] traditional visual attention model. Image from [IK01a].

even within feature channels. Typical normalizations not just try to normalize the value range but also try to highlight local maxima and suppress the often considerable noise in the feature maps [IK01a, IKN98, Fri06]. The feature maps can be weighted, for example, bottom-up by their uniqueness or top-down to incorporate task knowledge when fused into the final saliency map.

normalization

Although the saliency map can serve as input to subsequent processing operations, *e.g.* as a relevance map for image regions, many applications require a trajectory of image regions similar to human saccades. Saccadic movement of the human eye is an essential part of the human visual system and critical to focus and resolve objects. By moving the eye, the small part of the scene that is fixated can be sensed with greater resolution, because it is projected on the central part of the retina, *i.e.* the fovea, which is responsible for highly resolved, sharp, non-peripheral vision. To serially attend image regions, the saliency map’s local maxima are determined and sequentially attended, typically in the order of descending saliency. A major contribution of Koch and Ullman [KU85] was to show that serially extracting the local maxima can be implemented with biologically-motivated winner-take-all (WTA) neural networks. To serially shift

saccades

inhibition of return

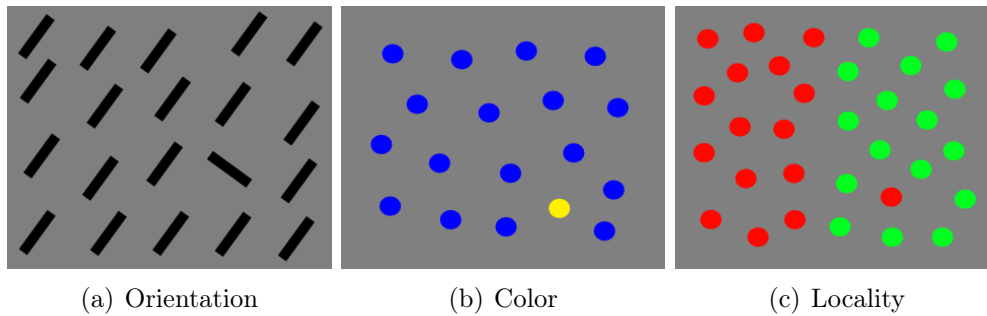


Figure 2.4.: Psychologically motivated test patterns that haven been and are still used to assess the capabilities of visual attention models [KF11]. The goal is to highlight the irregularities in the patterns.

the focus of attention, the saliency of an attended region is suppressed so that the return of the focus of attention to previously attended regions is inhibited.

The computational model as described so far mostly reflects bottom-up attention, *i.e.* it does not explicitly handle task-specific top-down information (*e.g.* as is given by a sentence that describes a searched object such as “search for the red ball”). The most common approach to integrate top-down information is control the influence of the feature maps during the fusion and adapt the weights in such way that feature maps that are likely to highlight distractors are suppressed [Wol94, NI07]. The weights can either be static or dynamic to adapt the model to specific scenarios [XZKB10]. Additionally, it is possible to integrate further, more specialized feature maps that encode, for example, faces, persons, or even cars [JEDT09, CFK09].

**Computational Models – Non-Traditional Approaches** Since human eye movements are controlled by visual attention, which can easily be observed, gaze trajectories have long served as basis to study visual attention and aspects of human cognition in general. For example, in 1967, Yarbus showed that eye movements depend on the task that is given to a person [Yar67]. Consequently, the main goal of psychological models is to explain and predict eye movements that are recorded during eye tracking experiments. However, due to the lack of modern computerized eye tracking equipment, the abilities of visual attention models where for a long time assessed by testing whether or not they were able to replicate effects that have been observed on psychological test patterns, see Fig 2.4. In the last five years, several eye tracking datasets have been made publicly available to evaluate visual attention models (*e.g.*, [KNd08, BT09, CFK09, JEDT09]; Winkler and Subramanian provide an up-to-date overview of eye tracking datasets [WS13]). Among other aspects, such easily accessible datasets and the resulting quantitative comparability of test results has lead to a plethora of novel algorithms such as, for example, attention by information maximization [BT09], saliency using natural statistics [ZTM<sup>+</sup>08], graph-based visual saliency

[HKP07], context-aware saliency [GZMT12, GZMT10], and Judd *et al.*'s machine learning model [JEDT09]. Interestingly, Borji *et al.* recently evaluated many proposed visual saliency algorithms on eye tracking data [BI13, BSI13b].

However, although being often evaluated on eye tracking data, most recently proposed models do not try to implement or explain any psychological or neurobiological models (*e.g.*, [HHK12, HZ07]). However, a biological plausibility can sometimes be discovered later (*e.g.*, [BZ09]). One such recent trend are spectral saliency models that were first proposed by Hou *et al.* [HZ07]. These models operate in the image's frequency spectrum and exploit the well-known effect that spectral whitening of signals will "accentuate lines, edges and other narrow events without modifying their position" [OL81]. Since these models are based on the fast Fourier transform (FFT), they combine state-of-the-art results in predicting where people look with the computational efficiency inherited from the FFT. Please note that spectral saliency models are discussed in detail in Sec. 3.1.1.

spectral models

Another recent trend is to use machine learning techniques to learn to predict where humans look, which was first proposed by Judd *et al.* [JEDT09]. Most saliency models that rely on machine learning are either pixel- or patch-based. Pixel-based approaches have in common with the traditional structure of computational models that they calculate a collection of feature maps. Then, classification or regression methods such as, for example, support vector machines [JEDT09] or boosting [Bor12] can be trained to learn how to optimally fuse the individual feature maps into the final saliency map. Patch-based approaches compare image patches against each other to calculate the saliency of each patch. For example, it is possible to rank the image patches by their uniqueness and assign a high saliency to patches that contain features that are rarely seen across the image [LXG12]. However, all approaches that rely on machine learning have the disadvantage that they require enough training data, which can be problematic, because most datasets consist of a very limited number of eye tracked images.

machine learning

**Computational Models – Salient Object Detection** Recently, Liu *et al.* adapted the traditional definition of visual saliency by incorporating the high level concept of a salient object into the process of visual attention computation [LSZ<sup>+</sup>07]. A "salient object" is defined as being the object in an image that attracts most of the user's interest such as, for example, the man, the cross, the baseball players and the flowers that are shown in Fig. 2.5. Accordingly, Liu *et al.* [LSZ<sup>+</sup>07] defined the task of "salient object detection" as the binary labeling problem of separating the salient object from the background. Here, it is important to note that the selection of a salient object happens consciously by the user whereas the gaze trajectories, which are recorded with eye trackers, are the result of mostly unconscious processes. Consequently, considering that salient objects naturally attract human gaze [ESP08], salient object detection

salient object

salient object detection



Figure 2.5.: Example images from Achanta *et al.*'s and Liu *et al.*'s salient object detection dataset [AS10, LSZ<sup>+</sup>07].

and predicting where people look are very closely related yet different tasks with different evaluation measures and characteristics.

Since the ties of salient object detection to psychology and neurobiology are relatively loose, a wide variety of models has been proposed in recent years that are even less restricted by biological principles than traditional visual saliency algorithms. Initially, Liu *et al.* [LSZ<sup>+</sup>07] combined multi-scale contrast, center-surround histograms, and color spatial-distributions with conditional random fields. Liu *et al.*'s ideas – a combination of histograms, segmentation, and machine learning – can still be found in most salient object detection algorithms. Alexe *et al.* [ADF10] combine traditional bottom-up saliency, color contrast, edge density, and superpixels in a Bayesian framework. Closely related to Bayesian surprise [IB06], Klein *et al.* [KF11] use the Kullback-Leibler divergence of the center and surround image patch histograms to calculate the saliency map, whereas Lu and Lim [LL12] calculate and invert the whole image's color histogram to predict the salient object. Achanta *et al.* [AHES09, AS10] rely on the difference of pixels to the average color and intensity value of an image patch or even the whole image. Cheng *et al.* [CZM<sup>+</sup>11] use segmentation and define each segments saliency based on the color difference and spatial distance to all other segments.

common ideas

### 2.1.2. Auditory Attention

Auditory attention is an important, complex system of bottom-up – *i.e.*, sound-based salience – and top-down – *i.e.*, task-dependent – aspects. Among other aspects, auditory attention assists in the computation of early auditory features



and acoustic scene analysis<sup>1</sup>, the identification and recognition of salient acoustic objects, enhancement of signal processing for the attended features or objects, and the planning of actions in response to incoming auditory information [FEDS07]. Moreover, auditory attention can be directed to a rich set of acoustic features including, among others, spatial location, auditory pitch, frequency or intensity, tone duration, timbre, speech versus non-speech, and characteristics of individual voices [FEDS07]. The best example for these abilities is the “cocktail party effect” [Che08], which illustrates that we are able to attend and selectively listen to different speakers in a crowded room that is filled with a multitude of ongoing conversations. Consequently, auditory attention influences many levels of auditory processing; ranging from processing in the cochlea to the association cortex. Not unlike the “what” and “where” pathways in the human brain’s visual system, there seem to be auditory “what” and “where” pathways, whose activation depends on whether an auditory task requires attending to an auditory feature or object or to a spatial location [ABGA04, DSCM07].

acoustic features

cocktail party effect

However, since auditory attention is an active research field in neurobiology, psychophysics, and psychology, it is only possible to provide a brief overview of selected aspects in the following. There exist however two detailed literature overviews: First, Hafter *et al.*’s review [HSL07] focuses on bottom-up aspects of auditory attention. Second, Fritz *et al.*’s survey [FEDS07] nicely presents aspects of top-down auditory and crossmodal attention. However, although there exists a large body of existing work, it is important to say that there are still many open research questions [FEDS07]. Some of these questions are directly related to the work presented in this dissertation such as, for example: How much of the brain’s acoustic novelty detection mechanisms can be explained by simple habituation mechanisms? What are the differences and similarities between visual and auditory attention? What is an appropriate computational model of auditory attention?

open questions

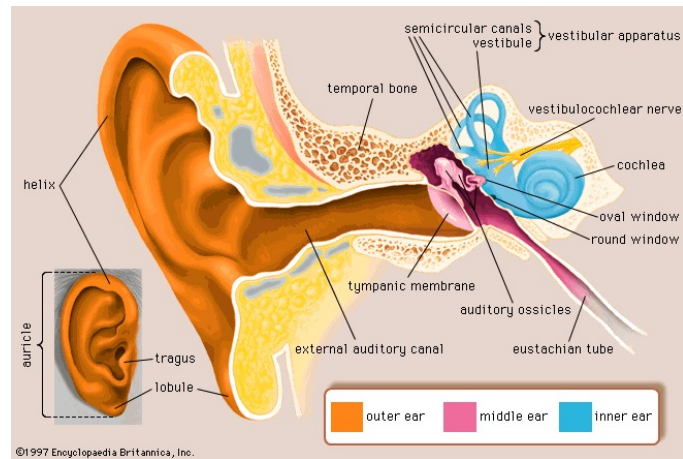
**How humans perceive sound** As shown in Fig. 2.6(a), the cochlea is a coiled system of three ducts: the scala vestibuli, the scala tympani, and the scala media. All of which are filled with lymphatic fluid. The cochlea contains a partition which is known as the “basilar membrane”, see Fig. 2.6(b). The basilar membrane is essential for our sense of hearing and consists of, most importantly, the scala media, the organ of Corti, the tectorial membrane, and the basilar membrane.

cochlea

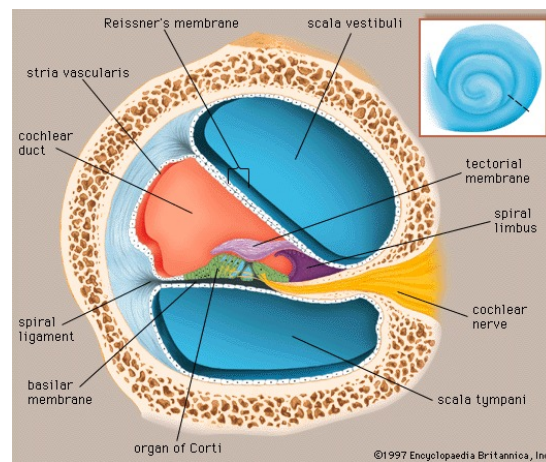
basilar membrane

Sound waves that reach the ear lead to oscillatory motions of the auditory ossicles. The oval window allows the transmission of this stimulus into the cochlea. In the cochlea, this stimulus sets the basilar membrane as well as the fluids in the scalae vestibuli and tympani in motion. The location of the maximal amplitude of the travelling wave that moves the basilar membrane

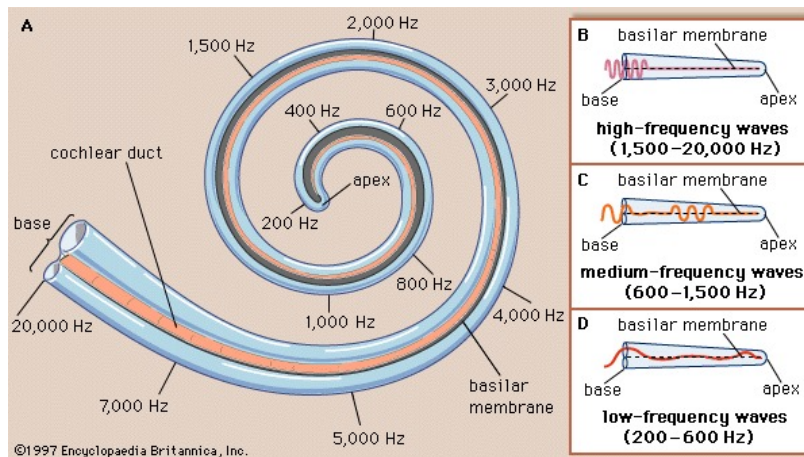
<sup>1</sup>Auditory scene analysis describes the process of segregating and grouping sounds from a mixture of sources to determine and represent relevant auditory streams or objects [Bre90].



(a) Structure of the human ear [Encc].



(b) Cross section of the cochlea [Encb].



(c) Basilar membrane stimulation at different frequencies [Enca].

Figure 2.6.: Images illustrating the structure and transmission of sounds in the human ear. Images from Encyclopaedia Britannica, Inc.

depends on the frequency of the incoming sound signal. In other words, the basilar membrane performs a frequency analysis of the incoming sound wave, see Fig. 2.6(c). The motion along the basilar membrane stimulates nerve cells that are located in the organ of Corti, see Fig. 2.6(b). These nerve cells send electrical signals to the brain, which are finally perceived as sound.

frequency  
analysis  
nerve cells

**Bottom-up auditory attention** As mentioned before, attentional effects in the human auditory system can occur at various levels of auditory processing. Interestingly, the earliest, mostly bottom-up attentional mechanisms can be observed already in the cochlea [FEDS07, DEHR07, HPSJ56].

bottom-up  
attention in  
cochlea

The ability to detect “novel”, “odd”, or “deviant” sounds amidst the environmental background noise is an important survival skill of humans and animals. Accordingly, the brain has evolved a sophisticated system to detect novel, odd, and deviant sounds. This system includes an automatic, pre-attentive component that analyzes stability and novelty of the acoustic streams within the acoustic scene, even for task-irrelevant acoustic streams [FEDS07, WTSH<sup>+</sup>03, WCS<sup>+</sup>05, Sus05].

The brain’s acoustic novelty detection system consists of an interconnected set of mechanisms, which includes “adaptive” neurons and a specialization of so-called “novelty” detection neurons. Here, novelty detection neurons specifically encode deviations from the pattern of preceding stimuli. There exist two alternative views on this “change detection” within the auditory scene, depending on where the triggered novelty responses arise in the brain. According to the first view, novelty signals can occur very early in the human auditory system [PGMC05] and suggest the possibility of subcortical pathways for change detection [FEDS07]. However, most research focuses on projections of current neural sound representations that are matched against incoming sounds [FEDS07]. In this view, the change detection system continuously monitors the auditory environment, tracks changes, and updates its representation of the acoustic scene [SW01]. Here, the matching and the novelty response is a largely pre-attentive mechanism, which however can be influenced by top-down mechanisms. It has been shown that this kind of signal mismatch detection can be triggered by deviations in stimulus frequency, intensity, duration or spatial location, or by irregularities in spectrotemporal sequences (over periods of up to 20 seconds), or even in patterns of complex sounds such as speech and music [FEDS07]. Once such a novel or odd stimulus is detected and marked, it can be analyzed by the auditory system to decide whether it should receive further attention or even trigger a behavioral response. Unfortunately, the exact neural basis of this impressive fast, pre-attentive change detection system has not conclusively been found so far.

**Computational models** In contrast to visual attention, hardly any computational auditory attention models exist (*cf.* [Kal09]). Most closely related to the work presented in this thesis is the model by Kayser *et al.* [KPLL05] and Kalinli and Narayanan [KN07]. In both models, Itti *et al.*’s [IKN98] visual

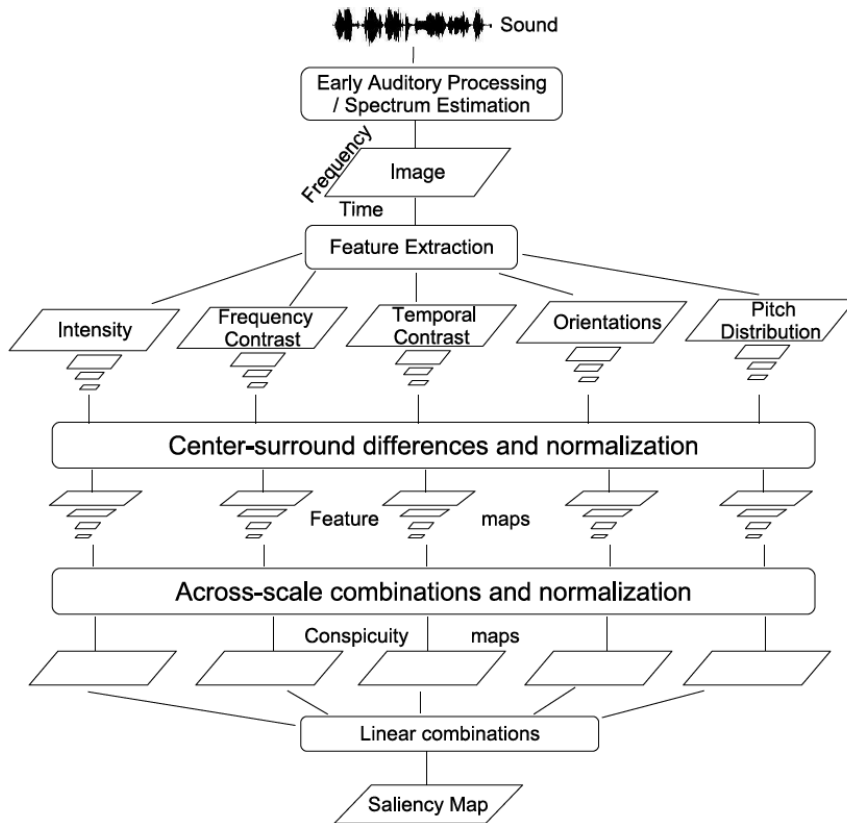


Figure 2.7.: Itti and Koch’s visual saliency model [IKN98] transferred to auditory saliency detection as has been proposed by Kayser *et al.* [KPLL05] and similarly by Kalinli and Narayanan [KN07]. Image from [Kal09].

saliency model, see Fig. 2.3, is applied on a map representation of the acoustic signal’s frequency spectrum, see Fig. 2.7, which is equivalent or very similar to the signal’s spectrogram. This model has been successfully applied and extended for speech processing by Kalinli and Narayanan [KN09, Kal09, KN07], where Kalinli and Narayanan focus on integrating top-down influences in the auditory attention model. Using the auditory spectrum of incoming sound as the basis for bottom-up auditory attention mimics “the process from basilar membrane to the cochlear nucleus in the human auditory system” [Kal09]. Transferring Itti *et al.*’s visual model to auditory signals is a radical implementation of the idea that the human visual and auditory systems have many similarities. But, it is not clear whether there exists an accessible time-frequency memory in early audition as is implied by the model’s time-frequency map, see Fig. 2.7.

### 2.1.3. Multimodal Attention

**Crossmodal integration** There exist substantial similarities between the visual and auditory attention systems: Most importantly, both consist of bottom-

up and top-down components and there appear to be specialized “what” and “where” processes. Since a few years, there is an increasing amount of experimental results that show that all sensory processing in the human brain is in fact multisensory [GS06]. For example, it has been shown that lipreading [CBB<sup>+</sup>97] or the observation of piano playing without hearing the sound [HEA<sup>+</sup>05] can activate areas in the auditory cortex.

Several studies have shown that the presence of a visual stimulus or attending a visual task can draw away attention from an auditory stimulus, which is indicated by a decreased activity in the auditory cortex (*e.g.*, [LBW<sup>+</sup>02, WBB<sup>+</sup>96]). Similarly, auditory attention can negatively influence visual attention. In fact, it was shown that there exists a reciprocal inverse relationship between auditory and visual activation, which means that increases in visual activation correlate with a decrease in auditory activation and vice versa. A very interesting study was performed by Weissman *et al.* [WWW04]. Weissman *et al.* created a conflict between auditory and visual target stimuli, and crossmodal distractors. They observed that when the “distracting stimulus in the task-irrelevant sensory channel is increased, there was a compensatory increase in selective attention to the target in the relevant channel and a corresponding increase in activation in the relevant sensory cortex” [FEDS07]. This suggests that it is likely that there exists a top-down mechanism that regulates the relative strengths of the sensory channels.

How auditory and visual sensory information interact for the control of overt attention, *i.e.* directing the sensory organs toward interesting stimuli, has recently been investigated by Onat *et al.* [OLK07]. Onat *et al.* performed eye tracking studies in which the participants were listening to sounds coming from different directions. It was shown that eye fixation probabilities increase toward the location where the sound originates, which means – unsurprisingly – that the selection of fixation points depends on auditory and visually salient stimuli. Furthermore, Onat *et al.* used the data to test several biologically plausible crossmodal integration mechanisms and found “that a linear combination of both unimodal saliencies provides a good model for this integration process” [OLK07]. Interestingly, such a linear combination is not just optimal in an information theoretic sense (see [OLK07]), but it also allows to adjust the relative strength of the sensory channels. However, this model assumes the existence of a 2-dimensional auditory saliency map that encodes where salient stimuli occur in the scene and how salient these stimuli are, which can be directly fused with the visual saliency map to form a joint audio-visual saliency map.

**High-level influences** Not just crossmodal effects can influence what is interesting in one modality. Instead, there exist many top-down signals that can direct attention toward specific targets (*e.g.*, [Ban04, Hob05, CC03, TTDC06, STET01, WHK<sup>+</sup>04, NI07]).

Verbal descriptions of object properties can directly influence what is perceived as being perceptually salient (*e.g.*, [STET01, WHK<sup>+</sup>04, NI07]). For example, it has been shown that knowledge about an object’s visual appearance can influence the perceptual saliency to highlight an object that we are actively searching, *i.e.* in a so-called visual search task. But, only specific information that refers to primitive preattentive features allows such attentional guidance [WHK<sup>+</sup>04]. Accordingly, if we have a good visual impression or memory of the target that we are looking for (*e.g.*, we have just seen it a few moments ago) or if we at least know the target’s color, then we can find the target faster. In these cases, the visual saliency will be guided in such way that it stronger highlights image regions that exhibit the target’s preattentive features. In contrast, for example, categorical information about the search target (*e.g.*, search for an animal) typically does not provide such top-down guidance (see [WHK<sup>+</sup>04]).

Interestingly, certain features that would typically be associated with high-level vision tasks can attract our low-level attention independent of task. Most importantly, it has be shown that faces and face-like patterns attract the gaze of infants as young as 6 weeks, *i.e.* before they can consciously perceive the category of faces [SS06]. The fact that the gaze and, consequently, interest of infants is attracted by face-like patterns seems to be an important aspect of early infant development, especially for social signals and processes (see, *e.g.*, [KJS<sup>+</sup>02]). Interestingly, infants show the ability to follow the observed gaze direction of caregivers at an age of 6 months [Hob05]. If people talk about objects that are part of the environment, where and at what people are looking at is related to the object that is being talked about. Consequently, the ability to follow the caregiver’s gaze makes it possible for an infant to associate what it sees with the words it hears, an important ability to learn a language. Similar to gaze but more direct and less subtle, infants also soon develop the ability to interpret pointing gestures (see [LT09]). Accordingly, pointing gestures and gaze are both non-verbal signals that direct the attention toward a spatial region of interest (see, *e.g.*, [Ban04, LB05, LT09]). This is an essential aspect in natural interaction, because it makes it possible to direct and coordinate the attention of interacting persons and, thus, helps to establish a joint focus of attention. In other words, such non-verbal signals are used to influence where an interaction partner is looking in order to direct his gaze toward a specific object that is or will become the subject of the conversation. Consequently, the generation and interpretation of such signals is fundamental for “learning, language, and sophisticated social competencies” [MN07a].

## 2.2 Applications of Attention Models

---

Knowing in advance what people might find interesting and attend to is an important information that can be integrated into many applications. Images and videos can be compressed better, street signs can be designed to immediately

grab the attention, and advertisement can put stronger emphasis on the intended message. Furthermore, having an estimate of what is probably a relevant signal in a data stream allows us to focus computational algorithms. This way, machine learning can learn better models from less data, class-independent object detection as well as object recognition can be improved, and robots are able to process incoming sensory information in real-time despite limited computational power.

### 2.2.1. Image Processing and Computer Vision

Image and video compression algorithms can improve the perceptual quality of compressed images and videos by allocating more bits to code image regions that exhibit a high perceptual saliency [GZ10, OBH<sup>+</sup>01]. This way, image regions that are likely to attract the viewers' interest are less compressed and thus show fewer disturbing alterations such as compression artifacts. Ouerhani *et al.* [OBH<sup>+</sup>01] implement such an adaptive coding scheme that favors the allocation of a higher number of bits to those image regions that are more conspicuous to the human visual system. The compressed image files are fully compatible with the JPEG standard. An alternative approach was recently proposed by Hadizadeh and Bajic [HB13]. Their method uses saliency to automatically reduce potentially attention-grabbing coding artifacts in regions of interest.

image  
compression

Visual attention and object recognition are tightly linked processes in human perception. Accordingly, although most models of visual attention and object recognition are separated, there is an increasing interest in integrating both processes to increase the performance of computer vision systems. Initial approaches tried to use attention as a front-end to detect salient objects or keypoint locations. Miao *et al.* [MPI01] use an attentional front-end with the biologically motivated object recognition system HMAX [RP99]. Walther and Koch [WK06] combine an attention system with a SIFT-based object recognition [Low04] and demonstrate that they are able to improve object recognition performance. Going a step further, Walther and Koch [WK06] suggest a unifying attention and object recognition framework. In this framework, the HMAX object recognition is modulated to suppress or enhance image locations and features depending on the spatial attention.

object  
recognition

Related to such attentional front-ends for object recognition, principles of visual attention have recently been integrated into approaches for general, class independent object detection [ADF10]. This way, sampling windows can be distributed according to the "objectness" distribution and used as location priors for class-specific object detectors. This can greatly reduce the necessary number of windows evaluated by class-specific object detectors as has been shown in the PASCAL Visual Object Classes (VOC) challenge 2007 [EVGW<sup>+</sup>]. Interestingly, going in the other direction, high-level object detectors are being integrated into saliency models to model, for example, that the human visual attention is attracted by faces and face-like patterns. For this purpose, some models integrate

object  
detection

detectors for faces, the horizon, persons and even cars [CHEK07, JEDT09]. This shows that attention and object recognition might grow together in the future.

one-shot learning Saliency has also been employed as a spatial prior to learn object attributes, categories, or classes from weakly labeled images. For example, Fei-Fei Li *et al.*'s [FFFP03] approach to “one-shot learning” uses Kadir and Brady’s saliency detector [KB01] to sample features at highly salient locations. The most salient regions are clustered over location and scale to give a reasonable number of distinctive features per image.

### 2.2.2. Audio Processing

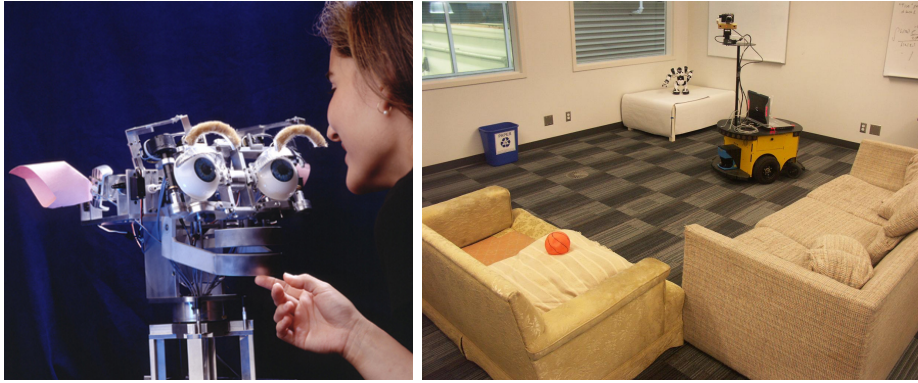
soundscape design human audio analysis speech processing In contrast to applications in computer vision, only few applications of acoustic or auditory saliency models have been explored so far. Coensel and Botteldooren [CB10] propose to use an auditory attention model in soundscape design to assess how specific sounds can mask unwanted environmental sounds. Lin *et al.* [LZG<sup>+</sup>12] as well as, in principle, Kalinli and Narayanan [KN07, KN09] use Itti’s classic visual saliency model [IKN98] to highlight visually salient patterns in the spectrogram. Lin *et al.* [LZG<sup>+</sup>12] fuse the spectrogram’s saliency map with the original spectrogram and use the resulting saliency-maximized audio spectrogram to enable faster than real-time detection of audio events by human audio analysts. Kalinli and Narayanan [KN07] use the spectrogram’s saliency map to detect prominent syllable and word locations in speech, achieving close to human performance. The task of syllable detection was chosen by the authors to investigate low-level auditory saliency models, because during speech perception, a particular phoneme or syllable can be perceived to be more salient than the others due to the coarticulation between phonemes, and other factors such as the accent, and physical and emotional state of the talker [KN07].

### 2.2.3. Robotics

In addition to reducing computational requirements by focusing on the most salient stimuli, robots can benefit from attentional mechanisms at several conceptual levels [Fri11]. On a low level, attention can be used for salient landmark detection and subsequent scene recognition and localization. On a mid level, attention can serve as a pre-processing step for object recognition. On a high level, attention can be implemented in a human-like fashion to guide actions and mimic human behavior, for example, during object manipulation or human-robot interaction.

localization Salient landmarks are excellent candidates for localization, because they are visually outstanding and distinctive, often having unique features. This makes them easy to (re-)detect and allows for a very sparse set of localization landmarks that can easily be detected, accessed in memory, and matched in real-time. The ARK project [NJW<sup>+</sup>98] is one of the earliest projects that investigated the use of salient landmarks for localization. The localization was based on manually





(a) Kismet: An attentive social robot [BS99]. Image from [Bre]. (b) Curious George: Attentive exploration robot [MFL<sup>+</sup>08]. Image from [MFL<sup>+</sup>07].

Figure 2.8.: Attentive robot systems.

generated maps of static obstacles and natural visual landmarks. Siagian and Itti [SI09] presented an integrated system for coarse global localization based on the “gist” of the scene and fine localization within a scene using salient landmarks. Frintrop and Jensfelt [FJ08] combined attention and salient landmark detection with simultaneous localization and mapping (SLAM). The attention system VOCUS [Fri06] detects salient regions. These regions are tracked and matched to all entries in a database of previously seen landmarks to estimate a 3D position.

The main difference between robotic applications and, for example, image processing is that a robot can move its body parts to interact with its environment and influence what it perceives. This way, robots can control their geometric parameters, *e.g.* where it looks, and manipulate the environment to improve the perception quality of specific stimuli [AWB88]. This can be implemented with an attentive two-step object detection and recognition mechanism: First, regions of interest are detected in a peripheral vision system based on visual saliency and a coarse view of the scene. Second, the robot then investigates each region of interest by focusing its sensors on the target object, which provides high-resolution images for object recognition (*e.g.*, [MFL<sup>+</sup>08, GAK<sup>+</sup>07]). It is noteworthy to say that using this strategy Meger *et al.*'s robot “Curious George” [MFL<sup>+</sup>08], see Fig. 2.8, won the 2007 and 2008 Semantic Robotic Vision Challenge [Uni].

A common assumption in the field of socially interactive robots is that “humans prefer to interact with machines in the same way that they interact with other people” [FND03]. This is based on the observation that humans tend to treat robots like people and, as a consequence, tend to expect human-like behavior from robots [FND03, NM00]. According to this assumption, a computational attention system that mimics how humans direct their attention can facilitate human-robot interaction. For example, this idea has been implemented in the

active  
perception

human-robot  
interaction

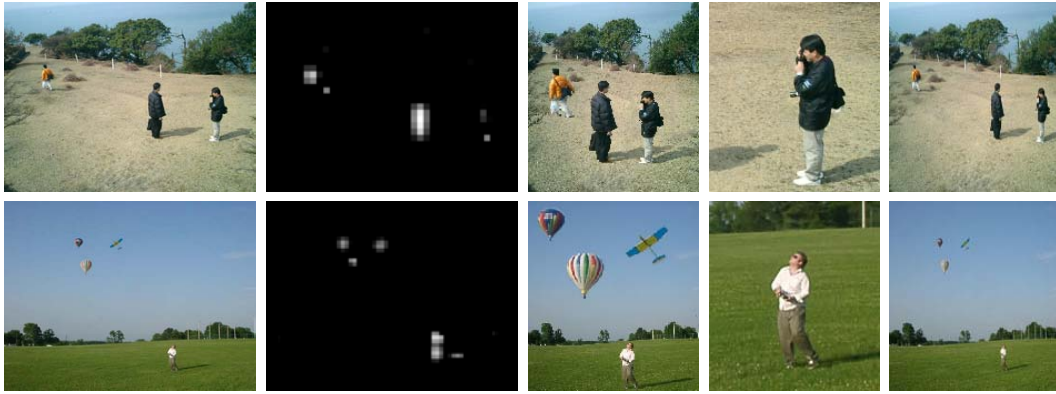


Figure 2.9.: Left-to-right: Original image, importance map, retargeted image, cropped image, and scaled image. Images from [SLNG07].

social robot Kismet, see Fig. 2.8, whose gaze is controlled by a visual attention system [BS99].

#### 2.2.4. Computer Graphics

retargeting

Naturally, knowing what attracts the viewer’s attention is important when automatically generating or manipulating images. For example, it is possible to automatically crop an image to only present the most relevant content to a user and/or act as a thumbnail [SLBJ03, SAD<sup>+</sup>06, CXF<sup>+</sup>03]. Similarly, content-aware media retargeting automatically changes the aspect ratio of images and videos to optimize the presentation of visual content across platforms and screen sizes [SLNG07, AS07, RSA08, GZMT10]. For this purpose, saliency models are used to automatically determine image regions that are likely to contain relevant information. Depending on their estimated importance, image regions are then deleted or morphed so that the resized image best portrays the most relevant information, see Fig. 2.9.

#### 2.2.5. Design, Marketing, and Advertisement

There exist several companies such as, for example, SMIVision [SMI] and Gaze-Hawk [Gaz] that offer eye tracking experiments as a service. This enables companies to analyze how people view their webpage, advertisement, or image and video footage, see Fig. 2.10. Other companies such as Google have their own in-house laboratories and solutions to perform eye tracking experiments and research [Goo].

With increasingly powerful computational attention models that predict human fixations, it becomes possible to reduce the need for expensive and intrusive eye tracking experiments. In 2013, 3M has started to offer its visual attention service [3M] that uses a computational attention model as a cheaper and faster

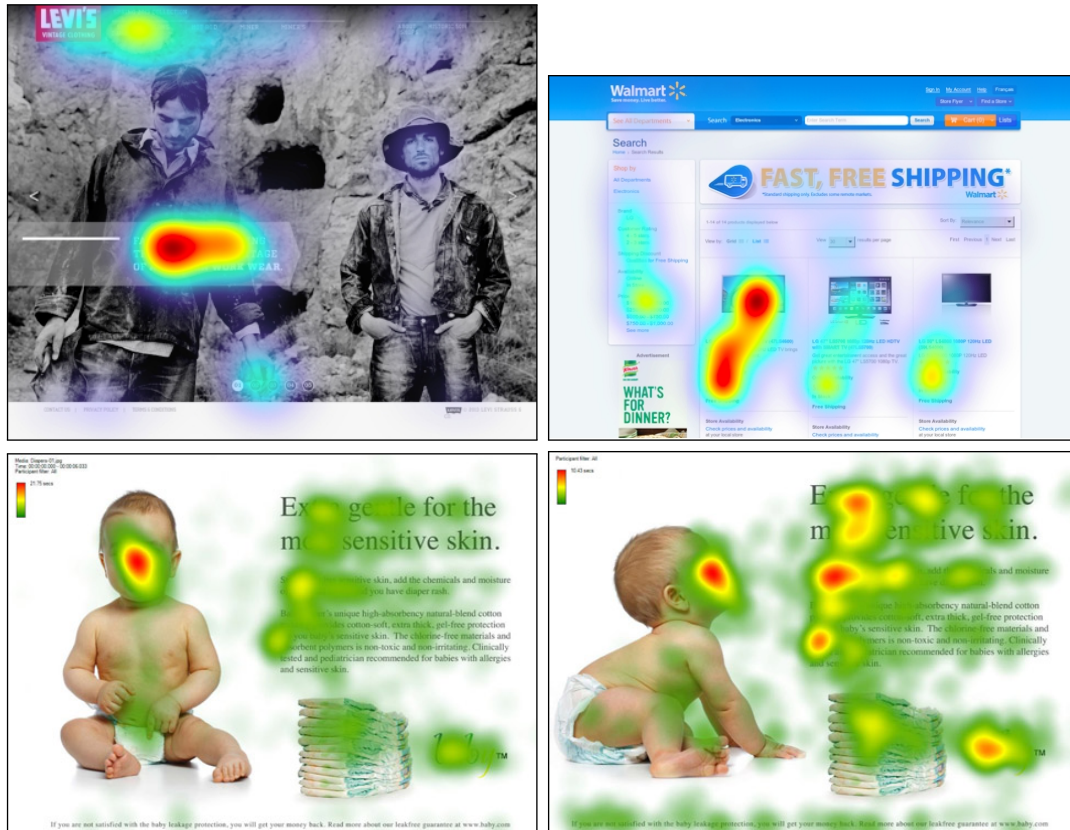


Figure 2.10.: Eye tracking experiments are used to optimize the layout of websites and advertisement. Images from [Usa] and [Eye].

alternative to eye tracking experiments. Potential usage scenarios as proposed by 3M are in-store merchandising, packaging, advertising, web and banner advertisement, and video analysis [3M].



# 3

## Bottom-up Audio-Visual Attention for Scene Exploration

We can differentiate between two attentional mechanisms: First, overt attention directs the sense organs toward salient stimuli to optimize the perception quality. For example, it controls human eye movements in order to project objects of interest onto the fovea of the eye. Second, covert attention focuses the mental processing (*e.g.*, object recognition) of sensory information on the salient stimuli. This is necessary to achieve a high reactivity despite the brain’s limited computational resources that are otherwise unable to process the full amount of incoming sensory information. And, it has been formally shown that this mechanism is an essential aspect of human perception, because it transforms several NP-complete perceptual tasks (*e.g.*, bottom-up perceptual search) into computationally tractable problems [Tso89, Tso95].

Since robots have to deal with limited computational resources and the fact that its sensor orientation influences the quality of incoming sensory information, biologically-inspired models of attention have attracted an increasing interest in the field of robotics to optimize the use of resources and improve perception in complex environments. In this chapter, we present how we implemented bottom-up audio-visual overt attention on the head of KIT’s robot platform ARMAR [ARA<sup>+</sup>06, AWA<sup>+</sup>08]. Our work ranges from developing novel auditory and improved visual saliency models over defining a modality independent 3-dimensional (3D) saliency representation to actually planning in which order the robot should attend salient stimuli. All our methods have in common that they are designed with computational efficiency in mind, because the robot should be able to quickly react to changes in its environment and – in general – a sensible attentional mechanism should not require more resources than it can save.

We first have to define what kind of signals should attract the robot’s attention, *i.e.* what is perceptually “salient”. For this purpose, we developed novel auditory and visual saliency models: Our auditory saliency model is based on Itti and Baldi’s surprise model [IB06]. According to this model, auditory stimuli that are unexpected given the prior frequency distribution are defined as being salient. This model has a biological foundation, because the spectrogram is similar in function to the human Basilar membrane [SIK11] while surprise is related to early

sensory neurons [IB06]. Our visual saliency model also relies on the frequency spectrum and suppresses the image’s amplitude components to highlight salient image regions. This so-called spectral whitening [OL81] accentuates image regions that depict edges and narrow events that stand out from their surround, which can be related back to Treisman’s feature integration theory principles [TG80]. To improve the performance of visual saliency models, we propose to decorrelate the image’s color information. This method not just improves the performance of our algorithms, but also the performance of several other visual saliency algorithms as we have shown on three datasets with respect to three evaluation measures. Furthermore, since it has been suggested that there exists a bottom-up attention mechanism for faces [CFK08], we integrate face detection as a bottom-up visual saliency cue.

To model auditory and visually salient stimuli in a common representation, we build upon the notion of salient proto-objects [WK06] – a concept related to Treisman’s object files, see Ch. 2 – and derive a 3D parametric Gaussian proto-object model. Here, each salient proto-object encodes the perceptual saliency as well as the location and its rough extent of an area in space that is likely to contain an object of interest. To detect and extract salient visual proto-objects, we analyze the saliency map’s isophote curvature to extract salient peaks and their contributing pixels, which can then be used to fit a Gaussian model. Acoustic sound source localization is used to locate salient auditory stimuli and the localization uncertainty is encoded as the proto-object’s spatial extent, *i.e.* spatial variance. Since all auditory and visual stimuli are represented by 3D Gaussian weight functions, we are able to efficiently perform crossmodal clustering and proto-object fusion over space and time. The information contained in each cluster is then fused to implement a biologically-plausible crossmodal integration scheme [OLK07].

Given the salient audio-visual proto-objects, we are able use the encoded information about perceptual saliency and location to plan in which order the robot should attend and analyze the proto-objects. We implemented and compared three strategies: Attending the proto-objects in an order that minimizes ego-motion, attending the proto-objects in the order of decreasing saliency, and performing multiobjective optimization to find a trade-off that suits both criteria, *i.e.* minimize ego-motion while giving priority to highly salient regions.

We demonstrate the applicability of our system and its components in a series of quantitative and qualitative experiments. First, we test the performance of the proposed auditory and visual saliency models. Since our goal was to implement overt attention on a robotic platform, we validate our visual saliency model on human eye tracking data. We show that our approach to visual saliency is state-of-the-art in predicting where humans look in images. Since we could not rely on data analog to eye tracking data to validate our auditory saliency model, we follow a more practice-oriented approach and show that our model is able to reliably detect arbitrary salient auditory events. To validate the overall system behavior, we first performed a series of qualitative active perception experiments.

---

Second, we demonstrate that using multiobjective optimization we can effectively reduce the necessary ego-motion while still assigning a high priority to more salient proto-objects, which results in more efficient scan path patterns.

**Remainder** Complementary to our broad background presentation in Ch. 2, we provide a detailed overview of related work (Sec. 3.1) that is relevant to understand the contributions presented in this chapter. Then, we present and evaluate our visual (Sec. 3.2) and auditory (Sec. 3.3) saliency model. Subsequently, we describe our audio-visual saliency-driven scene exploration system (Sec. 3.4 and 3.5). We discuss how we map the detected salient acoustic events and visually salient regions in a common, modality-independent proto-object representation. Then, we describe how the common salient proto-object representation allows us to fuse this information across modalities. Afterwards, we explain how we can plan the robot’s eye movement based on our audio-visually salient proto-objects.

**Acknowledgment** The work described in this chapter contains results of a collaborative research effort. Fortunately, there exists a clear boundary as to who developed which aspects: We – that means me under the supervision of Rainer Stiefelhagen – focused our research on the attentional methods, which includes auditory and visual saliency models, the 3D representation based on proto-objects as well as biologically plausible crossmodal fusion. In parallel, Benjamin Kühn, supervised by Kristian Kroschel, focused his research on hierarchical, knowledge-driven audio-visual object analysis. Consequently, we focus on attentional aspects in the following and would like to refer the interested reader to Benjamin Kühn’s work for details on the analysis. An exception is the multiobjective exploration path optimization, which represents a joint effort. Since we cannot truly separate the work on the multiobjective exploration path planning, we present the whole approach for the sake of completeness, but we would kindly ask the reader to keep in mind that this part has been joint work.

## 3.1 Related Work and Contributions

---

In the following, we first present the state-of-the-art – excluding our work that is presented in this thesis – for each of the affected research topics. Then, after each topic’s overview, we discuss our contribution with respect to the state-of-the-art.

### 3.1.1. Spectral Visual Saliency

The first spectral approach for visual saliency detection was presented in 2007 by Hou *et al.* [HZ07]. Since then, several spectral saliency models have been proposed (see, *e.g.*, [BZ09, GZ10, GMZ08, PI08b, HZ07, AS10, LLAH11]). Hou *et al.* proposed to use the Fourier transform to calculate the visual saliency of an image. To this end, – processing each color channel separately – the image is Fourier transformed and the magnitude components are attenuated. Then, the inverse Fourier transform is calculated using the manipulated magnitude components in combination with the original phase angles. The saliency map is obtained by calculating the absolute value of each pixel of this inverse transformed image and subsequent Gaussian smoothing. This way Hou *et al.* achieved state-of-the-art performance for salient region (proto-object) detection and psychological test patterns. However, although Hou *et al.* were the first to propose this method for saliency detection, it has been known for at least three decades that suppressing the magnitude components in the frequency domain highlights signal components such as lines, edges, or narrow events (see [OL81, HBD75]).

In 2008 [PI08b], Peters *et al.* analyzed the role of Fourier phase information in predicting visual saliency. They extended the model of Hou *et al.* by linearly combining the saliency of the image at several scales. Then, they analyzed how well this model predicts eye fixations and found that “saliency maps from this model significantly predicted the free-viewing gaze patterns of four observers for 337 images of natural outdoor scenes, fractals, and aerial imagery” [PI08b].

Also in 2008 [GMZ08], Guo *et al.* proposed the use of quaternions as a holistic color image representation for spectral saliency calculation. This was possible because quaternions provide a powerful algebra that allows to realize a hypercomplex Fourier transform [Ell93], which was first demonstrated to be applicable for color image processing by Sangwine [San96, SE00]. Thus, Guo *et al.* were able to Fourier transform the image as a whole and did not have to process each color channel separately. Furthermore, this made it possible to use the scalar part of the quaternion image as 4<sup>th</sup> channel to integrate a motion component. However, in contrast to Hou *et al.*, Guo *et al.* did not preserve any magnitude information and perform a whitening instead. Most interestingly, Guo *et al.* were able to determine salient people in videos and outperformed the models of Itti *et al.* [IKN98] and Walther *et al.* [WK06]. In 2010, a multiresolution attention selection mechanism was introduced, but the definition of the main saliency model remained unchanged [GZ10]. However, most



interestingly, further experiments demonstrated that the approach outperformed several established approaches in predicting eye gaze on still images.

In 2009 [BZ09], Bian *et al.* adapted the work of Guo *et al.* by weighting the quaternion components. Furthermore, they provide a biological justification for spectral visual saliency models and – without any detailed explanation – proposed the use of the YUV color space, in contrast to the use of the previously applied intensity and color opponents (ICOPP) [GMZ08, GZ10], and RGB [HZ07]. This made it possible to outperform the models of Bruce *et al.* [BT09], Gao *et al.* [GMV08], Walther and Koch [WK06], and Itti and Koch [IKN98] when predicting human eye fixations on video sequences.

In 2012 [HHK12], Hou *et al.* proposed and theoretically analyzed the use of the discrete cosine transform (DCT) for spectral saliency detection. They showed that this approach outperforms all other evaluated approaches – including the algorithms of Itti and Koch [IKN98], Bruce and Tsotsos [BT09], Harel *et al.* [HKP07], and Zhang *et al.* [ZTM<sup>+</sup>08] – in predicting human eye fixations on the well-known Toronto dataset [BT09]. Furthermore, Hou *et al.* pointed out the importance of choosing an appropriate color space.

**Contributions:** We combine and extend several aspects of spectral saliency detection algorithms. Analog to Guo *et al.*'s [GMZ08] adaptation of Hou *et al.*'s spectral residual saliency algorithm [HZ07], we extend Hou *et al.*'s DCT image signature approach [HHK12] and use quaternions to represent and process color images in a holistic framework. Consequently, we apply the quaternion discrete cosine transform (QDCT) and signum function to calculate the visual saliency. Furthermore, we integrate and investigate the influence of quaternion component weights as proposed by Bian *et al.* [BZ09], adapt the multiscale model by Peters *et al.* [PI08b], and propose the use of the quaternion eigenaxis and eigenangle for saliency algorithms that rely on the quaternion Fourier transform (*e.g.*, [HZ07, GZ10, GMZ08]). This way, we were able to improve the state-of-the-art in predicting where humans look on three eye tracking datasets – proving the outstanding performance of spectral models for this task, which was not conclusively shown before.

### 3.1.2. Visual Saliency and Color Spaces

As has been noted for spectral saliency models (see Sec. 3.1.1), the chosen base color space can have a significant influence on the performance of bottom-up visual saliency models. Accordingly, different color spaces have been used for saliency models (see also Sec. 3.1.1) such as, for example, RGB (*e.g.*, [HZ07]), CIE Lab (*e.g.*, [HHK12]), and ICOPP (*e.g.*, [GMZ08, GZ10]). The most prominent color space is probably red-green-blue (RGB), which is an additive color space that is suited for most of today's displays (see [Pas08]). The YUV model defines a color space in terms of one luma (Y) and two chrominance (UV) components. Similarly, the CIE 1976 Lab color space has been designed to

approximate human vision and aspires to perceptual uniformity. The simple intensity and red-green/blue-yellow color opponent (ICOPP) model is often used in conjunction with saliency models (*e.g.*, [GMZ08]). The LMS color space models the response of the three types of cones of the human eye, which are named after their sensitivity for long, medium and short wavelengths [SG31]. Whereas Geusebroek *et al.*'s Gaussian color space [GvdBSG01] represents an extension of the differential geometry framework into the spatio-spectral domain.

Decorrelation of color information has been successfully applied for several applications, *e.g.* texture analysis and synthesis [LSL00, HB95], color enhancement [GKW87], and color transfer [RP11]. More importantly, it is highly related to the human visual system and techniques such as the zero-phase transform (ZCA) have been developed and proposed to model aspects of the human visual system [BS97]. Buchsbaum and Gottschalk [BG83] and Ruderman *et al.* [RCC98] found that linear decorrelation of LMS cone responses at a point matches the opponent color coding in the human visual system. However, when modeling the human visual system it is mostly applied in the context of spatio-chromatic decorrelation, *i.e.* local (center-surround) contrast filter operations [BSF11, RCC98, BS97]. Since decorrelation is an important aspect of the human visual system, it has also been part of a few visual saliency models. Duan *et al.* [DWM<sup>+</sup>11] explored the use of principal component analysis (PCA) on image patches, which is closely related Zhou *et al.*'s approach [ZJY12] in which the image is first segmented into patches and then the PCA is used to reduce the patch dimensions to throw out dimensions that are basically noise for the saliency calculation. Similarly, Wu *et al.* [WCD<sup>+</sup>13] propose to use the PCA to attenuate noise as well as to reduce computational complexity. Luo *et al.* [LLLN12] also use the PCA on a block-wise level to differentiate between salient objects and background.

**Contributions:** As a result of our experience with quaternion-based spectral algorithms, we wanted to try the opposite approach: Instead of using quaternions to represent and process the image's color information holistically, we try to decorrelate the information in the color components. To this end, we propose to use a global image-dependent decorrelated color space for visual saliency detection. This way, we are able to improve the performance of all eight visual saliency algorithms that we tested: Itti and Koch's classic model [IKN98], Harel's graph-based visual saliency [HKP07], Hou and Zhang's pure Fourier transform algorithm [HZ07], Hou *et al.*'s DCT image signature [HHK12], Lu and Lim's histogram-based approach [LL12], Achanta's frequency-tuned approach [AHES09], and our own QDCT image signature and quaternion Fourier transform with eigenaxis/eigenangle algorithms.

### 3.1.3. Visual Saliency and Faces

Studies have shown that – independent of the subject's task – when looking at natural images the gaze of observers is attracted to faces (see [CFK08, SS06]).

Even more, there exists evidence that the gaze of infants is attracted by face-like patterns before they can consciously perceive the category of faces [SS06], which is supported by studies of infants as young as 6 weeks that suggest that faces are visually captivating [CC03]. This seems to play a crucial role in early development, especially emotion and social processing (see, *e.g.*, [KJS<sup>+</sup>02]). This early attraction and inability to avoid looking at face-like patterns suggests that there exist bottom-up attention mechanisms for faces [CFK08]. To model this influence, Cerf *et al.* combined traditional visual saliency models – Harel’s graph-based visual saliency (GBVS) [HKP07] and Itti and Koch’s model [IKN98] – with face detections provided by the well-known Viola-Jones detector [CHEK07, CFK09].

**Contributions:** We build on Cerf *et al.*’s work and integrate a scalable Gaussian face model based on modified census transform (MCT) face detection [FE04] into our state-of-the-art low-level visual saliency model. This way, we are able to improve the state-of-the-art in predicting where people look in the presence of faces. Furthermore, considering the face detections and bottom-up visual saliency as two modalities, we investigate the influence of different biologically plausible combination schemes (see [OLK07]).

#### 3.1.4. Auditory Saliency

As has already been addressed in Sec. 2.1.2, in contrast to the vast amount of proposed visual saliency models (*cf.* [FRC10, Tso11]), only few computational bottom-up auditory attention models exist. Most closely related to our work is the model described by Kayser *et al.* [KPLL05], see Fig. 2.7 on page 30, which has later been adopted by Kalinli and Narayanan [KN09]. This model is based on the well-established visual saliency model of Itti and Koch [IKN98] and, most notably, has been successfully applied to speech processing by Kalinli *et al.* [KN09] and, in principle, by Lin *et al.* [LZG<sup>+</sup>12] to allow for faster human acoustic event detection through audio visualization.

**Contributions:** The application of Itti and Koch’s visual saliency model to spectrograms has several drawbacks: First and most importantly, it requires that the spectrogram has elements of the future to detect salient events in the present, which prohibits online detection and – as a consequence – quick reactions to salient acoustic events. This is caused by the inherent down-scaling and filtering in Itti and Koch’s model, which makes precise localization of salient stimuli at the borders problematic. Second, Itti and Koch’s model is computationally expensive, because it requires the calculation and combination of a considerable amount of 2D feature maps at each time step. Third, although Itti and Koch’s saliency model represents an outstanding historical accomplishment, it can hardly

be said to be state-of-the-art [BSI13b]. To account for these drawbacks, we developed auditory Bayesian surprise, see Sec. 3.3.

### 3.1.5. Audio-Visual Saliency-based Exploration

To realize overt audio-visual attention, it is not sufficient to just determine auditory or visually salient stimuli, but it is also necessary to meaningfully and efficiently integrate the information from both modalities. This is a topic that seems to attract increasing attention, however only a relatively modest number of theoretical studies (*e.g.*, [OLK07]) and models have been proposed so far (*e.g.*, [RMDB<sup>+</sup>13, SPG12]). The proposed models rely on the existence of a 2-dimensional (2D) audio-visual saliency map [RMDB<sup>+</sup>13, OLK07], although it is unclear whether a similar representation exists in the human brain and how such a 2D spatial auditory saliency map can be calculated. Furthermore, it is also unclear how such a map could be updated in the presence of ego-motion. However, when realizing overt attention it is important to consider that each shift of the overt focus of attention leads to ego-motion, which partially renders the previously calculated information obsolete [BKMG10]. Accordingly, it is necessary to enable storing and updating the saliency as well as object information in the presence of ego-motion that are caused by overt attention shifts.

Saliency-based overt attention, *i.e.* directing the robot sensors toward salient stimuli, and saliency-based scene exploration has been addressed by several authors in recent years (*e.g.* [MFL<sup>+</sup>07, BKMG10, RLB<sup>+</sup>08, XCKB09, VCSS01, FPB06, DBZ07, YGMG13]). Almost all state-of-the-art systems only consider visual attention (*e.g.* [MFL<sup>+</sup>07, BKMG10, DBZ07, OMS08]), which – among other drawbacks – makes it impossible to react on salient events outside the visual field of view (*cf.* [SRP<sup>+</sup>09]). Most related to our work on audio-visual attention are the approaches by Ruesch *et al.* [RLB<sup>+</sup>08], who implement audio-visual attention for the “iCub” [Rob] robot platform, and Schauerte *et al.* [SRP<sup>+</sup>09], who implement an audio-visually attentive smart room. Both systems use common visual saliency algorithms (see Sec. 2.1.1) and the energy of the audio signal as primitive auditory attention model, due to the absence of applicable, more elaborate auditory saliency models. Ruesch *et al.* [RLB<sup>+</sup>08] use a linear combination of audio-visual stimuli, whereas Schauerte *et al.* [SRP<sup>+</sup>09] use Fuzzy logic [Zad65] to implement a diverse set of audio-visual combinations, including linear combinations. More importantly, Ruesch *et al.* [RLB<sup>+</sup>08] rely on an ego-centric spatial grid representation, *i.e.* azimuth-elevation maps, following the idea of a sensory ego-sphere in which the reference coordinate system is anchored to a fixed point on the robot’s body [FPB06]. Schauerte *et al.* [SRP<sup>+</sup>09] use a 3D voxel representation, which is similar to Meger *et al.*’s 2D occupancy grid representation [MFL<sup>+</sup>07]. Meger *et al.*’s [MFL<sup>+</sup>07] and Schauerte *et al.*’s [SRP<sup>+</sup>09] representation anchor the reference coordinate system to a fixed point in the scene to allow for ego-motion.

In most publications on overt attention, the order in which the objects in the scene are attended is solely based on the perceptual saliency (see, *e.g.*, [RLB<sup>+</sup>08, BZCM08, IK00]; [SRP<sup>+</sup>09]). Accordingly, in each focus of attention selection step, the location with the highest saliency gains the focus of attention and an inhibition of return mechanism ensures that salient regions are not visited twice. However, in many practical applications, it is beneficial to incorporate other aspects into the decision which location to attend next; for example, sensor coverage planning to maximize the coverage of previously unseen areas [MFL<sup>+</sup>07], top-down target information for visual search [OMS08, WAD09, Wel11], transsaccadic memory consistency [WAD11], or a task-dependent spatial bias [DBZ07]. Most related to our work, Saidi *et al.* [SSYK07] and Andreopoulos *et al.* [AHW<sup>+</sup>10] use rating functions for object search such as, most importantly, a motion cost function for sensor alignment.

**Contributions:** Not unlike Kahneman and Treisman’s object files [KTG92], in coherence theory of visual cognition, proto-objects are volatile units of information that can be accessed by selective attention and subsequently validated as actual objects [WK06]. Our audio-visual saliency representation relies on the concept of audio-visual proto-objects, since we propose a parametric object-centred crossmodal 3D model that is based on Gaussian weight functions to represent salient proto-object regions. For this purpose, we use the visual saliency maps’ isophote curvature and stereo vision (see [LHR05]) to extract visual proto-object regions and use sound source localization and salient acoustic event detection to model auditory salient proto-objects. This proto-object model allows for efficient representation, fusion, and update of information. By treating the proto-object regions as primitive, uncategorized object entities in our world model, it is also the foundation to implement our exploration strategies and object-based inhibition of return. Apart from this seamless model integration, our parametric representation has further practical advantages compared to grid representations. Most importantly, every spatial grid representation leads to a spatial quantization and consequently localization error. To reduce this error and increase the model’s quality, it is necessary to increase the grid resolution which typically has a quadratic or cubic impact on the run-time of algorithms that operate on 2D pixels or 3D voxels, respectively. In contrast, our model does not have a quantization error and the run-time of, for example, crossmodal saliency fusion only depends on the number of salient proto-objects, which according to the definition of salient signals is relatively small.

To plan which location to attend next, we use a flexible, multiobjective exploration strategy based on the salient proto-object regions. In our current implementation, our multiobjective target function considers two criteria: the audio-visual saliency and the required head ego-motion. This way, we are able to implement a solution that substantially reduces head ego-motion while it still strongly favors to attend highly salient regions as fast as possible. Furthermore,

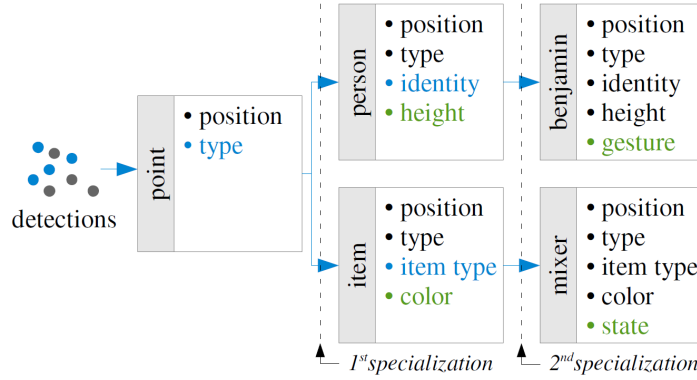


Figure 3.1.: An example to illustrate the principle of the hierarchical object analysis. “Blue” attributes trigger a refinement in the hierarchy and “green” attributes supply additional information about an entity. This illustration is best viewed in color.

the chosen formulation makes it easily possible to integrate additional target criteria, *e.g.* task-specific influences, in the future. Similar to human behavior (*cf.* [Hen03, Wel11]), our exploration considers all salient regions that are present in the short-time memory of our world model, even if they are currently outside the robot’s view. An integrated tracking of proto-objects – which can be linked to already attended objects – makes it possible to detect changes in object saliency and to distinguish novel proto-objects from already attended objects. This way, this makes it possible to seamlessly implement object-based inhibition of return (IoR), which is consistent with human behavior [TDW91] and has the advantage that we are able to realize IoR even for moving targets.

### 3.1.6. Scene Analysis

Fusing the information of different sensors and sensor modalities in order to analyze a scene has been addressed throughout the years in several application areas (see, *e.g.*, [Ess00, MSKK10, KBS<sup>+</sup>10, HY09, HL08]). We build on Machmer *et al.*’s hierarchical, knowledge-driven audio-visual scene analysis approach [MSKK10] that follows an integrated bottom-up and top-down analysis methodology (see [HL08, HY09]). In this framework, the multimodal classification and fusion at each level of the knowledge hierarchy is done bottom-up whereas the appropriate selection of classification algorithms is done in a top-down fashion. The basis for this exploration and analysis is an object-based world model as proposed by Kühn *et al.* [KBS<sup>+</sup>10], which provides an uncertainty-based description for every object attribute. A notable feature of the chosen object analysis approach is that it facilitates the dynamic adjustment of object-specific tracking parameters, *e.g.* for mean shift [CM02], depending on the classification result, *e.g.* person or object specific parameters.

**Contributions:** We integrated saliency-driven, iterative scene exploration into a hierarchical, knowledge-driven audio-visual scene analysis approach that was first presented by Machmer *et al.* [MSKK10], see Fig. 3.1. This, in principle, consistently implements many of the ideas expressed in Treisman’s psychological attention model, see Fig. 2.1, and has not been done to this extent by any other research group so far.

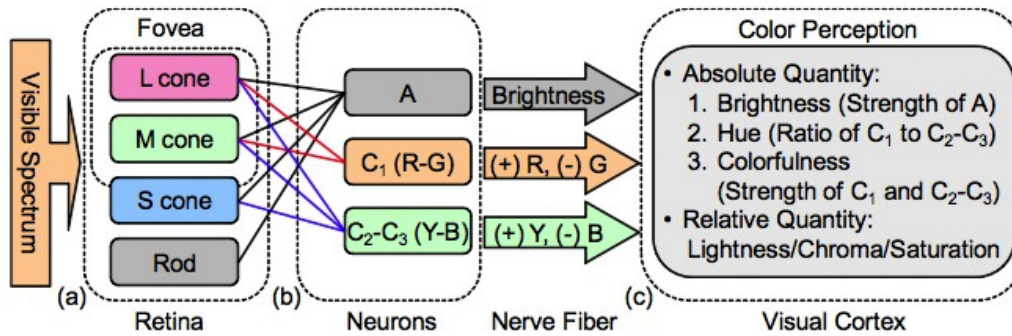


Figure 3.2.: The opponent color model as input to the visual cortex. Image from Wikimedia [Wik].

## 3.2 Visual Attention

Various saliency models have been proposed (see Sec. 2.1.1 and 3.1.1) that vary substantially in what they highlight as being “salient”, *i.e.* what should or is most likely to attract the attention. However, in principle, all saliency models share the same principles and target description, *i.e.* use a set of image features and contrast measures to highlight “sparse” image regions. Such sparse image regions contain or consist of rare features or other irregularities that let them visually “pop out” from the surrounding image context.

In this context, the most fundamental features are color features and edge orientations (see Fig. 2.1, 2.3 and 2.4). Here, the opponent color theory [Her64] forms the theoretical justification for the widely applied opponent color model. In this model, it is suggested that the human visual system interprets information about color based on three opponent channels: red versus green, blue versus yellow, and black versus white, see Fig. 3.2. The latter is achromatic and consequently encodes luminance while the other components encode chrominance. The color opponents can be seen to represent nearly orthogonal axes in an image-independent color space in which red/green, blue/yellow, and white/black form the start/end points of each axis, because – under normal viewing conditions – there exists no hue that humans could describe as a mixture of opponent hues. For example, there exists no hue that appears at the same time red and green (*i.e.*, “redgreen’ish”) or yellow and blue (*i.e.*, “yellowblue’ish”) to a human observer. Compression and efficient coding of sensory signals is another approach to address this aspect and in this context it was found that decorrelation of LMS cone responses at a point matches the opponent color coding in the human visual system [BG83, RCC98]. We further investigate this topic and will first witness the influence of color space on spectral saliency models (Sec. 3.2.1) before we investigate color space decorrelation as a preprocessing or feature encoding step for visual saliency detection (Sec. 3.2.2).



To determine what “pops out” of the feature maps, we rely on spectral visual saliency models. Such models were first proposed to detect proto-objects in images [HZ07] and subsequently it was shown that spectral models also provide an outstanding performance in predicting where people look (*e.g.*, [BZ09, HHK12]). However, real-valued spectral saliency models have – like many other methods – the problem that they process the color channels independently, which can lead to a loss or misrepresentation of information that can result in a suboptimal performance (see, *e.g.*, Fig. 3.3). As an alternative, it is possible to represent images as quaternion matrices and use the quaternion algebra to process the color information as a whole (*e.g.*, [ES07, SE00, BZ09]), *i.e.* holistically. We present how we can calculate the visual saliency based on the quaternion discrete cosine transform and, since not all components might have the same importance for visual saliency, that weighting the quaternion components can improve the performance.

Human attention is not just sensitive to low-level bottom-up features such as, most importantly, color and intensity contrast. Instead, there exists evidence that some complex cues can as well attract human attention independent of task, which suggests that such cues are bottom-up and not top-down features. Most importantly, it has been shown that faces and face-like patterns attract human attention [CFK08]. Consequently, in addition to color features, we integrate face detection as an attentional bottom-up cue in our model (Sec. 3.2.3).

**Remainder** The remainder of this section is organized as follows: First, we present real-valued and quaternion-based spectral saliency detection (Sec. 3.2.1). Then, we introduce color space decorrelation to boost the performance of several visual saliency algorithms (Sec. 3.2.2). Finally, we show how we integrate the influence of faces into our visual attention model (Sec. 3.2.3).

### 3.2.1. Spectral Visual Saliency

Hou *et al.* introduced the spectral residual saliency model to detect salient proto-objects in images [HZ07], which also form a foundation for our audio-visual exploration system as will be addressed in Sec. 3.4. Spectral saliency is based on the idea that “statistical singularities in the spectrum may be responsible for anomalous regions in the image, where proto-objects pop up” [HZ07]. To detect salient image regions, Hou *et al.* [HZ07] attenuate the magnitude in the Fourier frequency spectrum, which in its extreme form leads to a phase-only reconstruction of the image – *i.e.*, the phase-only Fourier transformed signal with unity magnitude – which is known as spectral whitening [OL81]. Although the application of this idea to visual saliency detection was novel, the principle has been widely known for a long time in signal processing theory. As Oppenheim described it [OL81], “since the spectral magnitude of speech and pictures tends to

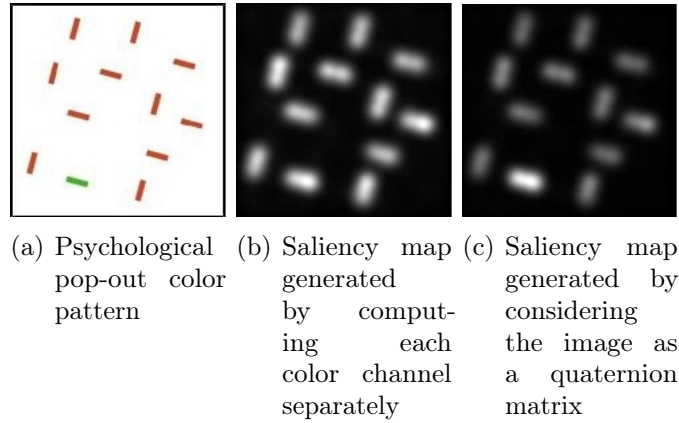


Figure 3.3.: A simple, illustrated example of the disadvantage of processing an image’s color channels separately. Image from [BZ09]. This illustration is best viewed in color.

fall off at high frequencies, the phase-only signal  $f_p(x)$ <sup>1</sup> will, among other effects, experience a high-frequency emphasis which will accentuate lines, edges and other narrow events without modifying their position”. Back in 1981, computational visual saliency has not been an active research field and accordingly Oppenheim focused on other applications such as, *e.g.*, image coding and reconstruction. However, since the basic principle of visual saliency models is to highlight such edges and sparse, narrow image regions, what Oppenheim described in 1981 was a visual saliency model that should become the state-of-the-art 25 years later.

It is possible to calculate the spectral saliency based on the Fourier transform [HZ07] as well as on the cosine transform [HHK12]. Unfortunately, if we want to calculate the spectral saliency for color images, it is necessary to process each image channel separately and subsequently fuse the information. However, since the color space components and consequently the information across the image channels is not independent, this means that color information is involuntarily mishandled or even lost, see Fig. 3.3. An interesting development with regard to this problem is the use of quaternions as a holistic representation to process color images [ES07, GMZ08, BZ09]. The quaternion algebra makes it possible to process color images as a whole without the need to process the image channels separately and, in consequence, tear apart the color information. Interestingly, since the Fourier transform and the cosine transform are also well-defined in the quaternion algebra, we can holistically calculate the spectral saliency based on quaternion color images.

<sup>1</sup>Oppenheim refers to  $\mathcal{F}[f_p(x)] = \frac{1}{|F(\omega)|} \mathcal{F}[f(x)]$  with  $F(\omega) = \mathcal{F}[f](\omega)$ .

### A. Real-valued Spectral Saliency

**Spectral residual and whitening** Given a single-channel image  $I_1$ , we can calculate the phase angle  $P$  and amplitude  $A$  of the image’s Fourier frequency spectrum

$$P = \Phi(\mathcal{F}(I_1)) \quad (3.1)$$

$$A = |\mathcal{F}(I_1)| \quad (3.2)$$

The spectral residual saliency map  $S_{\text{FFT}}$  [HZ07] of the image can then be calculated according to

$$S_{\text{FFT}} = \mathcal{S}_{\text{FFT}}(I_1) = g * |\mathcal{F}^{-1}\{e^{R+iP}\}| \quad \text{with} \quad (3.3)$$

$$L(x, y) = \log A(x, y) \quad \text{and} \quad (3.4)$$

$$R(x, y) = L(x, y) - [h * L](x, y) \quad (3.5)$$

Here,  $\mathcal{F}$  denotes the Fourier transform;  $g$  and  $h$  are Gaussian filter kernels.  $h$  is applied to subtract the smoothed log magnitude, *i.e.*  $h * L$ , from the raw log magnitude  $L$ , which forms the “spectral residual”  $R$ . In principle, this process implements a local contrast operation in the log magnitude matrix, whose strength is defined by the variance  $\sigma_h$  of the Gaussian filter  $h$ .

Shortly after Hou *et al.*’s method was proposed, Guo *et al.* showed [GMZ08] that the influence of the spectral residual itself is negligible in many situations. This means that  $R$  in Eq. 3.3 can be removed

$$S_{\text{PFT}} = \mathcal{S}_{\text{PFT}}(I) = g * |\mathcal{F}^{-1}\{e^{iP}\}|, \quad (3.6)$$

which leads to spectral whitening and is commonly referred to as “pure Fourier transform”. However, spectral whitening can be seen as an extreme case of the spectral residual, because the spectral residual  $R$  approaches 0 when  $\sigma_h$  approaches 0, *i.e.*  $\lim_{\sigma_h \rightarrow 0^+} R = 0$

If we want to process multi-channel color images  $I$ , it is necessary to calculate the saliency of each image channel  $I_c$  and subsequently fuse the maps, because the 2D Fourier transform is only defined for single-channel images. Consequently, the real-valued spectral saliency for color channel images is defined as [HHK12]

$$S_{\text{FFT}}^C = \mathcal{S}_{\text{FFT}}^C(I) = g * \sum_{1 \leq c \leq C} \mathcal{S}_{\text{FFT}}(I_c) \quad (3.7)$$

**DCT image signature** The visual saliency based on discrete cosine transform (DCT) image signatures  $S_{\text{DCT}}$  for a multi-channel image  $I$  is defined as follows [HHK12]:

$$S_{\text{DCT}}^C = \mathcal{S}_{\text{DCT}}^C(I) = g * \sum_{1 \leq c \leq C} [T(I_c) \circ T(I_c)] \quad \text{with} \quad (3.8)$$

$$T(I_c) = \mathcal{D}^{-1}(\text{sgn}(\mathcal{D}(I_c))), \quad (3.9)$$

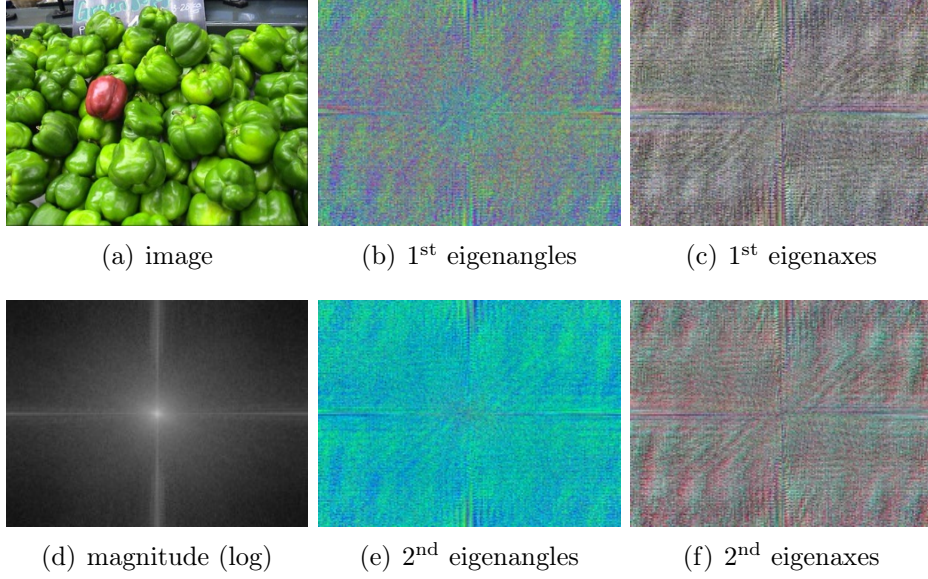


Figure 3.4.: Visualization [ES07] of the quaternion Fourier spectrum of an example image for two transformation axes (1<sup>st</sup> & 2<sup>nd</sup>). This illustration is best viewed in color.

where  $I_c$  is the  $c^{\text{th}}$  image channel,  $\circ$  denotes the Hadamard – *i.e.*, element-wise – product,  $\text{sgn}$  is the signum function,  $\mathcal{D}$  denotes the DCT, and  $g$  is typically a Gaussian smoothing filter. Most notably, it has been formally shown that the DCT image signatures, *i.e.*  $\text{sgn}(\mathcal{D}(I_c))$ , suppress the background and are likely to highlight sparse salient features and objects [HHK12].

## B. Quaternion Image Processing

Quaternions form a 4-dimensional (4D) algebra  $\mathbb{H}$  over the real numbers and are in principle an extension of the 2D complex numbers [Ham66]. A quaternion  $q$  is defined as  $q = a + bi + cj + dk \in \mathbb{H}$  with  $a, b, c, d \in \mathbb{R}$ , where  $i, j$ , and  $k$  provide the basis to define the (Hamilton) product of two quaternions  $q_1$  and  $q_2$  ( $q_1, q_2 \in \mathbb{H}$ ):

$$q_1 q_2 = (a_1 + b_1 i + c_1 j + d_1 k)(a_2 + b_2 i + c_2 j + d_2 k), \quad (3.10)$$

where  $i^2 = j^2 = k^2 = ijk = -1$ . Since, for example, by definition  $ij = k$  while  $ji = -k$  the Hamilton product is not commutative. Accordingly, we have to distinguish between left-sided and right-sided multiplications (marked by L and R, respectively, in the following). A quaternion  $q$  is called real, if  $x = a + 0i + 0j + 0k$ , and pure (imaginary), if  $q = 0 + bi + cj + dk$ . We can define the operators  $S(q) = a$  and  $V(q) = bi + cj + dk$  that extract the scalar

part and the imaginary part of a quaternion  $q = a + bi + cj + dk$ , respectively. As for complex numbers, we can define conjugate quaternions  $\bar{q}$

$$\bar{q} = a - bi - cj - dk \quad (3.11)$$

as well as the norm  $|q|$

$$|q| = \sqrt{q \cdot \bar{q}}. \quad (3.12)$$

Here, a unit quaternion is defined as being a quaternion of norm one. Furthermore, we can define the quaternion scalar product  $*$ :  $\mathbb{H} \times \mathbb{H} \rightarrow \mathbb{R}$

$$s = q_1 * q_2 = a_1 a_2 + b_1 b_2 + c_1 c_2 + d_1 d_2. \quad (3.13)$$

**Eigenaxis and eigenangle** Euler’s formula for the polar representation using the complex exponential generalizes to a (hypercomplex) quaternion form

$$e^{\mu\Phi} = \cos \Phi + \mu \sin \Phi, \quad (3.14)$$

where  $\mu$  is a unit pure quaternion (see [SE00] and [GZ10]). Consequently, any quaternion  $q$  may be represented in a polar representation such as:

$$q = |q|e^{\gamma\Phi} \quad (3.15)$$

with the norm  $|q|$ , its “eigenaxis”  $\gamma$

$$\gamma = f_\gamma(q) = \frac{V(q)}{|V(q)|}, \quad (3.16)$$

and the corresponding “eigenangle”  $\Phi$

$$\Phi = f_\Phi(q) = \arctan \left( \frac{|V(q)| \operatorname{sgn}(V(q) * \gamma)}{S(q)} \right) \quad (3.17)$$

with respect to the eigenaxis  $\gamma$ , which is a unit pure quaternion, and where  $\operatorname{sgn}(\cdot)$  is the signum function (see [SE00]). The eigenaxis  $\gamma$  specifies the quaternion direction in the 3D space of the imaginary, vector part and can be seen as being a generalization of the imaginary unit of complex numbers. Analogously, the eigenangle  $\Phi$  corresponds to the argument of a complex number.

**Quaternion images** Every image  $\mathbf{I} \in \mathbb{R}^{M \times N \times C}$  with at most 4 color components, *i.e.*  $C \leq 4$ , can be represented using a  $M \times N$  quaternion matrix

$$\mathbf{I}_Q = I_4 + I_1 i + I_2 j + I_3 k \quad (3.18)$$

$$= I_4 + I_1 i + (I_2 + I_3 i) j \quad (\text{symplectic form}), \quad (3.19)$$

where  $\mathbf{I}_c$  denotes the  $M \times N$  matrix of the  $c^{\text{th}}$  image channel. It is common to represent the (potential) 4<sup>th</sup> image channel as the scalar part (see, *e.g.*, [SE00]), because when using this definition it is possible to work with pure quaternions for the most common color spaces such as, *e.g.*, RGB, YUV and Lab.

**Quaternion discrete Fourier transform** We can transform a  $M \times N$  quaternion matrix  $\mathbf{f}$  using the definition of the quaternion Fourier transform  $\mathcal{F}_Q^L$  [ES07]:

$$\begin{aligned}\mathcal{F}_Q^L[f](u, v) &= F_Q^L(u, v) \\ F_Q^L(u, v) &= \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-\eta 2\pi((mv/M)+(nu/N))} f(m, n),\end{aligned}\tag{3.20}$$

see Fig. 3.4 for an example. The corresponding inverse quaternion discrete Fourier transform  $\mathcal{F}_Q^{-L}$  is defined as:

$$\begin{aligned}\mathcal{F}_Q^{-L}[F](m, n) &= f_Q^L(m, n) \\ f_Q^L(m, n) &= \frac{1}{\sqrt{MN}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} e^{\eta 2\pi((mv/M)+(nu/N))} F(u, v).\end{aligned}\tag{3.21}$$

Here,  $\eta$  is a unit pure quaternion, *i.e.*  $\eta^2 = -1$ , that serves as an axis and determines a direction in the color space. Although the choice of  $\eta$  is arbitrary, it is not without consequence (see [ES07, Sec. V]). For example, in RGB a good axis candidate would be the “gray line” and thus  $\eta = (i + j + k)/\sqrt{3}$ . In fact, as discussed by Ell and Sangwine [ES07], this would decompose the image into luminance and chrominance components.

**Quaternion discrete cosine transform** Following the definition of the quaternion DCT [FH08], we can transform the  $M \times N$  quaternion matrix  $f$ :

$$\mathcal{D}_Q^L[f](p, q) = \alpha_p^M \alpha_q^N \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \eta f(m, n) \beta_{p,m}^M \beta_{q,n}^N\tag{3.22}$$

$$\mathcal{D}_Q^R[f](p, q) = \alpha_p^M \alpha_q^N \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m, n) \beta_{p,m}^M \beta_{q,n}^N \eta,\tag{3.23}$$

where  $\eta$  is again a unit (pure) quaternion that serves as DCT axis. In accordance with the definition of the traditional type-II DCT, we define  $\alpha$  and  $N$  as follows<sup>2</sup>:

$$\alpha_p^M = \begin{cases} \sqrt{\frac{1}{M}} & \text{for } p = 0 \\ \sqrt{\frac{2}{M}} & \text{for } p \neq 0 \end{cases}\tag{3.24}$$

$$\beta_{p,m}^M = \cos \left[ \frac{\pi}{M} \left( m + \frac{1}{2} \right) p \right].\tag{3.25}$$

<sup>2</sup>From a visual saliency perspective, it is not essential to define the case in  $\alpha$  that handles  $p = 0$ . However, this makes the DCT-II matrix orthogonal, but breaks the direct correspondence with a real-even DFT of half-shifted input. Even more, it is possible to entirely operate without normalization, *i.e.* remove the  $\alpha$  terms, which results in a scale change that is irrelevant for saliency calculation.

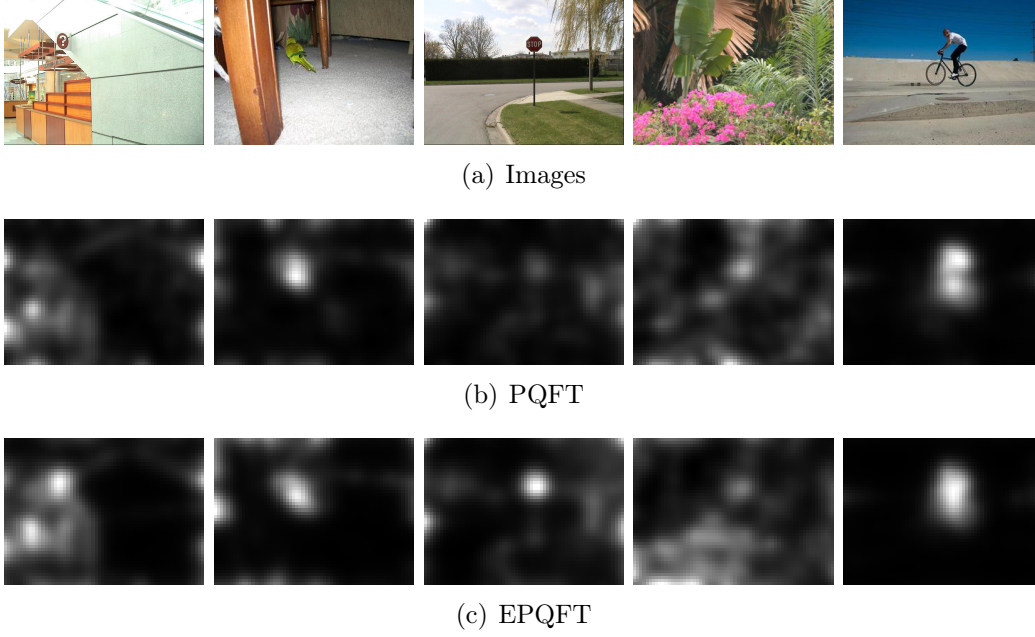


Figure 3.5.: Example images (a) that illustrate the difference between PQFT (b) and our EPQFT (c) saliency maps.

Consequently, the corresponding inverse quaternion DCT is defined as follows:

$$\mathcal{D}_Q^{-L}[F](m, n) = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \alpha_p^M \alpha_q^N \eta F(p, q) \beta_{p,q}^M \beta_{m,n}^N \quad (3.26)$$

$$\mathcal{D}_Q^{-R}[F](m, n) = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \alpha_p^M \alpha_q^N F(p, q) \beta_{p,q}^M \beta_{m,n}^N \eta. \quad (3.27)$$

Again, the choice of the axis  $\eta$  is arbitrary (see [ES07]).

As can be seen when comparing Eq. 3.20 and 3.22, the definition of  $\mathcal{D}_Q^L$  is substantially different from  $\mathcal{F}_Q^L$ , because the factors  $\beta_{u,m}^M$  are real-valued instead of the hypercomplex terms of  $\mathcal{F}_Q^L$ . However, both definitions share the concept of a unit pure quaternion  $\eta$  that serves as a transformation axis.

### C. Quaternion-based Spectral Saliency

**Eigenaxis and -angle Fourier spectral saliency** Similar to the real-numbered definition of the spectral residual by Hou *et al.* [HZ07], let  $A_Q$  denote the amplitude,  $E_\gamma$  the eigenaxes, and the eigenangles  $E_\Theta$  (see Sec. 3.2.1.B) of the quaternion image  $I_Q$ :

$$E_\gamma(x, y) = f_\gamma(I_Q(x, y)) \quad (3.28)$$

$$E_\Theta(x, y) = f_\Theta(I_Q(x, y)) \quad (3.29)$$

$$A_Q(x, y) = |I_Q(x, y)|. \quad (3.30)$$

Then, we calculate the log amplitude and a low-pass filtered log amplitude using a Gaussian filter  $h_{\sigma_A}$  with the standard deviation  $\sigma_A$  to obtain the spectral residual  $R_Q$ :

$$L_Q(x, y) = \log A_Q(x, y) \quad (3.31)$$

$$R_Q(x, y) = L_Q(x, y) - [h_{\sigma_A} * L_Q](x, y). \quad (3.32)$$

Finally, we can calculate the Eigen spectral residual (ESR) saliency map  $S_{\text{ESR}}$  using the spectral residual  $R_Q$ , the eigenaxis  $E_\gamma$ , and the eigenangle  $E_\Theta$ :

$$S_{\text{ESR}} = \mathcal{S}_{\text{ESR}}(I_Q) = h_{\sigma_S} * |\mathcal{F}_Q^{-L} [e^{R_Q + E_\gamma \circ E_\Theta}]|, \quad (3.33)$$

where  $\circ$  denotes the Hadamard product and  $h_{\sigma_S}$  is a real-valued Gauss filter with standard deviation  $\sigma_S$ . If  $\sigma_A$  approaches zero, then the spectral residual  $R_Q$  will become 0, *i.e.*  $\lim_{\sigma_A \rightarrow 0^+} R_Q(x, y) = 0$ , in which case we refer to the model as the Eigen pure quaternion Fourier transform (EPQFT).

If the input image is a single-channel image, then the quaternion definitions and equations are reduced to their real-valued counterparts, in which case Eq. 3.33 is identical to the single-channel real-numbered definitions by Hou *et al.* [HZ07] and Guo *et al.* [GMZ08]. Our ESR and EPQFT definition that is presented in Eq. 3.33 differs from Guo's pure quaternion Fourier transform (PQFT) [GMZ08] definition in two aspects: First, it – in principle – preserves Hou's spectral residual definition [HZ07]. Second, it relies on the combination of the eigenaxes and eigenangles instead of the combination of a single unit pure quaternion and the corresponding phase spectrum (see [GZ10, Eq. 16] and [GMZ08, Eq. 20]), see Fig. 3.5 for an illustration.

**Quaternion DCT image signature saliency** The signum function for quaternions can be considered as the quaternion's “direction” and is defined as follows:

$$\text{sgn}(x) = \begin{cases} \frac{x_0}{|x|} + \frac{x_1}{|x|}i + \frac{x_2}{|x|}j + \frac{x_3}{|x|}k & \text{for } |x| \neq 0 \\ 0 & \text{for } |x| = 0. \end{cases} \quad (3.34)$$

Given that definition, we can transfer the single-channel definition of the DCT signature and derive the visual saliency  $S_{\text{QDCT}}$  using the quaternion DCT signature

$$S_{\text{QDCT}} = \mathcal{S}_{\text{QDCT}}(I_Q) = g * [T(I_Q) \circ \bar{T}(I_Q)] \quad \text{with} \quad (3.35)$$

$$T(I_Q) = \mathcal{D}_Q^{-L}(\text{sgn}(\mathcal{D}_Q^L(I_Q))), \quad (3.36)$$

where again  $h_{\sigma_S}$  is a smoothing Gauss filter with standard deviation  $\sigma_S$ .



### D. Weighted Quaternion Components

As proposed by Bian *et al.* [BZ09], and related to the recent trend to learn feature dimension weights (see, *e.g.*, [ZK11]), we can model the relative importance of the color space components for the visual saliency by introducing a quaternion component weight vector  $w = [w_1 \ w_2 \ w_3 \ w_4]^T$  and adapting Eq. 3.18 appropriately:

$$I_Q = w_4 I_4 + w_1 I_1 i + w_2 I_2 j + w_3 I_3 k. \quad (3.37)$$

In case of equal influence of each color component, *i.e.* uniform weights, Eq. 3.18 is a scaled version of Eq. 3.37, which is practically equivalent for our application.

### E. Multiple Scales

The above spectral saliency definitions only consider a fixed, single scale (see, *e.g.*, [BZ09, GZ10, GMZ08, HHK12]). But, the scale is an important parameter when calculating the visual saliency and an integral part of many saliency models (see, *e.g.*, [FRC10]). For spectral approaches the scale is (implicitly) defined by the resolution of the image  $I_Q$  (see, *e.g.*, [JDT11]). Consequently, as proposed by Peters and Itti [PI08b], it is possible to calculate a multiscale saliency map  $S^M$  by combining the spectral saliency of the image at different image scales. Let  $I_Q^m$  denote the quaternion image at scale  $m \in M$ , then

$$S^M = \mathcal{S}^M(I_Q) = h_{\sigma_M} * \sum_{m \in M} \phi_r(\mathcal{S}(I_Q^m)), \quad (3.38)$$

where  $\phi_r$  rescales the matrix to the target saliency map resolution  $r$  and  $h_{\sigma_M}$  is an additional, optional Gauss filter.

### F. Evaluation

To evaluate the considered saliency algorithms, we use the following eye tracking datasets: Bruce/Toronto [BT09], Kootstra [KNd08], and Judd/MIT [JEDT09]. As evaluation measure, we rely on the AUC, because it is the most widely applied and accepted evaluation measure.

**Datasets** In the last five years, several eye tracking datasets have been made publicly available to evaluate visual attention models (*e.g.*, [KNd08, BT09, CFK09, JEDT09]; see [WS13]). These easily accessible datasets and the resulting quantitative comparability can be seen as the fuel that has led to the plethora of novel visual saliency algorithms. Most importantly, the datasets differ in the choice of images (see Fig. 3.6), the number of images, and the number of observers. While the first aspect defines what can be evaluated (*i.e.*, are top-down or specific dominant bottom-up influences present?), a higher number of images and observers leads to more robust evaluation results, because it reduces the influence of “noise”. Here, “free-viewing” refers to a scenario in which the human



Figure 3.6.: Example images from the visual saliency evaluation datasets.

subjects are not assigned with a task that could lead to a substantial influence of top-down attentional cues such as, for example, to drive a car [BSI13a].

Bruce and Tsotsos’s “Toronto” dataset [BT09] is probably the most widely-used dataset to evaluate visual saliency models. It contains 120 color images ( $681 \times 511$  px) depicting indoor and outdoor scenes. Two image categories are dominant within the Toronto dataset: street scenes and object shots, see Fig. 3.6(a). The dataset contains eye tracking data of 20 subjects (4 seconds, free-viewing).

Judd *et al.*’s “MIT” dataset [JEDT09] contains 1003 images (varying resolutions) selected from Flickr and the LabelMe database, see Fig. 3.6(b). Accordingly, the dataset contains huge variations in the depicted scenes. However, there are two very frequent image types: images that depict landscapes and images that show people. The images (variable resolution) were shown to 15 subjects for 3 seconds with a 1 second gray screen between each two. Eye tracking data was recorded for 15 subjects (3 seconds, free-viewing).

Kootstra *et al.*’s dataset [KNd08] contains 100 images ( $1024 \times 768$  px; collected from the McGill calibrated color image database [OK04]). It contains images from five image categories, close-up as well as landscape images, and images with and without a strong photographer bias, see Fig. 3.6(c). This substantial data variability makes it particularly hard, because it is difficult to find algorithms

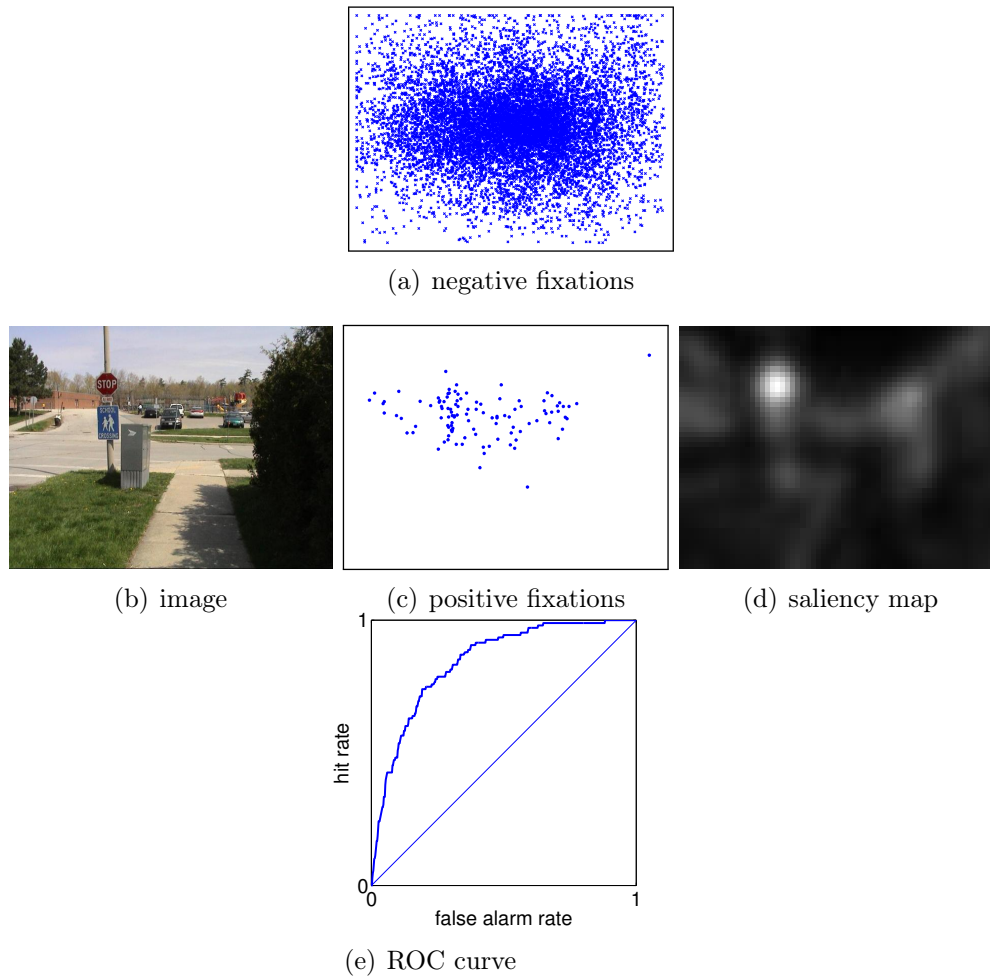


Figure 3.7.: Example image from the Bruce/Toronto dataset to illustrate the components involved when calculating the shuffled, bias-corrected AUC evaluation measure.

that perform well on all of these image types. The images were shown to 31 subjects (free-viewing).

**Evaluation measure** New saliency evaluation measures are proposed regularly and existing measures are sometimes adapted (*e.g.*, [AS13, RDM<sup>+</sup>13, BSI13b, ZK11, JEDT09, PIIK05, PLN02]). Riche *et al.* [RDM<sup>+</sup>13] group visual saliency evaluation measures into three classes. First, value-based metrics such as, for example, normalized scanpath saliency [PIIK05, PLN02], the percentile metric [PI08a], and the percentage of fixations [TOCH06]. Second, several metrics that rely on the area under the receiver operator characteristic curve (*e.g.*, [JEDT09, ZK11, BSI13b]), all of which fall into the group of location-based metrics. Third, there exist distribution-based metrics such as, for example,

the correlation coefficient [JOvW<sup>+</sup>05, RBC06], the Kullback-Leibler divergence [MCB06, TBG05], and the earth mover’s distance [JDT12]. Of these evaluation measures, the dominating and most widely applied evaluation measure is the bias-correcting AUC, which – most importantly – has the distinct advantage that it compensates spatial dataset biases (an aspect that we will encounter again in Ch. 4).

The shuffled, bias-correcting area under the receiver operator characteristic curve (AUROC) calculation (see, *e.g.*, [HHK12]) – commonly referred to as the AUC evaluation measure – tries to compensate for biases such as, *e.g.*, the center-bias that is commonly found in eye tracking datasets. To this end, we define a positive and a negative set of eye fixations for each image, see Fig. 3.7. The positive sample set contains the fixation points of all subjects on that image. The negative sample set contains the union of all eye fixation points across all other images from the same dataset. To calculate the AUROC, we can threshold each saliency map and the resulting binary map can be seen as being a binary classifier that tries to classify positive and negative samples. Sweeping over all thresholds leads to the ROC curves and defines the the area under the ROC curve. When using the AUROC as a measure, the chance level is 0.5 (random classifier), values  $< 0.5$  indicate negative correlation, values  $> 0.5$  represent positive correlation, and an AUROC of 1 means perfect classification. For eye-fixation prediction the maximally achievable, ideal AUROC is typically substantially lower than 1 (*e.g.*,  $\sim 0.88$  on the Bruce/Toronto dataset,  $\sim 0.62$  on Kootstra, and  $\sim 0.90$  on Judd/MIT). The ideal AUROC is calculated by predicting the fixations of one individual using the fixations of other individuals on the same image. In is necessary to say that the calculation of an ideal AUROC requires a Gaussian filter step. Accordingly, the results in the literature can slightly differ due to different filter parameters and should be seen to serve as guiding values of estimated upper baselines. In some publications, authors normalize the results based on the chance and ideal AUROC values (*e.g.*, [ZK11]). However, the most common practice is to report the original AUROC results and, consequently, we follow this established convention in the following. Thus, when interpreting the results, it is important to consider that the actual value range is practically limited by chance at 0.5 (lower baseline) and the ideal AUROC (upper baseline).

**Baseline algorithms and results** Tab. 3.1 shows the results of several baseline algorithms on the three datasets that do not involve face detection, *i.e.* Kootstra, Judd/MIT, and Bruce/Toronto. The algorithms are: Itti and Koch’s model [IKN98] as implemented by the Harel *et al.* (IK’98) and additionally by Itti’s iLab Neuromorphic Vision Toolkit (iNVT’98), Harel *et al.*’s graph-based visual saliency (GBVS’07; [HKP07]), Bruce and Tsotsos’s attention using information maximization (AIM’09; [BT09]), Judd *et al.*’s linear support vector machine (SVM) approach (JEDA’09; [JEDT09]), Goferman *et al.*’s context-aware saliency (CAS’12; [GZMT12]), and Lu and Lim’s color histogram saliency

	Toronto	Kootstra	Judd
CAS'12	0.6921	0.6033	0.6623
CCH'12	0.6663	0.5838	0.6481
JEDA'09	0.6249	0.5497	0.6655
AIM'09	0.6663	0.5747	0.6379
GBVS'07	0.6607	0.5586	0.5844
IK'98	0.6455	0.5742	0.6365
iNVT'98	0.5442	0.5185	0.5365
Chance	0.5	0.5	0.5
Ideal	$\sim 0.88$	$\sim 0.62$	$\sim 0.90$

Table 3.1.: AUC Performance of well-known visual saliency algorithms on the three most-commonly used benchmark datasets.

(CCH'12; [LL12]). Please note that you can find results for further algorithms in, for example, Borji *et al.*'s quantitative saliency evaluation papers (*e.g.*, [BSI13b]). Apart from minor differences, the reported results should be comparable, due to the shared underlying evaluation measure implementation.

**Algorithms** As real-valued spectral saliency algorithms, we evaluate: Hou *et al.*'s spectral residual saliency (SR'07; [HZ07]) and its variant spectral whitening, which is also known as pure Fourier transform and was proposed by Guo *et al.* (PFT'07; [GMZ08, GZ10]). Furthermore, we evaluate Hou *et al.*'s DCT signature saliency (DCT'11; [HHK12]).

As quaternion-based algorithms, we evaluate: Guo *et al.*'s original pure quaternion Fourier transform (PQFT'08; [GMZ08]), which is the quaternion-based counterpart of PFT'07. Our own quaternion-based algorithms, *i.e.* Eigen pure quaternion Fourier transform (EPQFT), which is related to PFT'07 and PQFT'08, Eigen spectral residual (ESR), which is related to SR'07, and quaternion discrete cosine transform image signature saliency (QDCT), which is the quaternion-based counterpart of DCT'11. A preceding  $\Delta$  – imagine a stylized image pyramid – marks algorithms that we evaluated with multiple scales.

We evaluate how well the proposed algorithms perform for all color spaces that have been applied in the literature related to spectral saliency detection: RGB (*e.g.*, [HHK12, HZ07]), ICOPP (*e.g.*, [GZ10, GMZ08]), YUV (*e.g.*, [BZ09]), and CIE Lab (*e.g.*, [HHK12]).

**Parameters** We kept the image resolution fixed at  $64 \times 48$  px in the evaluation, because in preparatory pilot experiments this resolution has constantly shown to provide very good results on all datasets and is the resolution most widely used in the literature (see, *e.g.*, [HHK12]). For multiscale approaches  $64 \times 48$  px is consequently the base resolution. For the Gaussian filtering of the saliency maps,

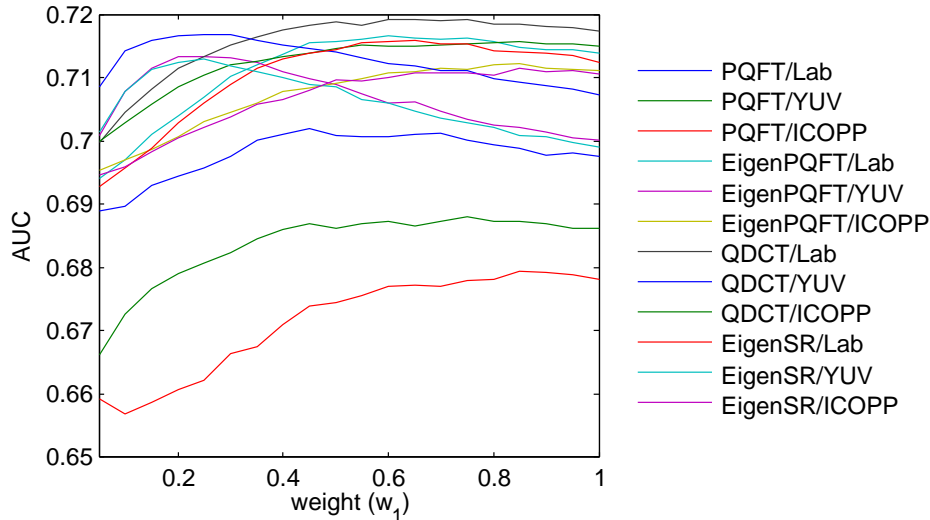


Figure 3.8.: Example of the influence of quaternion color component weights on the AUC performance for QDCT, EPQFT, ESR, and PQFT’08 on the Bruce/Toronto dataset.

we use the fast recursive filter implementation by Geusebroek *et al.* [GSvdW03]. We optimized the filter parameters for all algorithms.

**Results** First, we can see that the performance depends substantially on the dataset, see Tab. 3.2, which is not surprising given their different ideal AUCs. We can rank the datasets by the maximum area under the ROC curve that spectral algorithms achieved and obtain the following descending order: Bruce/Toronto, Judd/MIT, and Kootstra. This order can most likely be explained with the different characteristics of the images in each dataset. Two image categories are dominant within the Bruce/Toronto dataset: street scenes and objects. Within these categories, the images have relatively similar characteristics. The Judd/MIT dataset contains many images from two categories: images that depict landscapes and images that show people. The second category is problematic for low-level approaches that do not consider higher-level influences on visual attention such as, *e.g.*, the presence of people and faces in images (we will address this aspect in Sec. 3.2.3). This also is the reason why Judd *et al.*’s JEDA’09 model performs particularly well on this dataset, see Tab. 3.1. The Kootstra dataset exhibits the highest data variability. It contains five image categories, close-up as well as landscape images, and images with and without a strong photographer bias. Furthermore, we have to consider that the Kootstra dataset has an extremely low ideal AUC of  $\sim 0.62$ . Accordingly, the dataset contains many images in which an image’s recorded gaze patterns vary substantially between persons, which drastically limits the achievable performance.

If we compare the influence of color spaces, then RGB is the color space that leads to the worst performance on all datasets. This is interesting, because it is the only color space in our evaluation that does not try to separate luminance from chrominance information. Interestingly, it appears that the performance difference with respect to the other color spaces (*i.e.*, RGB vs Lab, YUV, or ICOPP) is slightly less for quaternion-based approaches than for real-valued approaches. In other words, quaternion-based approaches seem to be able to achieve better results based on RGB than their real-valued counterparts (see, *e.g.*, the results achieved on the Bruce/Toronto dataset; especially, DCT'11 vs QDCT and PFT'07 vs EPQFT). Most interestingly, as we have mentioned in Sec. 3.2.1.B, the quaternion axis transformation can decompose the RGB color space into luminance and chrominance components [ES07]. Accordingly, it is likely that – at least for RGB as a basis – quaternion-based algorithms benefit from their ability to create an intermediate color space that separates luminance from chrominance information.

Within each color space and across all datasets the performance ranking of the algorithms is relatively stable, see Tab. 3.2. We can observe that without color component weights the performance of the quaternion-based approaches may be lower than the performance of their real-valued counterparts. The extent of this effect depends on the color space as well as on the algorithm. For example, for QDCT this effect does not exist on the RGB and ICOPP color spaces and for Lab only on the Kootstra dataset. However, over all datasets this effect is most apparent for the YUV color space. But, the YUV color space is also the color space that profits most from non-uniform quaternion component weights, see Fig. 3.8, which indicates that the unweighted influence of the luminance component is too high. When weighted appropriately, as mentioned before, we achieve the overall best results using the YUV color space. The influence of quaternion component weights is considerable and depends on the color space, see Tab. 3.2 and Fig. 3.8. As mentioned it is most important for the YUV color space. However, it is also considerable for Lab and ICOPP. Most importantly, we can observe that the best weights are relatively constant over all datasets.

The importance of multiple scales depends on the dataset. The influence is small for the Bruce/Toronto dataset, which can be explained by the fact that the resolution of  $64 \times 48$  pixels is nearly optimal for this dataset (see [HHK12]). On the Kootstra dataset the influence is also relatively small, which may be due to the heterogeneous image data. The highest influence of multiple scales can be seen on the Judd/MIT dataset (*e.g.*, compare  $\Delta$ QDCT with QDCT).

With the exception of Judd *et al.*'s model on the Judd/MIT dataset – as has been discussed earlier –, our quaternion-based spectral approaches are able to perform better than all evaluated non-spectral baseline algorithms (compare Tab. 3.1 and Tab. 3.2). We achieve the best single-scale as well as multiscale performance with the QDCT approach. With respect to their overall performance we can rank the algorithms as follows: QDCT, EPQFT, ESR, and PQFT'08. Especially QDCT performs consistently better than its real-valued counterpart

Method	Toronto				Kootstra				Judd			
	Lab	YUV	ICP	RGB	Lab	YUV	ICP	RGB	Lab	YUV	ICP	RGB
Optimal Color Component Weights												
$\Delta$ QDCT	0.7201	0.7188	0.7174	0.7091	0.6104	0.6125	0.6110	0.6007	0.6589	0.6751	0.6712	0.6622
QDCT	0.7195	0.7170	0.7158	0.7066	0.6085	0.6119	0.6106	0.5994	0.6528	0.6656	0.6623	0.6552
$\Delta$ EPQFT	0.7183	0.7160	0.7144	0.7035	0.6053	0.6082	0.6064	0.5963	0.6527	0.6658	0.6617	0.6559
EPQFT	0.7180	0.7137	0.7122	0.7006	0.6058	0.6073	0.6063	0.5934	0.6483	0.6611	0.6568	0.6493
$\Delta$ ESR	0.7175	0.7153	0.7133	0.7014	0.6050	0.6077	0.6056	0.5941	0.6508	0.6649	0.6603	0.6534
ESR	0.7162	0.7129	0.7112	0.6990	0.6038	0.6068	0.6044	0.5912	0.6467	0.6601	0.6554	0.6470
$\Delta$ PQFT <sup>08</sup>	0.7085	0.6969	0.6927	0.6930	0.5943	0.5994	0.5922	0.5868	0.6467	0.6503	0.6429	0.6468
PQFT <sup>08</sup>	0.7042	0.6881	0.6826	0.6891	0.5930	0.5970	0.5913	0.5861	0.6404	0.6416	0.6379	0.6398
PQFT <sup>08</sup> /Bian [BZ09]	0.7035	0.6880	0.6817	0.6884	0.5928	0.5961	0.5911	0.5861	0.6404	0.6411	0.6375	0.6396
Uniform Color Component Weights												
$\Delta$ QDCT	0.7191	0.7107	0.7070	0.7088	0.6050	0.6036	0.6078	0.6002	0.6539	0.6648	0.6618	0.6620
QDCT	0.7180	0.7079	0.7039	0.7056	0.6036	0.6005	0.6079	0.5987	0.6517	0.6572	0.6552	0.6551
$\Delta$ EPQFT	0.7148	0.7030	0.7024	0.7026	0.6005	0.5963	0.6045	0.5959	0.6490	0.6530	0.6548	0.6556
EPQFT	0.7141	0.7006	0.6982	0.7006	0.5984	0.5939	0.6023	0.5934	0.6461	0.6496	0.6518	0.6491
$\Delta$ ESR	0.7142	0.7135	0.7006	0.7013	0.6003	0.5951	0.6028	0.5937	0.6477	0.6504	0.6534	0.6531
ESR	0.7132	0.6998	0.6969	0.6988	0.5975	0.5930	0.6007	0.5909	0.6448	0.6486	0.6502	0.6466
$\Delta$ PQFT <sup>08</sup>	0.7022	0.6925	0.6868	0.6927	0.5803	0.5826	0.5877	0.5850	0.6431	0.6441	0.6380	0.6465
PQFT <sup>08</sup> [GMZ08]	0.6974	0.6858	0.6796	0.6884	0.5788	0.5808	0.5860	0.5846	0.6368	0.6368	0.6271	0.6396
Non-Quaternion Spectral Algorithms												
DCT <sup>11</sup> [HHK12]	0.7137	0.7131	0.7014	0.6941	0.6052	0.6089	0.6049	0.5907	0.6465	0.6604	0.6556	0.6461
$\Delta$ PFT <sup>07</sup> [P108b]	0.7177	0.7170	0.7079	0.7014	0.6072	0.6107	0.6084	0.5945	0.6502	0.6601	0.6583	0.6523
PFT <sup>07</sup> [GMZ08]	0.7140	0.7120	0.7025	0.6958	0.6057	0.6079	0.6058	0.5908	0.6445	0.6590	0.6572	0.6446
SR <sup>07</sup> [HZ07]	0.7156	0.7144	0.7051	0.6983	0.6059	0.6090	0.6061	0.5916	0.6462	0.6599	0.6573	0.6461

Table 3.2.: AUC performance of the evaluated spectral algorithms. The performance of non-spectral baseline algorithms is presented in Tab. 3.1.



DCT’11, see Tab. 3.2, whereas the situation is not as clear for EPQFT. However, although PFT’07 can achieve slightly better results than EPQFT on the Kootstra dataset, overall EPQFT provides a better performance. Furthermore, we can see that our EPQFT is a substantial improvement over Guo’s PQFT’08 and, in contrast to PQFT’08, it is also able to achieve a better performance than the non-spectral baseline algorithms on all three datasets.

In summary, based on a combination of state-of-the-art quaternion-based saliency detection, quaternion component weights, and multiple scales, we are able to improve the state-of-the-art in predicting human eye fixation patterns.

**Runtime considerations** Spectral saliency algorithms inherit the  $O(N \log_2 N)$  computational complexity of the discrete fast Fourier transform and in practice also benefit from highly optimized fast Fourier implementations. The (quaternion) FFT- and DCT-based models that we evaluated can be implemented to operate in less than one millisecond (single-scale) on an off the shelf PC. For example, in our implementation of (quaternion) DCT image signatures, we use a hard-coded  $64 \times 48$  real DCT-II and DCT-III – the latter is used to calculate the inverse – implementation and are able to calculate the bottom-up saliency map in 0.4 ms on an Intel Core i5 with 2.67 GHz (single-threaded; double-precision). This time excludes the time to subsample or resize the image, which depends on the input image resolution, but includes the time for Gauss filtering. This computational efficiency is an important aspect for practical applications and is only a fraction of the computational requirements of most other visual saliency algorithms. For example, assuming a run-time of 1 ms as baseline, the implementations of Judd *et al.*’s JEDA’09 [JEDT09], Goferman *et al.*’s CAS’12 [GZMT12], and Bruce and Tsotsos’s AIM’09 [BT09] are more than 30,000 $\times$ , 40,000 $\times$ , and 100,000 $\times$  slower, respectively. Furthermore, our implementation is 20 – 50 $\times$  faster than previously reported for spectral saliency algorithms (see [HHK12, Table II] and [GMZ08, Table 3]) and substantially faster than other run-time optimized visual saliency implementations such as, most importantly, Xu *et al.*’s multi-GPU implementation of Itti and Koch’s saliency model (see [XPKB09, Table II]).

### 3.2.2. Color Space Decorrelation

As we have seen previously (Sec. 3.2.1), the input color space influences the performance of spectral saliency algorithms. In this context, we noticed that the color spaces that separate lightness from chrominance information (*e.g.*, CIE Lab and YUV) lead to a better performance than RGB. And, we related the relatively good performance of quaternion-based spectral algorithms on the RGB color space to the quaternion axis transformation, which can decompose an RGB image’s color information into luminance and chrominance components [ES07]. Interestingly, it is known in research fields such as, *e.g.*, color enhancement that the first principal axis – *i.e.*, the first component after applying the PCA –

of an image’s or image patch’s color information describes the major lightness fluctuations in the scene [GKW87], while the second principal axis describes deviations from the mean color. Furthermore, decorrelation of color information has been successfully applied for several applications such as, *e.g.*, texture analysis and synthesis [LSL00, HB95], color enhancement [GKW87], and color transfer [RP11]. For example, it forms the basis of the well-known decorrelation stretch method to image color enhancement (*cf.* [All96]).

Interestingly, evidence suggests that specific signals in the human visual system are subject to decorrelation. For example, spatial decorrelation such as lateral inhibition operations is evident in the human vision system. Particularly, this type of spatial decorrelation results in the visual illusion of Mach bands [Rat65], which exaggerates the contrast between edges of slightly differing shades of gray. Buchsbaum and Gottschalk [BG83] and Ruderman *et al.* [RCC98] found that linear decorrelation of LMS cone responses at a point matches the opponent color coding in the human visual system. Such decorrelation is beneficial for the human visual system, because adjacent spots on the retina will often perceive very similar values, since adjacent image regions tend to be highly correlated in intensity and color. Transmitting this highly-redundant raw sensory information from the eye to the brain would be wasteful and instead the opponent color coding can be seen as performing a decorrelation operation that leads to a less redundant, more efficient image representation. This follows the efficient coding hypothesis of sensory information in the brain [Bar61], according to which the visual system should encode the information presented at the retina with as little redundancy as possible.

We motivated quaternion-based image processing with the wish to being able to process an image’s color information holistically, see Sec. 3.2.1 and Fig. 3.3. We did this, because we did not want to process image channels separately and, in consequence, tear apart the color information. Interestingly, we can see color decorrelation as the opposite approach to this holistic idea, because we use decorrelation to make the information that is encoded in the individual color channels as independent or decorrelated as possible. However, this suggests that color decorrelation has the potential to improve the performance of real-valued saliency algorithms that process color image channels separately.

Thus, in the following, we investigate how we use color decorrelation to provide a better feature space for a diverse set of bottom-up, low-level visual saliency algorithms.

### A. Decorrelation

Let  $I \in \mathbb{R}^{M \times N \times K}$  be the matrix that represents an  $M \times N$  image in a color space with  $K$  components, *i.e.* the image has  $C = K$  color channels. Reshaping the image matrix and subtracting the image’s mean color, we represent the image’s mean centered color information in a color matrix  $X \in \mathbb{R}^{MN \times K}$ .

In general, a matrix  $W$  is a decorrelation matrix, if the covariance matrix of the transformed output  $Y = XW$  satisfies

$$YY^T = \text{diagonal matrix} \quad . \quad (3.39)$$

In general, there will be many decorrelation matrices  $W$  that satisfy Eq. 3.39 and decorrelate [BS97].

The most common approach to decorrelation is the whitening transform, which diagonalizes the empirical sample covariance matrix according to

$$YY^T = C' = WCW^T = I \quad , \quad (3.40)$$

where  $C$  is the color covariance matrix

$$C = \frac{1}{MN} \sum_{i=1}^{MN} x_i x_i^T \text{ with } X = \begin{pmatrix} x_1 \\ \dots \\ x_{MN} \end{pmatrix} \quad . \quad (3.41)$$

Here,  $C'$  is the covariance of  $Y$ , *i.e.* of the data after the whitening transform  $Y = XW$ . As can be seen in Eq. 3.39 and Eq. 3.40, by definition the covariance matrix after a whitening transform equates to the identity matrix, whereas the covariance matrix after an arbitrary decorrelation transform can be any diagonal matrix. Still, there exist multiple solutions for  $W$ . The principal component analysis (PCA) computes the projection according to

$$W_{\text{PCA}} = \Sigma^{-1/2} U^T \quad . \quad (3.42)$$

Here, the eigenvectors of the covariance matrix are the columns of  $\Sigma$  and  $U$  is the diagonal matrix of eigenvalues. As an alternative, the zero-phase transform (ZCA) [BS97] calculates  $W$  according to the symmetrical solution

$$W_{\text{ZCA}} = U \Sigma^{-1/2} U^T \quad . \quad (3.43)$$

The dimensionality preserving color space transform is then given by

$$Y = XW \quad (3.44)$$

and results in the score matrix  $Y$  that represents the projection of the image.

Interestingly, the ZCA was introduced by Bell and Sejnowski [BS97] to model local decorrelation in the human visual system. Although the difference in Eq. 3.42 and Eq. 3.43 seems small, the solutions produced by the PCA and ZCA are substantially different (see, *e.g.*, [BS97, Fig. 3]). Interestingly, the ZCA's additional rotation by  $U$ , *i.e.*  $W_{\text{ZCA}} = UW_{\text{PCA}}$ , causes the whitened data  $Y_{\text{ZCA}} = XW_{\text{ZCA}}$  to be as close to the original data as possible.

We reshape the score matrix  $Y$  so that it spatially corresponds with the original image and this way obtain our color decorrelated image representation  $I_{\text{PCA}} \in \mathbb{R}^{M \times N \times K}$ . Finally, we normalize each color channel's value range to the unit interval  $[0, 1]$ . Although not necessary for all saliency algorithms, it is a beneficial step for algorithms that are sensitive to range differences between color components such as, *e.g.*, Achanta's frequency-tuned algorithm [AHES09]. We can then use the decorrelated image channels as a foundation – *i.e.*, in the sense of raw input or feature maps – for a wide range of visual saliency algorithms.






	better
	better or equal
	probably equal
	equal or worse
	worse

Table 3.3.: Statistical test result classes and visualization color chart.

## B. Quantitative Evaluation

As in the previous evaluation, we rely on the AUC evaluation measure (see Sec. 3.2.1.F), and evaluate on the Bruce/Toronto, Kootstra, and Judd/MIT eye tracking datasets (see Sec. 3.2.1.F). Consequently, the baseline results are again shown in Tab. 3.1. Since color spaces and consequently color decorrelation is such a fundamental aspect that can influence a wide range of algorithms, we take extra precaution to ensure the validity of our claims: First, we evaluate how color space decorrelation influences the performance of eight algorithms. Second, we introduce statistical tests to test the performance of each algorithm on the original color space against the performance based on the image-specific decorrelated color space. Third, although we focus on the AUC evaluation measure in the main body of this thesis, we do not just rely on the AUC as single evaluation measure in this case and, consequently, we present additional results for the NSS and CC evaluation measures in Appx. C.

At this point, we would like to note that this evaluation’s goal is not to assess whether ZCA is a better decorrelation method compared to PCA. Instead, we use two decorrelation methods to indicate that color decorrelation itself is beneficial, independent of a single, specific decorrelation algorithm.

**Statistical tests** We perform statistical significance tests to determine whether or not observed performance differences are significant. Therefore, we record each algorithms prediction for every image and use the evaluation measurements (*e.g.*, AUC) as input data for the statistical tests. We rely on three pairwise, two-sample t-tests to categorize the results: First, we perform a two-tailed test to check whether the compared errors come from distributions with different means (*i.e.*,  $\mathcal{H}_=$ : “means are equal”). Analogously, second, we perform a left-tailed test to check whether an algorithm’s error distribution’s mode is greater (*i.e.*,  $\mathcal{H}_>$ : “mean is greater”) and, third, a right-tailed test to check whether an algorithm’s error distribution’s mode is lower (*i.e.*,  $\mathcal{H}_<$ : “mean is lower”). All tests are performed at a confidence level of 95%, *i.e.*,  $\alpha = 5\%$ .

To simplify the presentation and discussion, we group the test results into five classes, see Tab. 3.3: “Better” means that the hypotheses of equal and worse mean error were rejected. “Better or equal” means that only the hypothesis of a worse mean error could be rejected. “Probably equal” means that no hypothesis could be rejected. “Equal or worse” means that the hypothesis of a better mean

error was rejected. “Worse” means that the hypotheses of equal and better mean error were rejected. Here, “better” and “worse” are defined on the desired characteristic or optimum of the target evaluation measure. For example, we would like to maximize the AUROC and accordingly a higher mean is defined as being better.

**Algorithms** We adapted the following visual saliency algorithms to evaluate the effect of color space decorrelation: We evaluate Itti and Koch’s model (IK’98; [IKN98]) and Harel’s graph-based visual saliency (GBVS’07; [HKP07]). For this purpose, we build on Harel’s implementation, in which both models share the same grounding feature maps that can encode color or orientation information. We evaluate pure Fourier transform (PFT’07; [HZ07]) by Hou and Zhang and DCT image signatures (DCT’11; [HHK12]) by Hou *et al.* Naturally, we also evaluate our own quaternion-based DCT image signatures (QDCT) and EigenPQFT (EPQFT) algorithms, see Sec. 3.2.1. All these algorithms have in common that they are spectral visual saliency algorithms, the first two operate on real-valued images and the latter two process quaternion images. Furthermore, we evaluate the effect on Achanta *et al.*’s (AC’09; [AHES09]) method, which is based on each pixel’s deviation from the image’s mean color. We would like to note that Achanta *et al.*’s algorithm was developed for salient object detection and not eye fixation prediction. Consequently, we do not expect it to achieve state-of-the-art performance on gaze prediction datasets. Nonetheless, we decided to include Achanta *et al.*’s algorithm, because we wanted to evaluate a mix of algorithms that rely on different principles for saliency calculation. Additionally, we implemented and adapted Lu and Lim’s algorithm (CCH’12; [LL12]) that calculates the visual saliency based on the image’s color histogram. Of the above algorithms, IK’98 and GBVS’07 follow the traditional scheme of local center-surround contrast, whereas the spectral approaches (PFT’07, DCT’11, QDCT, and EPQFT), AC’09 and CCH’12 process the image globally.

**Parameters** As in our previous evaluation in Sec. 3.2.1.F, we use an image resolution of  $64 \times 48$  px for spectral saliency approaches. However, in contrast to our previous evaluation, we do not evaluate multiscale approaches, because we have already seen that the integration of multiple scales can further improve the performance. Instead, we focus on the influence of color decorrelation in the following. Therefore, we also use fixed algorithm parameters and do not optimize each algorithm’s parameters for each evaluated color space.

**Results** We present the achieved results for the Bruce/Toronto, Kootstra, and Judd/MIT dataset in Tab. 3.4(a), 3.4(b), and 3.4(c), respectively. To keep our main evaluation compact and readable, we only present the results for RGB, CIE Lab, and ICOPP as base color spaces and base our discussion on the AUC evaluation measure. Results for further color spaces (*e.g.*, Gauss [GvdBSG01])

(a) Bruce/Toronto dataset

AUC Method	RGB			Lab			ICOPP		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.6661	<b>0.7031</b>	0.6974	0.6979	0.7061	<b>0.7072</b>	0.6881	<b>0.7032</b>	0.7019
EPQFT	0.7003	0.7142	<b>0.7158</b>	0.7154	0.7180	<b>0.7212</b>	0.7112	<b>0.7118</b>	<b>0.7156</b>
PFT'07	0.6952	<b>0.7196</b>	0.7135	0.7141	<b>0.7226</b>	<b>0.7226</b>	0.7128	0.7179	<b>0.7189</b>
QDCT	0.7033	<b>0.7157</b>	0.7149	0.7158	0.7187	<b>0.7210</b>	0.7135	0.7140	<b>0.7175</b>
DCT'11	0.6915	<b>0.7196</b>	0.7121	0.7126	<b>0.7208</b>	0.7207	0.7114	<b>0.7184</b>	0.7166
AC'09	0.5406	0.5608	<b>0.5780</b>	0.5541	0.5609	<b>0.5735</b>	0.5510	0.5543	<b>0.5702</b>
GBVS'07	0.6030	<b>0.6620</b>	0.6614	0.6371	<b>0.6665</b>	0.6655	0.6374	<b>0.6637</b>	0.6617
IK'98	0.6410	0.6723	<b>0.6772</b>	0.6612	0.6734	<b>0.6814</b>	0.6636	0.6721	<b>0.6756</b>

(b) Kootstra dataset

AUC Method	RGB			Lab			ICOPP		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.5838	0.6030	<b>0.6045</b>	0.6018	<b>0.6043</b>	0.6037	0.6027	0.6040	<b>0.6042</b>
EPQFT	0.5955	0.6050	<b>0.6140</b>	0.6021	0.6032	<b>0.6069</b>	0.6016	0.6050	<b>0.6070</b>
PFT'07	0.5936	<b>0.6180</b>	0.6147	0.6087	0.6157	<b>0.6172</b>	0.6100	0.6159	<b>0.6190</b>
QDCT	0.5974	0.6068	<b>0.6148</b>	0.6041	0.6049	<b>0.6088</b>	0.6045	0.6069	<b>0.6092</b>
DCT'11	0.5891	<b>0.6148</b>	0.6143	0.6063	0.6126	<b>0.6147</b>	0.6074	0.6134	<b>0.6173</b>
AC'09	0.5415	0.5509	<b>0.5633</b>	0.5464	0.5487	<b>0.5544</b>	0.5463	0.5488	<b>0.5534</b>
GBVS'07	0.5584	<b>0.5897</b>	0.5879	0.5788	<b>0.5914</b>	0.5906	0.5764	<b>0.5912</b>	0.5901
IK'98	0.5740	0.5951	<b>0.5965</b>	0.5882	0.5936	<b>0.5950</b>	0.5881	0.5943	<b>0.5966</b>

(c) Judd/MIT dataset

AUC Method	RGB			Lab			ICOPP		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.6480	0.6696	<b>0.6708</b>	0.6674	<b>0.6733</b>	0.6722	0.6595	<b>0.6705</b>	0.6702
EPQFT	0.6484	0.6590	<b>0.6621</b>	0.6579	0.6581	<b>0.6609</b>	0.6547	0.6558	<b>0.6583</b>
PFT'07	0.6449	<b>0.6652</b>	0.6627	0.6597	<b>0.6653</b>	0.6650	0.6590	0.6639	<b>0.6647</b>
QDCT	0.6517	0.6608	<b>0.6625</b>	0.6599	0.6610	<b>0.6625</b>	0.6585	0.6593	<b>0.6613</b>
DCT'11	0.6440	<b>0.6641</b>	0.6608	0.6581	<b>0.6645</b>	0.6638	0.6577	<b>0.6632</b>	0.6627
AC'09	0.5306	0.5513	<b>0.5810</b>	0.5493	0.5514	<b>0.5592</b>	0.5452	0.5492	<b>0.5585</b>
GBVS'07	0.5846	<b>0.6343</b>	0.6327	0.6207	<b>0.6367</b>	0.6362	0.6162	<b>0.6349</b>	0.6342
IK'98	0.6367	0.6572	<b>0.6585</b>	0.6508	0.6581	<b>0.6582</b>	0.6493	0.6556	<b>0.6564</b>

Table 3.4.: Color space decorrelation results as quantified by the AUC evaluation measure. This table contains color coded information and is best seen in color. Please refer to Tab. 3.3 for a color legend.

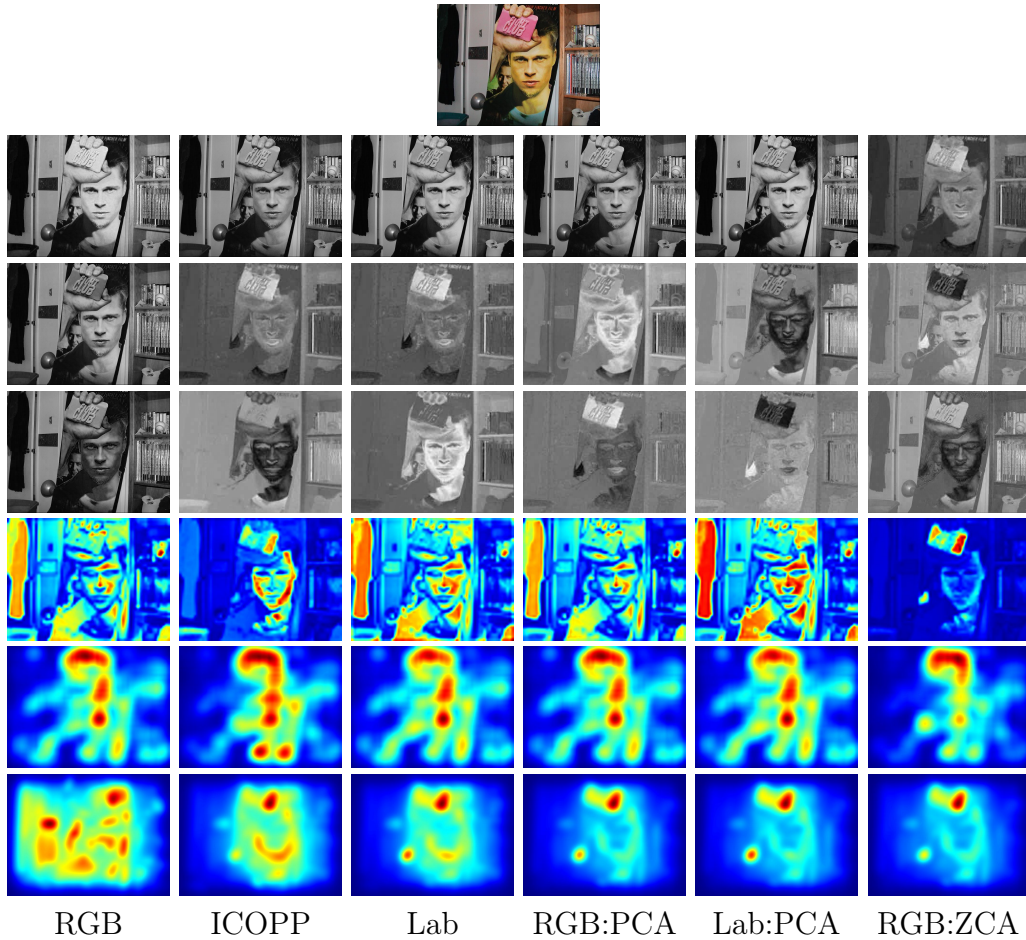


Figure 3.9.: First row: original image; Rows 2-4: 1st, 2nd, and 3rd color component; Rows 5-7: saliency maps for AC'09, QDCT, and GBVS'07 (see Sec. 3.2.2.B). This illustration is best viewed in color.

and LMS [SG31]) and evaluation measures (NSS and CC) are presented in Appx. C. Without going into any detail, the results on these additional color spaces and evaluation measures follow the trend that is visible in Tab. 3.4(a), 3.4(b), and 3.4(c) and thus further substantiate our claim that color space decorrelation is an efficient and robust preprocessing step for many low-level saliency detection algorithms.

As can be seen, the performance of all saliency algorithms improves, if we perform a color space decorrelation. This is independent of the base color space. Even in cases where our statistical tests do not indicate that color space decorrelation improves the results, the mean AUC based on the decorrelated color space is still slightly higher in all cases. Although both evaluated decorrelation methods perform very well, ZCA seems to be slightly better than PCA, because the ZCA leads to the best performance in 44 cases whereas PCA leads to the best performance in 26 cases (the performance is identical for Bruce/Toronto,

PFT'07, and Lab). However, this also seems to depend on the saliency algorithm, because DCT'11 and GBVS'07 appear to benefit more from PCA, since PCA leads to the better performance in 7 of 9 cases for DCT'11 and all 9 of 9 cases for GBVS'07.

Interestingly, color space decorrelation leads to better results than quaternion component weighting on the Bruce/Toronto and Kootstra datasets and roughly equal performance on the Judd/MIT dataset. However, although quaternion-based spectral approaches (QDCT and EPQFT) benefit from color space decorrelation, their real-valued counterparts (DCT'11 and PFT'07) seem to provide a slightly better performance in combination with color decorrelation.

In summary, we strongly suggest to perform color space decorrelation for saliency algorithms, because it can significantly increase the performance while it only requires modest computational resources as we will see in the following.

**Runtime considerations** We can calculate the PCA in 0.82 ms for a  $64 \times 48$  color image on an Intel Core i5 with 2.67 GHz (single-threaded; double-precision), in Matlab. Please note that  $64 \times 48$  px is the default resolution that we use to calculate spectral saliency maps. Here, we use a specialized implementation to calculate the eigenvalues and normalized eigenvectors of hermitian  $3 \times 3$  matrices based on the Jacobi algorithm. In general, the time to perform the color space decorrelation scales linearly with the number of pixels in the input image. Again, this time excludes the time to subsample or resize the image, which depends on the input image resolution, but includes the time that is required to apply the transformation to the image.

### C. Discussion

Given these results, we still have to address what the effects of color decorrelation are and why they can help computational saliency algorithms. For this purpose, we examine the intra and inter color component correlation of color spaces, which is shown for some exemplary color spaces in Tab. 3.5. Here, the intra color component correlation is the correlation of each color space's individual components (*e.g.*, the correlation of the Lab color space's L and a, L and b, or a and b channels). The inter color component correlation refers to the correlation of the components of different color spaces (*e.g.*, the correlation between RGB's R channel and Lab's a channel).

**Does the decorrelated color space depend on the input space?** First of all, the decorrelated color spaces are not independent from their base color spaces. This comes at no surprise, because – for example – an antecedent non-linear transformation such as, *e.g.*, a conversion from RGB to Lab can naturally lead to a different linear decorrelation result, which is illustrated by the low inter component correlation of RGB:PCA and Lab:PCA in Tab. 3.5. As a result,



	RGB			Lab			ICOPP			LMS			YUV			RGB:PCA			Lab:PCA			RGB:ZCA			Lab:ZCA		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
RGB	1.00	0.88	0.79	0.95	0.32	0.24	0.94	0.39	0.28	0.93	0.86	0.78	0.94	0.28	0.41	0.88	0.05	0.02	0.21	0.07	0.04	0.79	0.45	0.36	0.92	0.26	0.15
RGB	0.88	1.00	0.89	0.98	0.06	0.09	0.97	0.00	0.16	0.94	0.96	0.88	0.99	0.15	0.03	0.91	0.00	0.03	0.27	0.05	0.00	0.50	0.70	0.47	0.96	0.11	0.03
RGB	0.79	0.89	1.00	0.88	0.01	0.25	0.93	0.00	0.19	0.86	0.88	0.95	0.90	0.19	0.02	0.89	0.04	0.06	0.36	0.01	0.02	0.40	0.47	0.76	0.91	0.00	0.33
Lab	0.95	0.98	0.88	1.00	0.07	0.14	0.99	0.14	0.19	0.96	0.94	0.86	1.00	0.19	0.16	0.92	0.02	0.02	0.26	0.06	0.02	0.61	0.63	0.45	0.97	0.02	0.06
Lab	0.32	0.06	0.01	0.07	1.00	0.24	0.10	0.96	0.24	0.10	0.05	0.02	0.07	0.24	0.91	0.08	0.01	0.07	0.10	0.02	0.15	0.73	0.56	0.03	0.05	0.92	0.12
Lab	0.24	0.09	0.25	0.14	0.24	1.00	0.04	0.42	0.96	0.11	0.05	0.20	0.11	0.99	0.53	0.01	0.11	0.24	0.33	0.13	0.05	0.47	0.28	0.71	0.09	0.11	0.92
ICOPP	0.94	0.97	0.93	0.99	0.10	0.04	1.00	0.15	0.10	0.96	0.95	0.91	0.99	0.10	0.16	0.93	0.00	0.00	0.28	0.05	0.02	0.60	0.57	0.55	0.97	0.06	0.04
ICOPP	0.39	0.00	0.00	0.14	0.96	0.42	0.15	1.00	0.43	0.16	0.01	0.02	0.13	0.43	0.99	0.12	0.05	0.02	0.15	0.06	0.14	0.80	0.43	0.16	0.11	0.87	0.31
ICOPP	0.28	0.16	0.19	0.19	0.24	0.96	0.10	0.43	1.00	0.17	0.11	0.15	0.17	0.99	0.54	0.06	0.11	0.24	0.30	0.14	0.05	0.50	0.31	0.68	0.14	0.11	0.89
LMS	0.93	0.94	0.86	0.96	0.10	0.11	0.96	0.16	0.17	1.00	0.98	0.90	0.97	0.16	0.18	0.90	0.00	0.02	0.24	0.05	0.01	0.61	0.58	0.46	0.94	0.05	0.03
LMS	0.86	0.96	0.88	0.94	0.05	0.05	0.95	0.01	0.11	0.98	1.00	0.93	0.96	0.11	0.03	0.89	0.01	0.02	0.26	0.05	0.01	0.48	0.66	0.50	0.93	0.09	0.02
LMS	0.78	0.88	0.95	0.86	0.02	0.20	0.91	0.02	0.15	0.90	0.93	1.00	0.88	0.15	0.04	0.86	0.04	0.05	0.33	0.02	0.01	0.38	0.50	0.71	0.88	0.04	0.28
RGB:PCA	0.88	0.91	0.89	0.92	0.08	0.01	0.93	0.12	0.06	0.90	0.89	0.86	0.93	0.06	0.13	1.00	0.00	0.00	0.32	0.01	0.01	0.55	0.53	0.52	0.90	0.05	0.05
RGB:PCA	0.05	0.00	0.04	0.02	0.01	0.11	0.00	0.05	0.11	0.00	0.01	0.04	0.01	0.11	0.09	0.00	1.00	0.00	0.02	0.21	0.05	0.11	0.00	0.11	0.02	0.07	0.13
RGB:PCA	0.02	0.03	0.06	0.02	0.07	0.24	0.00	0.02	0.24	0.02	0.02	0.05	0.02	0.24	0.05	0.00	0.00	1.00	0.02	0.04	0.00	0.06	0.27	0.34	0.01	0.15	0.40
Lab:PCA	0.21	0.27	0.36	0.26	0.10	0.33	0.28	0.15	0.30	0.24	0.26	0.33	0.26	0.30	0.19	0.32	0.02	0.02	1.00	0.00	0.00	0.07	0.15	0.29	0.27	0.05	0.20
Lab:PCA	0.07	0.05	0.01	0.06	0.02	0.13	0.05	0.06	0.14	0.05	0.05	0.02	0.05	0.14	0.08	0.01	0.21	0.04	0.00	1.00	0.00	0.11	0.06	0.07	0.08	0.03	0.12
Lab:PCA	0.04	0.00	0.02	0.02	0.15	0.05	0.02	0.14	0.05	0.01	0.01	0.01	0.01	0.05	0.13	0.01	0.05	0.00	0.00	0.00	1.00	0.16	0.13	0.03	0.04	0.21	0.00
RGB:ZCA	0.79	0.50	0.40	0.61	0.73	0.47	0.60	0.80	0.50	0.61	0.48	0.38	0.60	0.49	0.82	0.55	0.11	0.06	0.07	0.11	0.16	1.00	0.00	0.00	0.58	0.71	0.38
RGB:ZCA	0.45	0.70	0.47	0.63	0.56	0.28	0.57	0.43	0.31	0.58	0.66	0.50	0.62	0.32	0.35	0.53	0.00	0.27	0.15	0.06	0.13	0.00	1.00	0.00	0.63	0.69	0.33
RGB:ZCA	0.36	0.47	0.76	0.45	0.03	0.71	0.55	0.16	0.68	0.46	0.50	0.71	0.48	0.69	0.25	0.52	0.11	0.34	0.29	0.07	0.03	0.00	0.00	1.00	0.51	0.04	0.84
Lab:ZCA	0.92	0.96	0.91	0.97	0.05	0.09	0.97	0.11	0.14	0.94	0.93	0.88	0.98	0.14	0.13	0.90	0.02	0.01	0.27	0.08	0.04	0.58	0.63	0.51	1.00	0.00	0.00
Lab:ZCA	0.26	0.11	0.00	0.02	0.92	0.11	0.06	0.87	0.11	0.05	0.09	0.04	0.02	0.10	0.82	0.05	0.07	0.15	0.05	0.03	0.21	0.71	0.69	0.04	0.00	1.00	0.00
Lab:ZCA	0.15	0.03	0.33	0.06	0.12	0.92	0.04	0.31	0.89	0.03	0.02	0.28	0.03	0.90	0.43	0.05	0.13	0.40	0.20	0.12	0.00	0.38	0.33	0.84	0.00	0.00	1.00

Table 3.5.: Mean correlation strength (*i.e.*, absolute correlation value) of color space components calculated over all images in the McGill calibrated color image database [OK04]. This table contains color coded information and is best seen in color.

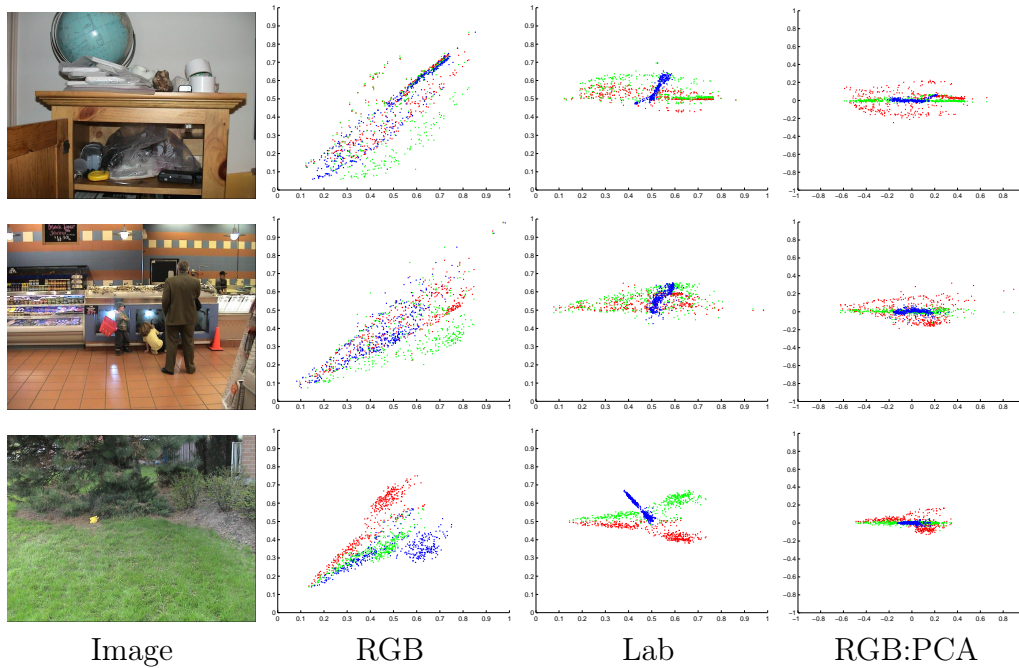


Figure 3.10.: The degree of intra color component correlation is indicated by the angle of rotation of each point cloud’s mean axis. Rotations close to  $0^\circ$  or  $90^\circ$  indicate uncorrelated data and rotations in between indicate degrees of correlation (red: 1st vs 2nd component; green: 1st vs 3rd; blue: 2nd vs 3rd). Visualization method according to Reinhard *et al.* [RAGS01]. This illustration is best viewed in color.

we have to neglect the notion of a unique, base color space independent color projection.

**Are PCA and ZCA different?** It becomes apparent in the rightmost column of Fig. 3.9 as well as by the inter color component correlation between RGB:ZCA and RGB or RGB:ZCA and RGB:PCA (see Tab. 3.5) that ZCA color projections differ substantially from PCA projections, because the ZCA does not seem to separate luminance and chrominance information. This is of interest, because it indicates that not the separation of color and luminance itself is the key to improve the performance, but the properties of decorrelated color information.

Furthermore, we can see that the color components of ZCA projections (*e.g.*, RGB:ZCA or Lab:ZCA) are highly correlated to their base color spaces, see Tab. 3.5, which stands in contrast to the behavior of PCA projections.

**What is the effect of decorrelation?** In fact, there are two aspects of color decorrelation that can influence saliency detection: First, the color information contained in the channels is as decorrelated and thus independent as possible.

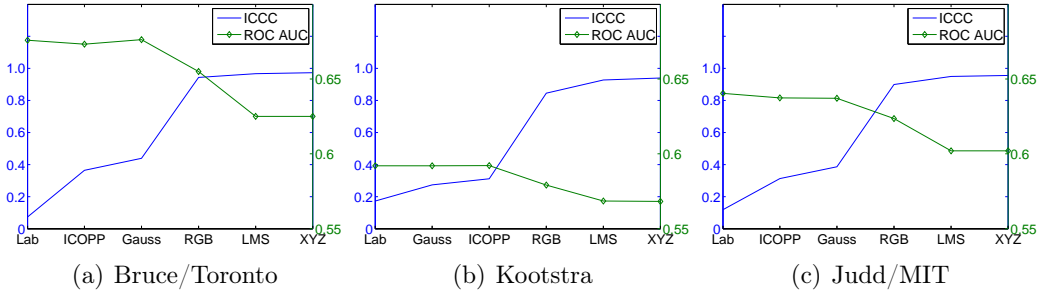


Figure 3.11.: Mean gaze prediction performance of the evaluated visual saliency algorithms (see Sec. 3.2.2.B) and the intra color component correlation (ICCC) for several well-known color spaces.

This naturally supports algorithms that process the color channels independently such as, *e.g.*, DCT’11. Second, algorithms that use color distances (*e.g.*, AC’09) benefit from the aspect that color decorrelation can enhance the contrast of highly correlated images, which is the foundation of the well-known decorrelation stretch color enhancement algorithm (see [GKW87]). In case of the PCA, this is due to the fact that the stretched and thus expanded color point cloud in the decorrelated space is less dense and spread more evenly over a wider volume of the available color space (see, *e.g.*, [GKW87]).

**Does intra color component correlation influence the performance of saliency algorithms?** In general, color spaces exhibit different degrees of intra color component correlation, an aspect that is illustrated in Fig. 3.10 and apparent in Tab. 3.5. Here, our image-specific color decorrelation forms an extreme case – *i.e.*, the intra color component correlation is zero – for which we have demonstrated that it can significantly increase the performance with respect to its base color space. Since it has been noted by several authors that the choice of color spaces directly influences the performance of visual saliency detection algorithms (*e.g.*, [HHK12]), an interesting question is whether or not these performance differences could be related to the color space’s degree of intra color component correlation.

To address this question, we calculated the mean performance over all evaluated visual saliency algorithms and the mean intra color component correlation for six well-known color spaces. As we can see in Fig. 3.11, there seems to exist a relation between the average saliency detection performance and the underlying color space’s correlation. To quantify this observation, we calculate the correlation between the intra color component correlation and the mean AUC, which is  $-0.8619$ ,  $-0.9598$ , and  $-0.9067$  on the Bruce/Toronto, Kootstra, and Judd/MIT dataset, respectively. Such an overall high negative correlation indicates that a lower visual saliency detection performance can be related to a higher intra color component correlation of the underlying color space.

### 3.2.3. Modeling the Influence of Faces

Up until this point, we have only considered the influence that low-level image features have on visual saliency. However, it has been shown that the gaze of human observers is attracted to faces, even if faces are not relevant for their given task [SS06]. The visual attraction of faces and face-like patterns can already be observed in infants as young as six weeks, which means that infants are attracted by faces before they are able to consciously perceive the category of faces [SS06]. This suggests nothing less than that there exists a bottom-up attentional mechanism for faces [CFK08]. This comes at no surprise, since the perception of the caregivers' faces is an important aspect in early human development, especially for emotion and social processing [KJS<sup>+</sup>02]. For example, observing the caregiver's face also means to observe the caregiver's eyes and consequently eye gaze, which is essential to start to follow the gaze direction (*cf.* Sec. 2.1.3 and Sec. 4.1). And, following the gaze direction while a caregiver talks about an object in the infant's environment is important, because the relation between gaze and objects is one of the early cues that allow a child to slowly associate spoken words with objects – a key ability to being able to learn a language.

Accordingly, we want to integrate the influence of faces into our computational bottom-up visual saliency model. To this end, we rely face detection and a Gaussian face map to model the presence of faces in an image. For this purpose, without going into any detail, we use Fröba and Ernst's face detection algorithm [FE04] that relies on the modified census transform (MCT) and is known to provide high performance face detections in combination with a very low false positive rate in varying illumination conditions. The output of the face detection algorithm is a set of bounding boxes, each of which reflects the position, size, and orientation of a detected face in the image.

#### A. Face Detection and the Face Conspicuity Map

In Cerf *et al.*'s model [CHEK07, CFK09], each detected face is modeled in the face conspicuity map by a circular 2D Gaussian weight function with a standard deviation of  $\sigma = \sqrt{(w + h)/4}$ , where  $w$  and  $h$  is the width and height, respectively, of Viola-Jones face detection's bounding box. We extend this model in two ways: First, we allow an in-plane rotation  $\theta$  of the face bounding boxes provided by our modified census transform (MCT) detectors. Then, we use an elliptical 2D Gaussian weight function  $g_0$ , where  $\sigma_u$  and  $\sigma_v$  is the standard deviation in the direction parallel and orthogonal, respectively, to the orientation  $\theta$ :

$$g_0(u, v, \sigma_u, \sigma_v) = \frac{1}{\sqrt{2\pi}\sigma_u} \exp\left\{-\frac{1}{2} \frac{u^2}{\sigma_u^2}\right\} \quad (3.45)$$

$$* \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left\{-\frac{1}{2} \frac{v^2}{\sigma_v^2}\right\},$$

where the  $u$ -axis corresponds to the direction of  $\theta$  and the  $v$ -axis is orthogonal to  $\theta$ , *i.e.*

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \hat{\theta}(x) \\ \hat{\theta}(y) \end{pmatrix}. \quad (3.46)$$

Accordingly, we can calculate the face conspicuity map  $S_F$

$$S_F(x, y) = \sum_{1 \leq i \leq N} g_0(\hat{\theta}(x - x_i), \hat{\theta}(y - y_i), \sigma_{u,i}, \sigma_{v,i}, \theta), \quad (3.47)$$

where  $(x_i, y_i)$  is the detected center of face  $i$  with orientation  $\theta_i$  and the standard deviations  $\sigma_{u,i}$  and  $\sigma_{v,i}$ . Since, depending on the detector training, the width and height of the bounding box may not be directly equivalent to the optimal standard deviation, we calculate  $\sigma_u$  and  $\sigma_v$  by scaling  $w$  and  $h$  with the scale factors  $s_w$  and  $s_h$  that we experimentally determined for our MCT detectors.

## B. Integration

Interpreting the calculated visual saliency map  $S_V$  and the face detections represented in  $S_F$  as two separate low-level modalities, we have to consider several biologically plausible multimodal integration schemes (*cf.* [OLK07]):

**Linear** We can use a linear combination

$$S_+ = w_V S_V + w_F S_F \quad (3.48)$$

as applied by Cerf *et al.* [CHEK07, CFK09]. However, in contrast to Cerf *et al.*, we analyze the weight space in order to determine weights that provide optimal performance in practical applications. Therefore, we normalize the value range of the saliency map  $S_V$  and use a convex combination, *i.e.*  $w_V + w_F = 1$  with  $w_V, w_F \in [0, 1]$ . From an information theoretic point of view, the linear combination is optimal in the sense that the information gain equals the sum of the unimodal information gains [OLK07].

**Sub-linear (late combination)** When considering a late combination scheme, no true crossmodal integration occurs. Instead, the candidate fixation points from the two unimodal saliency maps compete against each other. Given saliency maps, we can use the maximum to realize such a late combination scheme, resulting in a sub-linear combination

$$S_{\max} = \max \{S_V, S_F\}. \quad (3.49)$$

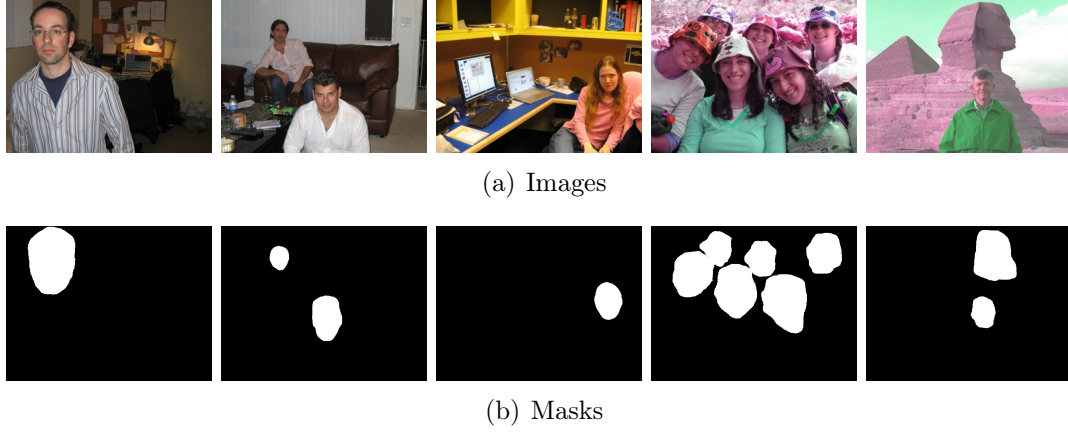


Figure 3.12.: Example images from the Cerf/FIFA dataset with their annotated face segments.

**Supra-linear (early interaction)** Early interaction assumes that there has been crossmodal sensory interaction at an early stage, before the saliency computation and focus of attention selection, which imposes an expansive non-linearity. As an alternative model, this can be realized using a multiplicative integration of the unimodal saliency maps

$$S_o = S_V \circ S_F. \quad (3.50)$$

**Quaternion face channel** From a technical perspective, if the image’s color space has less than 4 channels, we can also use the quaternion scalar part to explicitly represent faces and obtain an integrated or holistic quaternion-based saliency map

$$S_Q = \mathcal{S}_{\text{QDCT}}(I_{\text{QF}}) \quad \text{with} \quad (3.51)$$

$$I_{\text{QF}} = S_F + I_Q = S_F + I_1i + I_2j + I_3k. \quad (3.52)$$

### C. Evaluation

**Dataset** To evaluate the integration of faces and face detection, we rely on Cerf *et al.*’s Cerf/FIFA dataset [CFK09]. The dataset consists of eye tracking data (2 seconds, free-viewing) of 9 subjects for 200 ( $1024 \times 768$  px) images of which 157 contain one or more faces, see Fig. 3.12. Additionally, the dataset provides human annotations of the location and size of faces in the images, which can be used to evaluate the influence between perfect, *i.e.* manual, and automatic face detection.

**Procedure** To use the annotated face masks, see Fig. 3.12, as input to our and Cerf’s face model, we calculate the principal directions and size of each

binary face region. For this purpose, we fit a 2D ellipse that has same normalized second central moments (*i.e.*, spatial variance) as the region. Then, we use the ellipse’s major axis length as the face’s height and its minor axis length as the face’s width; *i.e.*, we assume that a typical face’s height is longer than its width. Furthermore, we assume that the ellipse’s rotation is identical to the face’s orientation.

**Algorithms** Since graph-based visual saliency (GBVS) was reported to perform better than Itti and Koch’s model [IKN98] when combined with face detections [CHEK07], we compare our system to GBVS’07. As an additional baseline, we include the results reported by Zhao and Koch [ZK11], who used an optimally weighted Itti-Koch model with center bias. We refrain from reporting the evaluation results for all previously evaluated saliency algorithms on the Cerf/FIFA dataset (see Sec. 3.2.1.F, Sec. 3.2.2.B, and Appx. C), because we would like to focus the evaluation on the integration of faces and, most importantly, we have already shown the state-of-the-art performance of spectral saliency detection, see Sec. 3.2.2. We report the results for QDCT, EPQFT, PFT’07, and DCT’11. The spectral saliency resolution is set to  $64 \times 48$  pixels, we rely on the Lab color space with ZCA decorrelation. and the Gaussian filter’s standard deviation is 2.5. The standard deviation was set based on the results of a preliminary experiment, in which we independently optimized the spectral saliency filter parameters and the face model parameters.

**Results** It can be seen in Tab. 3.6 that the face map itself has a considerable predictive power, which confirms the observation made by Cerf *et al.* [CHEK07]. In a few instances, we can even observe that the AUC is higher when using automatic, MCT-based face detection instead of optimal, annotated bounding boxes calculated from the manually annotated face regions. This can be explained by the fact that false positives usually occur on complex image patches that are also likely to attract the attention. Accordingly, false positives do not necessarily have a negative impact on the evaluation results. The linear combination of the bottom-up visual saliency and the face conspicuity map substantially improve the results and we achieve the best results with our scaled elliptical Gauss model. If we look at the results for the two best integration schemes, *i.e.* linear and late, we can see that our adapted face model is better in all cases but one (GBVS with late combination and face annotations).

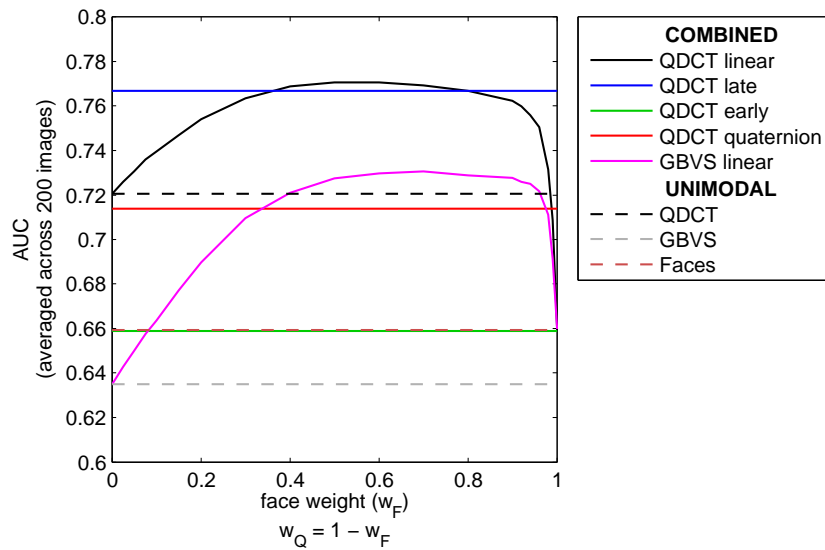
If we use the ideal, *i.e.* human, AUROC to calculate the normalized normalized area under the receiver operator characteristic curve (nAUROC) of our best result with MCT face detections, we obtain an nAUROC of 0.978 which is also higher than the most recently reported 0.962 by Zhao and Koch [ZK11, see Table 1].

The chosen multimodal integration scheme has a considerable influence on the performance, see Fig. 3.13 and Tab. 3.6. The linear combination achieves

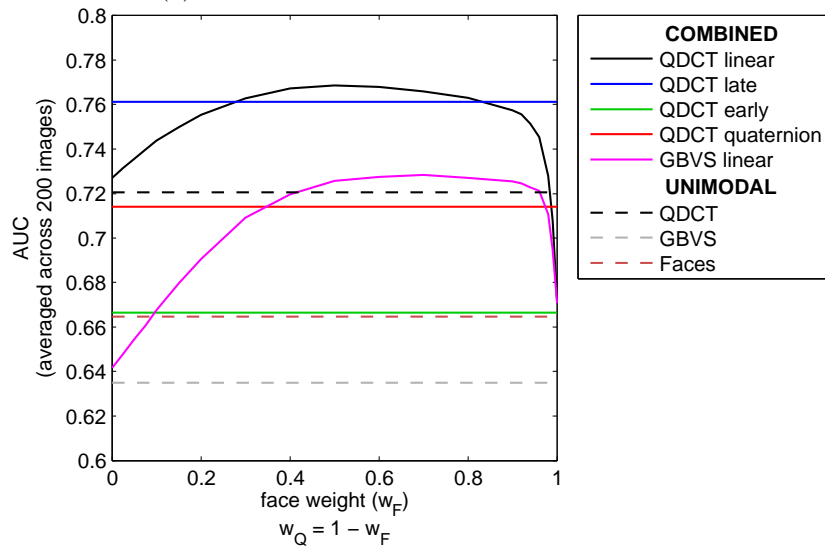
Face Detection Face Model	Annotated		MCT	
	Cerf	Our	Cerf	Our
Linear Combination				
EPQFT	0.7601	0.7697	0.7641	0.7685
PFT'07	0.7577	0.7677	0.7610	0.7666
QDCT	0.7611	<b>0.7706</b>	0.7634	<b>0.7686</b>
DCT'11	0.7593	0.7682	0.7615	0.7673
GBVS'07	0.7223	0.7306	0.7120	0.7284
Late Combination				
EPQFT	0.7632	0.7689	0.7529	<b>0.7617</b>
PFT'07	0.7594	<b>0.7667</b>	0.7487	0.7606
QDCT	0.7660	<b>0.7667</b>	0.7445	0.7613
DCT'11	0.7624	0.7655	0.7434	0.7600
GBVS'07	0.7238	0.7089	0.6632	0.7047
Early Interaction				
EPQFT	0.6581	0.6582	0.6355	0.6660
PFT'07	0.6586	0.6586	0.6365	0.6667
QDCT	0.6588	0.6589	0.6373	0.6666
DCT'11	<b>0.6593</b>	0.6588	0.6372	<b>0.6670</b>
GBVS'07	0.6537	0.6575	0.6366	0.6632
Quaternion Face Channel				
EPQFT	0.7111	0.7115	0.7118	0.7110
QDCT	<b>0.7140</b>	0.7138	0.7145	<b>0.7142</b>
Face-only				
Faces	0.6566	<b>0.6594</b>	0.6367	<b>0.6648</b>
Saliency-only				
EPQFT			0.7223	
QDCT			0.7205	
DCT'11			0.7204	
PFT'07			0.7199	
GBVS'07			0.6350	
Further Baseline				
Zhao and Koch*, 2011			0.7561	

Table 3.6.: AUC performance of the evaluated algorithms on the Cerf/FIFA dataset [CFK09]. The ideal, *i.e.* human, AUC is 0.786. \*: We used the ideal AUC to calculate our AUC for Zhao and Koch's reported result [ZK11].





(a) Our face model based on face annotations.



(b) Our face model based on MCT face detections.

Figure 3.13.: Illustration of the average AUC in dependency of the chosen face integration method on the Cerf/FIFA dataset [CFK09]. This illustration is best viewed in color.

the best performance, which is closely followed by the late integration scheme. The integration of the face conspicuity map in the quaternion image does not perform equally well. However, it still substantially outperforms the supra-linear combination, which performs worse than each unimodal map. This could be expected, because the supra-linear combination implies a logical “and”.

There is one question that we would like to discuss further: Is the linear combination scheme significantly better than the late combination scheme? To address this question, we resort to our array of statistical tests, see Sec. 3.2.2.B. Unfortunately, the question can not be answered easily and definitely. If we look at the results that we achieve with groundtruth annotations, we see that the results are mixed with beneficial cases for both integration schemes. For example, for QDCT late integration is beneficial in combination with Cerf’s face model while linear integration is beneficial for our face model. In both example cases the statistical tests leave not much room for interpretation. The p-values for our three t-tests (*i.e.*, higher, equal, and lower mean) are close to  $(0, 0, 1)$  in the first case and  $(1, 0, 0)$  in the second case. Here, a potential cause might be the sometimes distorted groundtruth segmentation masks, *e.g.* see the two rightmost images in Fig. 3.12. However, linear integration is better in all cases for MCT face detection. Given that the performance differences appear quite substantial for MCT detections, it comes at no surprise that the statistical tests indicate that linear integration is in fact “better” for all these cases, *i.e.* if we rely on MCT face detection. In combination with the fact that the performance is better for a relatively large value range of  $w_F$ , see Fig. 3.13, we can only suggest to use the linear integration for practical applications with automatic face detections.

## 3.3 Auditory Attention

---

Two fundamental concepts are involved in human bottom-up auditory attention that form the basis for our computational attention model (*cf.* Sec. 2.1.2): First, auditory attention relies on audio data that is subject to a frequency analysis that is realized by the basilar membrane. Second, the brain’s auditory attention system relies – among other aspects – on so-called novelty detection neurons that encode deviations from the pattern of preceding stimuli. We can model the first concept with common time-frequency analysis methods (Sec. 3.3.1.A). Here, it is interesting that by doing so, we can rely on computations that can in later stages be reused by other auditory tasks such as, *e.g.*, speech recognition or sound source localization. To model the second concept, we can assign a “surprise” neuron to each frequency, following Itti and Baldi’s theory of Bayesian surprise [IB06]. Such neurons probabilistically learn and adapt to changes of each frequency’s distribution over time. To detect novel, odd, or changed signal components, we can then measure how far a newly observed sample deviates from our learned pattern (Sec. 3.3.1.B), which in principle is similar to the brain’s signal mismatch detection mechanism.

### 3.3.1. Auditory Novelty Detection

In sensory neuroscience, it has been suggested that only unexpected information is transmitted from one stage to the next stage of neural processing [RB99]. According to this theory, the sensory cortex has evolved neural mechanisms to adapt to, predict, and suppress expected statistical regularities [OF96, MMKL99, DSMS02] to focus on events that are unpredictable and appear as being novel, odd, or “surprising”. It is intuitively clear that “surprising” signals and events can only occur in undeterministic environments. This means that surprise arises from the presence of uncertainty that can be caused by, for example, intrinsic stochasticity or missing information. Interestingly, it has been shown in probability and decision theory that the Bayesian theory of probability provides the only consistent and optimal theoretical framework to model and reason about uncertainty [Jay03, Cox64]. Accordingly, Itti and Baldi [IB06] suggested a Bayesian approach to model neural responses to surprising signals, see Fig. 3.14.

In the Bayesian probability framework, probabilities correspond to subjective degrees of beliefs (see, *e.g.*, [Gil00]) in models that are updated according to Bayes’ rule as new data is observed. According to Bayesian surprise, the background information of an observer is represented in the prior probability distribution  $\{P(M)\}_{M \in \mathcal{M}}$  over the models  $M$  in a model space  $\mathcal{M}$ . Given the prior distribution of beliefs, a new data observation  $D$  is used to update the

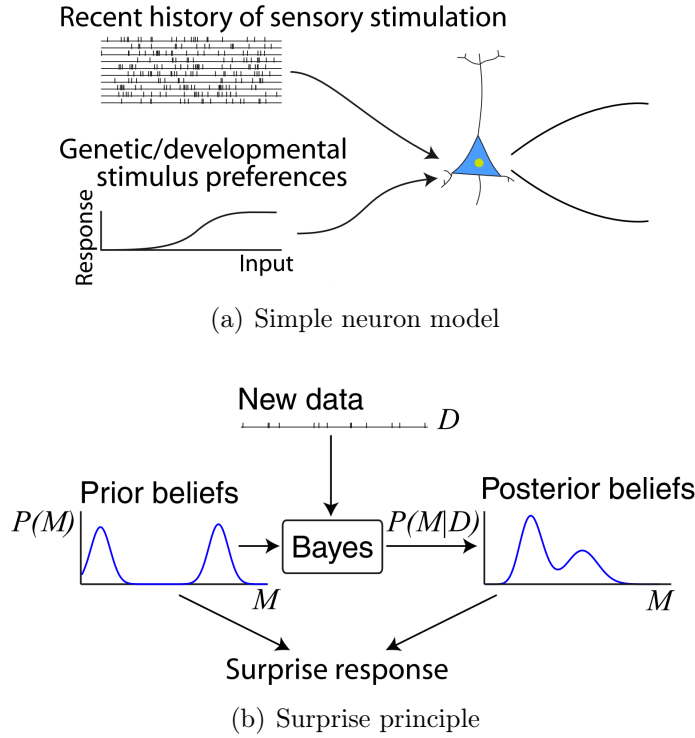


Figure 3.14.: Bayesian surprise as a probabilistic model for novelty detection on the basis of simple neurons. Images from [IB06].

prior distribution  $\{P(M)\}_{M \in \mathcal{M}}$  into the posterior distribution  $\{P(M|D)\}_{M \in \mathcal{M}}$  via Bayes' rule

$$\forall M \in \mathcal{M} : \quad P(M|D) = \frac{P(D|M)}{P(D)} P(M). \quad (3.53)$$

In this framework, the new data observation  $D$  carries no surprise, if it leaves the observer beliefs unaffected, *i.e.* the posterior is identical to the prior.  $D$  is surprising, if the posterior distribution after observing  $D$  significantly differs from its prior distribution. To formalize this, Itti and Baldi propose to use the Kullback-Leibler divergence (KLD) to measure the distance  $D_{\text{KL}}$  between the prior and posterior distribution

$$S(D, \mathcal{M}) = D_{\text{KL}}(P(M|D), P(M)) = \int_{\mathcal{M}} P(M|D) \log \frac{P(M|D)}{P(M)} dM. \quad (3.54)$$

The distance  $S(D, \mathcal{M})$  between the prior and posterior now quantifies how surprising observation  $M$  is.

Since the posterior distribution can be updated immediately after observing new data, it is clear that surprise is an attention model that is particularly suited to detect surprising events online and without delay in sensory streams such as, most importantly, audio and video streams.

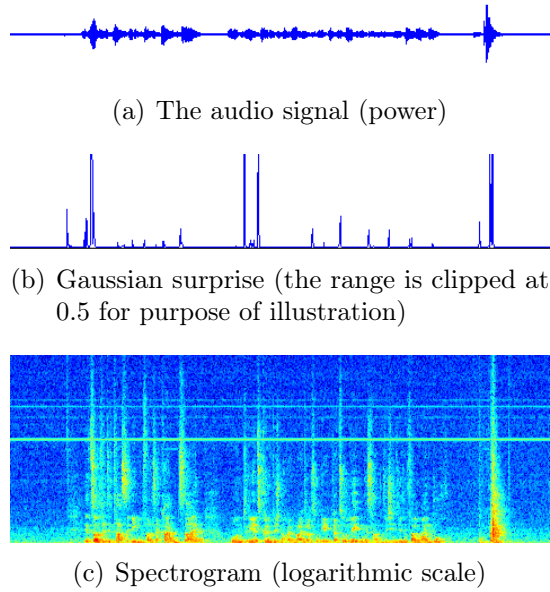


Figure 3.15.: An approximately ten second exemplary audio sequence in which a person speaks and places a solid object on a table at the end of the sequence. The measured audio signal (a), the resulting Gaussian auditory surprise (b), and the spectrogram (c).

### A. Time-Frequency Analysis and Bayesian Framework

We can use the short-time Fourier transform (STFT), short-time cosine transform (STCT), or the modified discrete cosine transform (MDCT) to calculate the spectrogram  $G(t, \omega) = |F(t, \omega)|^2$  of the windowed audio signal  $a(t)$ , where  $t$  and  $\omega$  denote the discrete time and frequency, respectively. Accordingly, at each time step  $t$ , the newly observed frequency data  $G(t, \omega)$  is used to update the prior probability distribution

$$\forall \omega \in \Omega : \quad P_{\text{prior}}^\omega = P(G(\cdot, \omega) | G(t-1, \omega), \dots, G(t-N, \omega)) \quad (3.55)$$

of each frequency and obtain the posterior distribution

$$\forall \omega \in \Omega : \quad P_{\text{post}}^\omega = P(G(\cdot, \omega) | G(t, \omega), G(t-1, \omega), \dots, G(t-N, \omega)) , \quad (3.56)$$

where  $N \in \{1, \dots, \infty\}$  allows additional control of the time behavior by limiting the history to  $N \neq \infty$  elements, if wanted. The history allows us to limit the influence of samples over time and consequently “forget” data, which is essential for the time behavior of the Gaussian surprise model.

## B. Auditory Surprise

**Gaussian model** Using the Gaussian distributions as model, we can calculate the auditory surprise  $S_A(t, \omega)$  for each frequency

$$S_A(t, \omega) = D_{\text{KL}}(P_{\text{post}}^\omega || P_{\text{prior}}^\omega) = \int P_{\text{post}}^\omega \log \frac{P_{\text{post}}^\omega}{P_{\text{prior}}^\omega} dg \quad (3.57)$$

$$= \frac{1}{2} \left[ \log \frac{|\Sigma_{\text{prior}}^\omega|}{|\Sigma_{\text{post}}^\omega|} + \text{Tr} \left[ \Sigma_{\text{prior}}^{\omega^{-1}} \Sigma_{\text{post}}^\omega \right] - I_D + \right. \\ \left. (\mu_{\text{post}}^\omega - \mu_{\text{prior}}^\omega)^T \Sigma_{\text{prior}}^{\omega^{-1}} (\mu_{\text{post}}^\omega - \mu_{\text{prior}}^\omega) \right] , \quad (3.58)$$

where  $\mu$  and  $\Sigma$  is the mean and variance, respectively, of the data in the considered time window.  $D_{\text{KL}}$  is the KLD and Eqn. 3.58 results from the closed form of  $D_{\text{KL}}$  for Gaussian distributions [HO07].

**Gamma model** The Gaussian model is extremely run-time efficient and in general performs well according to our experience. But, it has one main disadvantage: All elements inside the history window have equal weight. Instead of equal weights, it would be desirable that the weight and thus the influence of each observation slowly decreases over time to realize a “smooth” forgetting mechanism. Similar to the approach by Itti and Baldi for detecting surprising events in video streams [IB05], we can use the Gamma distribution as an alternative to the Gaussian distribution

$$P(x) = \gamma(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad (3.59)$$

with  $x \geq 0$ ,  $\alpha, \beta > 0$ , and Gamma function  $\Gamma$  to calculate the surprise.

Given a new observation  $G(t, \omega)$  and prior density  $P_{\text{prior}}^\omega = \gamma(\cdot; \alpha, \beta)$ , we calculate the posterior  $P_{\text{post}}^\omega = \gamma(\cdot; \alpha', \beta')$  using Bayes' rule

$$\alpha' = \alpha + G(t, \omega) \quad (3.60)$$

$$\beta' = \beta + 1 . \quad (3.61)$$

However, using this update rule would lead to an unbounded growth of the values over time. To avoid this behavior and reduce the relative importance of older observations, we integrate a decay factor  $0 < \zeta < 1$

$$\alpha' = \zeta \alpha + G(t, \omega) \quad (3.62)$$

$$\beta' = \zeta \beta + 1 . \quad (3.63)$$

This formulation preserves the prior's mean  $\mu = \frac{\alpha}{\beta} = \frac{\zeta \alpha}{\zeta \beta}$  but increases its variance, which however represents a relaxation of belief in the prior's precision after observing  $G(t, \omega)$ .

Now, we can calculate the surprise as follows

$$S_A(t, \omega) = D_{\text{KL}}(P_{\text{post}}^\omega || P_{\text{prior}}^\omega) = \int P_{\text{post}}^\omega \log \frac{P_{\text{post}}^\omega}{P_{\text{prior}}^\omega} dg \quad (3.64)$$

$$= \alpha' \log \frac{\beta}{\beta'} + \log \frac{\Gamma(\alpha')}{\Gamma(\alpha)} \quad (3.65)$$

$$+ \beta' \frac{\alpha}{\beta} + (\alpha - \alpha') \psi(\alpha) , \quad (3.66)$$

where  $\psi$  is the Digamma function. Unfortunately, the Gamma and Digamma functions  $\Gamma$  and  $\psi$ , respectively, do not have a closed form. But, there exist sufficiently accurate approximations (see, *e.g.*, [Ber76]), which however make the calculation slightly more complex than in the case of the Gaussian model.

### C. Across Frequency Combination

Finally, we calculate the auditory saliency  $S_A(t)$  as the mean over all frequencies

$$S_A(t) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} S_A(t, \omega) \quad . \quad (3.67)$$

We do not use an alternatively possible joint (*e.g.*, Dirichlet) model for the surprise calculation due to its computational complexity. Such a joint model would require the calculation of a general covariance matrix with every update. Given the typically large number of analyzed frequencies (*i.e.*,  $> 10000$ ), the associated computational complexity makes real-time processing impractical if not impossible.

### 3.3.2. Evaluation

In contrast to, for example, recording eye fixations as a measure of visual saliency (see Sec. 2.1.1), we can not simply observe and record humans to provide a measure of auditory saliency. Consequently, we follow a pragmatic, application-oriented evaluation approach that enables us to use existing acoustic event detection and classification datasets.

#### A. Evaluation Measure

Salient acoustic event detection has to suppress “uninteresting” audio data while highlighting potentially relevant and thus salient acoustic events. However, in contrast to classical acoustic event detection and classification, this consideration leads to a different evaluation methodology in which: First, a high recall is necessary, because we have to detect all prominent events so that they can be analyzed by later processing stages. Second, a high precision is of secondary interest, because we can tolerate false positives as long as we still filter the

signal in such a way that we achieve a net benefit when taking into account subsequent processing stages. We can realize this evaluation idea by using the well-established  $F_\beta$  score

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \quad (3.68)$$

$$F_\beta = \frac{(1 + \beta^2) \cdot \text{true pos.}}{(1 + \beta^2) \cdot \text{true pos.} + \beta^2 \cdot \text{false neg.} + \text{false pos.}} \quad (3.69)$$

as evaluation measure, where  $\beta$  “measures the effectiveness of retrieval with respect to a user who attaches  $\beta$  times as much importance to recall as precision” [vR79]. It is noteworthy that  $F_\beta$  is also the most commonly used evaluation measure for salient object detection in image processing applications (see Sec. 4.2.3.B).

## B. Evaluation Data

We use the CLEAR2007 acoustic event detection dataset for evaluation [CLE] (*cf.* [TMZ<sup>+</sup>07]). The dataset contains recordings of meetings in a smart room and its collection was supported by the European integrated project Computers in the Human Interaction Loop (CHIL) and the US National Institute of Standards and Technology (NIST). For each recording a human analyst marked and classified all occurring acoustic events that were remarkable enough to “pop-out” from the acoustic scene’s background noise. In total, 14 different classes of acoustic events were classified and flagged, including sudden “laughter”, “door knocks”, “phone ringing” and “key jingling”. Here, it is interesting to note that not all events could be identified by the human analyst, in which case they were labeled with “unknown”.

## C. Evaluation Parameters

For the time-frequency analysis, we set the window size to contain 1 second of audio data, which has a resolution of 22 kHz, and use 50 % overlap. We also experimented with applying various window functions (*e.g.*, Blackman, Gauss), but the resulting performance difference is mostly negligible, if the window functions’ parameters are well defined. We evaluated the performance for the modified discrete cosine transform (MDCT), short-time cosine transform (STCT), and short-time Fourier transform (STFT) to determine whether or not the Gamma distribution is beneficial for all of these transformations. We do this, because one aim is to produce as little run-time overhead as possible, which requires us to ideally rely on the transformation that is used for the subsequent processing steps such as, *e.g.*, sound source localization, event recognition, and/or speech recognition. We optimized the history size and forgetting parameter for the Gaussian and Gamma model, respectively, and report the results for the best choice



Algorithm	$F_1$	$F_2$	$F_4$
STFT + Gamma	0.7668	0.8924	0.9665
STCT + Gamma	0.7658	0.8916	0.9655
MDCT + Gamma	0.7644	0.8894	0.9647
STFT + Gaussian	0.7604	0.8832	0.9531
STCT + Gaussian	0.7612	0.8813	0.9529
MDCT + Gaussian	0.7613	0.8805	0.9538

Table 3.7.: Performance of the evaluated auditory surprise algorithms on CLEAR 2007 acoustic event detection data. The  $F_2$  and  $F_4$  scores are our main evaluation measure, because for our application a high recall is much more important than a high precision (we provide the  $F_1$  score mainly to serve as a reference). We can see that surprise is able to reliably detect arbitrary, interesting acoustic events.

## D. Results

As can be seen in Tab. 3.7, quantified using the  $F_1$ ,  $F_2$ , and  $F_4$  score, auditory surprise is able to efficiently detect arbitrary salient acoustic events. Although in general an  $F_1$  score of roughly 0.77 is far from perfect for precise event detection, we can see from the substantially higher  $F_2$  and  $F_4$  scores that we can efficiently detect most (salient) acoustic events, if we tolerate a certain amount of false positives. This nicely fulfills the target requirements for our application domain and comes at a low computational complexity, since Gaussian surprise allows us to process one minute of audio data in roughly 1.5 seconds. This makes it possible to process the incoming audio data stream in real-time, detect salient events online, and signal occurring salient events to subsequent stages with a minimum delay. Furthermore, since we calculate the surprise value for all frequencies that we subsequently combine, we can also determine which frequencies trigger the detection. An information that can be passed to subsequent stages to focus the processing on these frequencies. We can see in Tab. 3.7 that the Gamma distribution leads to a better performance compared to the Gauss distribution, independently of the preceding time-frequency transformation. This, however, comes at the cost of greater computational complexity.

## 3.4 Saliency-based Audio-Visual Exploration

In the previous sections, we have investigated saliency models to determine what attracts the auditory or visual attention (Sec. 3.2 and 3.3, respectively). However, to realize an attention system that sequentially focuses on salient regions in the scene – similar to human saccades –, we need to define a crossmodal representation, extract auditory and visually salient regions, sequentially shift the focus of attention, and keep track of attended objects to implement inhibition of return (*cf.* Sec. 2.1.1 and 2.1.3).

Attention forms a selective gating mechanisms that decides what will be processed by later stages. This process is often describes as a “spotlight” that enhances the processing in the attended [TG80, Pos80], *i.e.* “illuminated”, region. In a similar metaphor, attention can act like a “zoom lense” [ESJ86, SW87], because the size of the attended region can be adjusted depending on the task. However, most models do not consider the shape and extent of the attended object, which is essential to determine the area that has to be attended. And, experimental evidence suggests that attention can be tied to objects, object parts, and/or groups of objects [Dun84, EDR94, RLS98]. But, how can we attend to objects before we recognize them [WK06]?

One model that addresses this question has been introduced by Rensink [Ren00a, Ren00b]. Rensink describes “proto-objects” as volatile units of visual information that can be bound into a coherent and stable object when accessed by focused attention [WK06]. A related concept that we have already addressed earlier (see Sec. 2.1.1) are Kahneman and Treisman’s “object files” [KT00, KTG92]. The main difference between proto-objects and object files is the role of location in space. In Kahneman and Treisman’s object file model, the spatial location is just another property of an object, *i.e.* it is just another entry in the object’s file, see Fig. 2.1. In contrast, in Rensink’s proto-object model and coherence theory (see [SY06]), the spatial location serves an index that binds together various low-level features into proto-objects across space and time [Ren00a, WK06].

### 3.4.1. Gaussian Proto-Object Model

We rely on a probabilistic model to represent salient auditory, visual, and audio-visual proto-objects. In our model, every proto-object  $o \in \mathcal{O}$  is represented by a parametric Gaussian weight function

$$f_o^G(x) = \frac{s_o}{\sqrt{(2\pi)^3 \det(\Sigma_o)}} \exp\left(-\frac{1}{2}(x - \mu_o)^T \Sigma_o^{-1} (x - \mu_o)\right) \quad (3.70)$$

with  $x \in \mathbb{R}^3$ . Here, the 3D mode  $\mu_o \in \mathbb{R}^3$  represents the likely spatial center of the proto-object, the variance  $\Sigma_o$  reflects the spatial extent that means – more generally – the spatial area that likely contains the actual object, and  $s_o$  is the

proto-object’s saliency. In accordance to our Gaussian model, we can represent every proto-object  $o$  as a 3-tuple  $h_o$

$$h_o = (s_o, \mu_o, \Sigma_o) \in \mathcal{H} \quad . \quad (3.71)$$

### 3.4.2. Auditory Proto-Objects

In Sec. 3.3, we have discussed how we determine how salient a sound signal is at time  $t$ . However, in order to form a proto-object, we have to determine the coarse spatial area that likely contains the salient signal’s sound source.

#### A. Localization

We rely on the well-known steered response power (SRP) with phase transform (PHAT) sound source localization [MMS<sup>+</sup>09, DSB01]. The SRP-PHAT algorithm uses the inter-microphone time difference of arrival (TDOA) of sound signals, which is caused by the different distances the sound has to travel to reach each microphone, to estimate the location of the sound source. To this end, the following inter-microphone signal correlation function is used to determine TDOAs  $\tau$  of prominent signals at time  $t$

$$R_{ij}(t, \tau) = \int_{-\infty}^{\infty} \psi_{ij}^{\text{PHAT}}(t, \omega) F'_i(t, \omega) F'_j(t, \omega)^* e^{j\omega\tau} d\omega \quad , \quad (3.72)$$

where  $F'_i$  and  $F'_j$  are the STFT transformed signals of the audio signal at microphone  $i$  and  $j$ , respectively. The PHAT specific weighting function

$$\psi_{ij}^{\text{PHAT}}(t, \omega) = |F'_i(t, \omega) F'_j(t, \omega)^*|^{-1} \quad (3.73)$$

can be regarded as a whitening filter and is supposed to decrease the influence of noise and reverberations. Subsequently, we can use the estimated TDOAs to calculate the corresponding spatial positions in the environment.

#### B. Parametrization

Since the sound source localization is a process that exhibits a considerable amount of noise, we perform spatio-temporal clustering to remove outliers and improve the accuracy of the localization. Accordingly, we can use the mean of each cluster as the proto-object’s location estimate  $\mu_o$  and calculate the corresponding co-variance matrix  $\Sigma_o$ . Consequently, each detected acoustically salient proto-object  $o$  is described by its saliency  $s_o$ , the estimated location  $\mu_o$ , and the co-variance matrix  $\Sigma_o$  that encodes the spatial uncertainty.

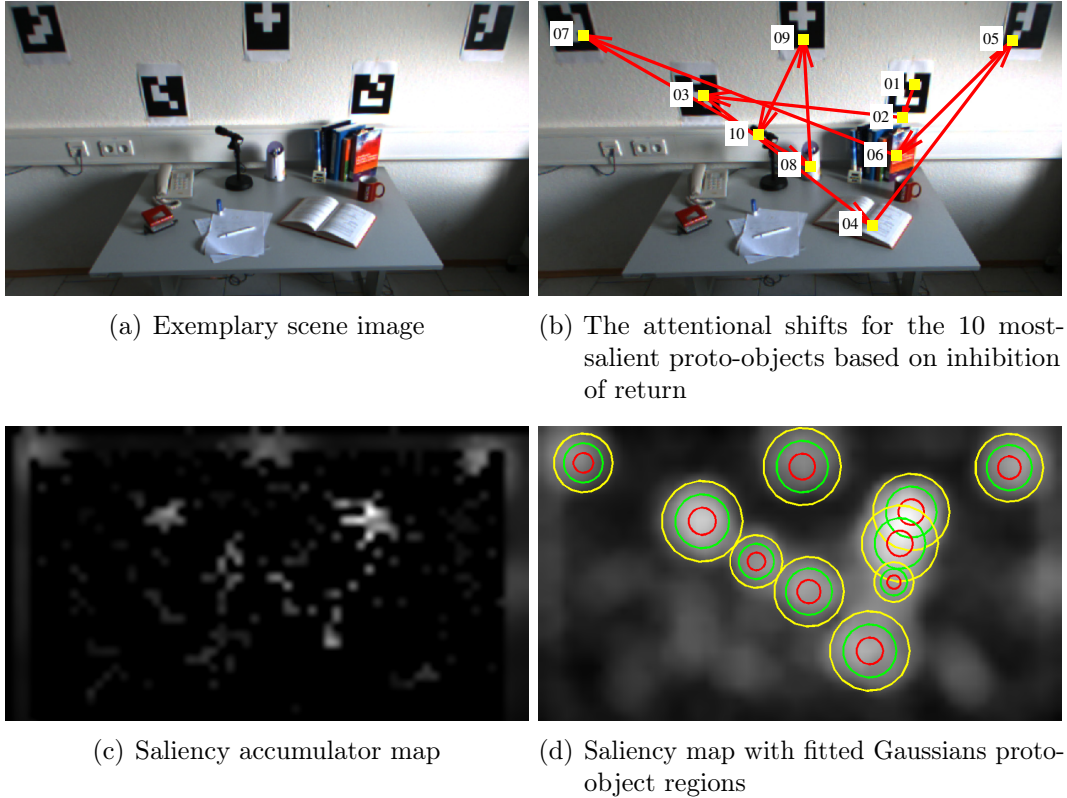


Figure 3.16.: An exemplary scene image (a), the saliency map with fitted Gaussian proto-object regions (d), the resulting accumulator (c), and the first 10 salient proto-object regions that are selected by the location-based inhibition of return (b). The estimated Gaussian weight descriptors are depicted as overlay on the saliency map (illustrated as circles with center  $\mu_o$  and radii  $r \in \{\sigma_o, 2\sigma_o, 3\sigma_o\}$  in {red, green, yellow}, respectively). Please note that the value range of the saliency map and the accumulator is attenuated for purpose of illustration. This illustration is best viewed in color.

### 3.4.3. Visual Proto-Objects

In Sec. 3.2, we have described how we calculate the visual saliency of an image. However, to represent this information in our proto-object model, we have to determine 2D proto-object regions in the saliency map. Then, we can use depth information that can be provided by, for example, stereo vision to form a 3D proto-object representation.

#### A. Proto-Object Regions

We analyze the saliency map’s isophote curvature (see [LHR05]) to estimate the proto-object regions, see Fig. 3.16. Here, isophotes are closed curves of

constant saliency within the saliency map. Assuming a roughly (semi-)circular structure of salient peaks, we can then determine the center of each salient peak as well as its corresponding pixels that we define as the pixels whose gradients point toward the peak center. This way, we are able to efficiently extract salient regions, even in the presence of highly varying spatial extent and value range, partially overlapping peaks and noise. To this end, we analyze the local isophote curvature  $\kappa$  of the visual saliency map  $S_V$

$$\kappa = -\frac{S_{cc}}{S_g} = -\frac{S_y^2 S_{xx} - 2S_x S_y S_{xy} + S_x^2 S_{yy}}{(S_x^2 + S_y^2)^{3/2}}, \quad (3.74)$$

where  $S_{cc}$  is the second derivative in the direction perpendicular to the gradient and  $S_g$  is the derivative in gradient direction<sup>3</sup>. Accordingly,  $S_x$ ,  $S_y$  and  $S_{xx}$ ,  $S_{xy}$ ,  $S_{yy}$  are the first and second derivatives in x and y direction, respectively. Exploiting that the local curvature is reciprocal to the (hypothetical) radius  $r$  of the circle that generated the saliency isoline of each pixel, *i.e.*

$$r(x, y) = \frac{1}{\kappa(x, y)}, \quad (3.75)$$

we can estimate the location of each peak's center. Therefore, we calculate the displacement vectors  $(D_x, D_y)$  with

$$D_x = \frac{S_x (S_x^2 + S_y^2)}{S_{cc}} \quad \text{and} \quad D_y = \frac{S_y (S_x^2 + S_y^2)}{S_{cc}} \quad (3.76)$$

and the resulting hypothetical peak centers  $(C_x, C_y)$  with

$$C_x = P_x - D_x \quad \text{and} \quad C_y = P_y - D_y, \quad (3.77)$$

where the matrices  $P_x$  and  $P_y$  represent the pixel abscissae and ordinates, *i.e.* the pixel  $(x, y)$  coordinates, respectively.

Thus, we can calculate a saliency accumulator map  $A_S$  in which each pixel votes for its corresponding center. The most salient regions, *i.e.* corresponding to the extents of the proto-objects in the image (see, *e.g.*, [HZ07]), in the saliency map can then be determined by selecting the pixels of the accumulator cells with the highest voting score, see Fig. 3.16(c). By choosing different weighting schemes for the voting, we are able to implement divers methods for assessing the saliency of each region. In the following, we use the saliency as weight and normalize each accumulator cell  $A_S(m, n)$  by division by the number of pixels that voted for the pixel, *i.e.*

$$A_S(m, n) = \frac{\sum_x \sum_y 1_m(C_x(x, y)) 1_n(C_y(x, y)) S_V(x, y)}{\sum_x \sum_y 1_m(C_x(x, y)) 1_n(C_y(x, y))}, \quad (3.78)$$

<sup>3</sup>Please note that all operations in Eq. 3.74 and 3.76 operate element-wise. We chose this simplified notation for its compactness and readability.

where  $1_x(y)$  is the indicator function with  $1_x(y) = 1$  iff  $x = y$  and  $1_x(y) = 0$  otherwise. However, due to noise and quantization effects, we additionally select pixels that voted for accumulator cells within a certain radius  $r$ , i.e.

$$A'_S(m, n) = \frac{\sum_x \sum_y 1_{m,n}^r(C_x(x, y), C_y(x, y)) S_V(x, y)}{\sum_x \sum_y 1_{m,n}^r(C_x(x, y), C_y(x, y))} \text{ with} \quad (3.79)$$

$$1_{m,n}^r(x, y) = \begin{cases} 1 & \text{if } \sqrt{(m-x)^2 + (n-y)^2} \leq r \\ 0 & \text{otherwise.} \end{cases} \quad (3.80)$$

Unfortunately, the initially selected pixels of our proto-object regions are contaminated with outliers caused by noise. Thus, we perform convex peeling (*cf.* [HA04]), a type 1, unsupervised clustering-based outlier detector to remove scattered outliers and eliminate regions whose percentage of detected outliers is too high.

To extract all salient proto-object regions that attract the focus of attention, we apply a location-based inhibition of return mechanism on the saliency map (see, *e.g.*, [RLB<sup>+</sup>08, IKN98]; [SRF10, SF10a]). To this end, we use the accumulator to select the most salient proto-object region and inhibit all pixels within the estimated outline by setting their saliency to zero. This process is repeated until no further prominent salient peaks are present in the map.

## B. Parametrization

For each extracted 2D salient proto-object region  $o \in \mathcal{O}_{2D}(S_V(t))$  within the visual saliency map  $S_V$  at time  $t$ , we derive a parametric description by fitting a Gaussian weight function  $f_o$ . We assume that the Gaussian weight function encodes two distinct aspects of information: the saliency  $s_o$  as well as the (uncertain) spatial location and extent of the object  $\mu_o$  and  $\Sigma_o$ , respectively. Consequently, we decompose the Gaussian weight function:

$$f_o^G(x) = \frac{s_o}{\sqrt{(2\pi)^D \det(\Sigma_o)}} \exp\left(-\frac{1}{2}(x - \mu_o)^T \Sigma_o^{-1} (x - \mu_o)\right) \quad (3.81)$$

with  $D = 2$ . Exploiting a stereo setup or other RGB-D sensors, we can estimate the depth and project the 2D model into 3D. This way, we obtain a 3D model for each visually salient proto-object region that follows the representation of the detected auditory salient events, see Sec. 3.4.2.B. However, we have to make assumptions about the shape, because the spatial extent of the object in direction of the optical axis can not be observed. Thus, we simplify the model and assume a spherical model in 3D and, accordingly, a circular outline in 2D, *i.e.*  $\Sigma_o = I_D \sigma_o$  with the unit matrix  $I_D$ .

### 3.4.4. Audio-Visual Fusion and Inhibition

#### A. Saliency Fusion

After the detection and parametrization of salient auditory and visual signals, we have a set of auditory  $\mathcal{H}_A$  and visual  $\mathcal{H}_V$  proto-objects represented in a Gaussian notation at each point in time  $t$

$$\{h_1, \dots, h_N\} = \mathcal{H}^t = \mathcal{H}_A^t \cup \mathcal{H}_V^t, \quad (3.82)$$

where each proto-object  $o_i \in \mathcal{O}$  is represented by a 3-tuple  $h_i$  consisting of its saliency  $s_{o_i}$ , spatial mean  $\mu_{o_i}$ , and spatial variance  $\Sigma_{o_i}$ , see Eq. 3.71. To reduce the influence of noise as well as to enable multimodal saliency fusion, we perform a cross-modal spatio-temporal mean shift clustering [CM02] of the auditory and visual Gaussian representatives. Accordingly, we obtain a set of audio-visual clusters  $C^t \in \mathcal{P}(\mathcal{H}^t)$ , each of which can be interpreted as a (saliency-weighted) Gaussian mixture model. Therefore, we interpret each cluster  $c \in C^t$  as a saliency-weighted Gaussian mixture model, that consists of auditory and/or visual proto-objects. This allows us to split each cluster  $c$  again into an auditory ( $c_A = c \cap \mathcal{H}_A^t$ ) and/or visual ( $c_V = c \cap \mathcal{H}_V^t$ ) sub-cluster and estimate the saliency for each modality separately. Subsequently, we consider a linear combination to integrate the audio-visual saliency

$$s_o = \frac{1}{2} \left( \sum_{o_k \in c_A} w_{o_k}^A f_{o_k}^G(\mu_o) + \sum_{o_l \in c_V} w_{o_l}^V f_{o_l}^G(\mu_o) \right), \quad (3.83)$$

using the modality specific weights  $w_{o_j}^A$  and  $w_{o_j}^V$  (analogous to Eq. 3.84). Consequently, we use the spatial mean of every proto-object within the cluster to estimate the position

$$\mu_o = \mathbb{E}[c] = \sum_{o_j \in c} \mu_{o_j} w_{o_j} \quad \text{with } w_{o_j} = \frac{s_{o_j}}{\sum_{o_i \in c} s_{o_i}}. \quad (3.84)$$

Finally, we determine the spatial variance of the cluster  $\Sigma_o$  by iteratively fusing the variance of the proto-objects

$$V_j = V_{j-1} - V_{j-1} (V_{j-1} + \Sigma_{o_j})^{-1} V_{j-1}, \quad \forall j=2, \dots, H \quad (3.85)$$

with  $V_1 = \Sigma_{o_1}$ ,  $\Sigma_o = V_H$ , and  $H = |c|$ . Accordingly, we are able to build a new audio-visual proto-object  $h_o = (s_o, \mu_o, \Sigma_o)$  with integrated saliency  $s_o$ , spatial mean  $\mu_o$  as well as spatial variance  $\Sigma_o$ .

We use a linear combination for crossmodal integration, because it has been shown to be a good model for human overt attention and is optimal according to information theoretic criteria [OLK07], see Sec. 2.1.3. However, other combination schemes (see, *e.g.*, [OLK07]) can be realized easily given the model and algorithmic framework.

## B. Object-based Inhibition of Return

An important additional feature of our spatio-temporal fusion in combination with the employed object-based world model, see [KBS<sup>+</sup>10], is the object-centric representation of salient regions. This allows us to use the euclidean distance metric to relate the current proto-objects with previous proto-object detections at previous time steps as well as already analyzed objects that are stored in a world model. This way, we can decide whether to create and attend a new proto-object or update the information of an already existing entity.

To iteratively attend and analyze the objects present in the scene, we use the detected salient proto-objects to realize an object-based inhibition of return mechanism. Therefore, at each decision cycle, the most salient proto-object cluster that is not related with an already attended and analyzed proto-object gains the overt focus of attention.

## C. Knowledge-driven Proto-object Analysis

After the sensors have been aligned with respect to the proto-object in the current overt focus of attention, the foveal cameras (see Fig. 3.17) are used to inspect the object. Therefore, we extend the multimodal knowledge-driven scene analysis and object-based world modeling system as presented by Machmer *et al.* [MSKK10] and Kühn *et al.* [KBS<sup>+</sup>10], to comply with our iterative, saliency-driven focus of attention and exploration mechanism. Most importantly, we replaced the detection and instantiation phase by regarding proto-objects as primitive candidates for world model entities. The attended proto-object region is instantiated as entity and subsequently hierarchically specialized and refined in a knowledge-driven model (see [MSKK10, KBS<sup>+</sup>10]). The analysis of each proto-object is finished, if no further refinement is possible, which marks the end of the decision cycle and initiates the next shift of attention. Within this framework, every entity is tracked which is an important feature of object-based inhibition of return.

### 3.4.5. Evaluation

#### A. Hardware and Software Setup

The sensor setup that was used for the evaluation of the presented system is shown in Fig. 3.17. The wide angle and foveal cameras have a focal length of 6 mm and 3.5 mm, respectively. The stereo baseline separation between each camera pair is 90 mm. The camera sensors provide a resolution of  $640 \times 480$  px at a frame rate of 30 Hz. In the evaluation only the front and side omnidirectional microphones are used (see Fig. 3.17). The distance between the side microphones is approximately 190 mm and the vertical distance between the front microphones is approximately 55 mm. The pan-tilt unit provides an angular resolution of



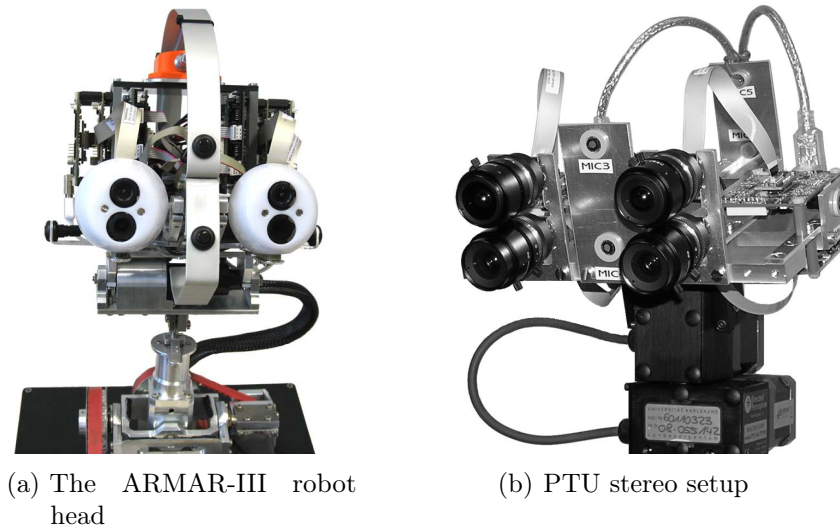


Figure 3.17.: The ARMAR-III humanoid robot head (a) and our pan-tilt-unit (PTU) stereo setup (b) provide 7 and 2 degrees of freedom, respectively. Both setups perceive their environment with 6 omnidirectional microphones (1 left, 1 right, 2 front, 2 rear) and 2 stereo camera pairs (coarse and fine view, respectively).

0.013° and is mounted on a tripod in such a way that the cameras are roughly on eye height of an averagely tall human to reflect a humanoid view of the scene.

The audio data is processed at a sampling rate of 48 kHz. A Blackman window with a size of 512 samples and 50% overlap is used to calculate the STFT for the Gaussian auditory surprise, which uses a history size of  $N = 128$ . In the following, the STFT  $F'$  of the sound source localization uses a lower temporal-resolution than the STFT  $F$  of the salient event detection. This is due to the fact that we require real-time performance and, on the one hand, want to detect short-timed salient events while, on the other hand, require sufficiently large temporal windows for robust correlations. Therefore, the window length of the localization is a multiple of the salient event detections' window length.



Figure 3.18.: Exemplary image of an object in the coarse and fine view, respectively.

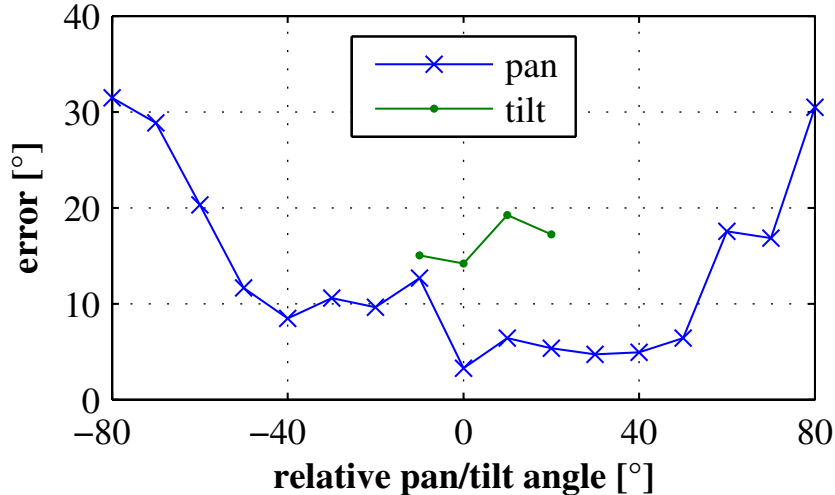


Figure 3.19.: Mean sound source localization error (in  $^{\circ}$ ) depending on the pan-tilt-orientation of the sensor setup.

Accordingly, we aggregate the auditory saliency of all detection windows that are located within the localization window. We use the maximum as aggregation function, because we want to react on short-timed salient events, instead of suppressing them.

## B. Evaluation Procedure and Measure

First of all, to demonstrate that overt attention is beneficial and justifies the required resources, we assess the impact of active sensor alignment on the perception quality (Sec. 3.4.5.C). While the improvement of the image data quality of objects in the focused foveal view compared to the coarse view is easily understandable (see Fig. 3.18), the impact on the acoustic perception depends on several factors, most importantly the sensor setup. Consequently, as reference we evaluate the acoustic localization error with respect to the pan-tilt orientation of our sensor setup relative to sound sources, *e.g.* household devices and speaking persons. For this purpose, the sound sources were placed at fixed locations and the localization was performed with pan-tilt orientations of  $\{-80^{\circ}, \dots, 80^{\circ}\} \times \{-30^{\circ}, \dots, 0^{\circ}\}$  in  $10^{\circ}$  steps (see Fig. 3.19). We only consider the angular error, because in our experience the camera-object distance error is too dependent on the algorithm parameters, implementation, and sampling rate.

We perform a couple of experiments to evaluate the behavior of the proposed system, because a quantitative, comparative method to evaluate the performance of an overt attention system does not exist (see [SS07, BKMG10]). In order to obtain a reliable impression of the performance of our system, we repeated every experiment multiple times with varying environmental conditions such as, *e.g.*, lighting, number of objects, distracting clutter, and timing of events.

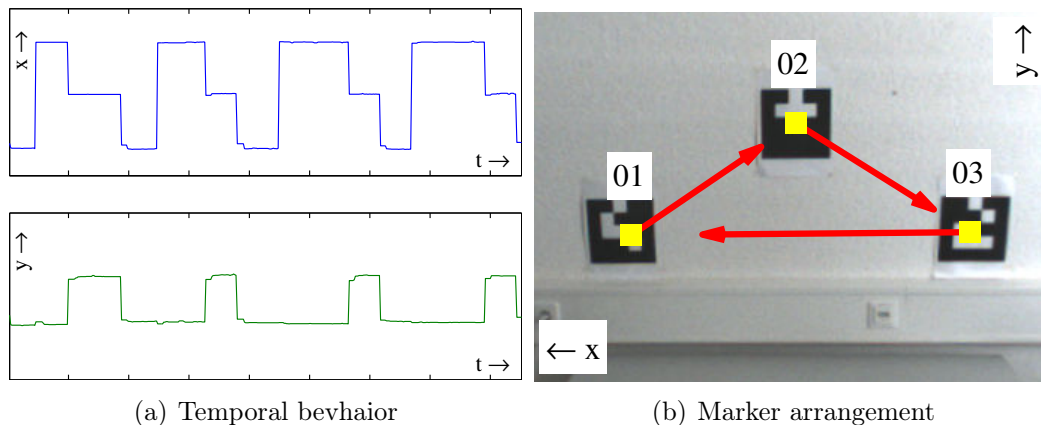


Figure 3.20.: A short temporal section of the attended x-y position (a) in the cyclic focus of attention shift experiment (see [RLB<sup>+</sup>08]). The positions correspond to the calibration marker locations (b) that lie on the same x-z plane.

Inspired by the evaluation procedures by Ruesch *et al.* [RLB<sup>+</sup>08] and Begum *et al.* [BKMG10], we investigate and discuss the performance of saliency-driven visual and multimodal scene exploration.

### C. Results and Discussion

**Audio-visual perception** As can be seen in the error curve depicted in Fig. 3.19, the angular localization error is minimal if the head faces the target object directly. This can be explained by the hardware setup in which the microphones are nearly arranged on a meridional plane. Interestingly, the curve shows a non-monotonic error progression, which is mainly caused by the hardware that interferes with the acoustic characteristic and perception, *e.g.* the cameras heavily influence the frontal microphones (see Fig. 3.17). Additionally, in Fig. 3.18 we show an example of the coarse and fine, *i.e.* foveal, view of a focused object to illustrate the improved visual perception, *i.e.* increased level of detail.

**Visual exploration I – FoA shift** In style of the experimental evaluation by Ruesch *et al.* [RLB<sup>+</sup>08, Sec. V–B], we mounted three salient calibration markers on the walls of an office environment and removed other distracting stimuli (see Fig. 3.20). In this experiment, we benefit from an object-specific lifetime that can be assigned to analyzed objects in our world model. Each object-specific lifetime is continuously evaluated and updated by, *e.g.*, taking the visibility into account. Thus, if an object has expired and is perceived as salient, it can regain the focus of attention. Driven by the implemented inhibition of return mechanism, the three salient marks are explored by shifting the overt

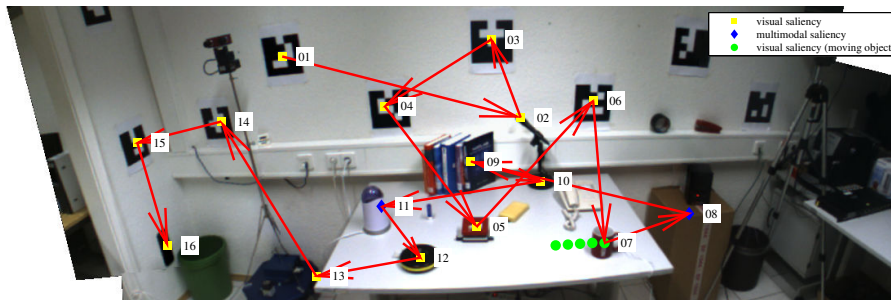


Figure 3.21.: An example of multimodal scene exploration: The focus of attention is shifted according to the numbers in the stitched image of the scene (only the first 15 shifts are shown). The yellow squares mark objects that attracted the focus solely due to their visual saliency whereas the blue squares (at 08 and 11) mark audio-visually caused shifts of attention. Furthermore, the green dotted lines (at 07) roughly indicate the trajectory of the moved object.

attention from one mark to the next most salient mark that is not inhibited. As expected, the achieved behavior corresponds to the cyclic behavior as described by Ruesch *et al.* [RLB<sup>+</sup>08]. Each attended mark is focused by controlling the pan-tilt-servos. The resulting trajectory is illustrated in Fig. 3.20.

**Visual exploration II – Object-based IoR** Once an object has been analyzed, it is tracked and inhibited – as long as the object has not been marked for re-focusing by higher-level processes – from gaining the overt focus of attention. To test the object-based inhibition of return mechanism, we perform experiments with moving objects in the scene. For this purpose, we place movable objects in the scene, start the exploration, and move objects after they have been analyzed. As expected, smoothly moving objects do not attract the focus, although they are moved to locations that have not been salient before. Naturally, this behavior even remains when motion is integrated as an additional saliency cue. Interestingly, objects that abruptly change their expected motion pattern attract the focus of attention again, because the tracking of the entity in the object-based world modeling system fails due to the unexpected movement (see Sec. 3.4.4.C). Although this could be seen as a technical deficit, this behavior is desired for an attention-based system and can be biologically motivated (*cf.* [HKM<sup>+</sup>09]).

**Multimodal exploration I – FoA shift** Following the experimental procedure of Ruesch *et al.* [RLB<sup>+</sup>08, Sec. V–C], we examine the behavior in scenes with acoustic stimuli. Therefore, we extend the scenario of the previous experiment (Sec. 3.4.5.C) and add a single visible sound source, *e.g.* a blender or a talking person. Our system explores the environment based on visual saliency

until the acoustic stimulus begins and the sound source directly gains the focus of attention.

**Multimodal exploration II – Scene** Finally, we unite the previously isolated experiments and assess the performance on more complex scenes with several objects, object motion, and auditory stimuli (please see Fig. 3.21 for an exemplary scene). The system is capable of handling these situations according to our expectations. Most importantly, objects that are auditory and visually salient tend to attract the saliency even if they are not the most salient point in each modality. Furthermore, salient sound sources outside the visual field of view compete with visually salient stimuli and both are able to attract the overt focus of attention due to the normalized value ranges (see Sec. 3.4.4.A).

## 3.5 Multiobjective Exploration Path

Iteratively attending the most salient region that has not been attended yet is the classical approach to saliency-based overt and covert attention, see Sec. 3.4. However, in many situations, it is advisable to integrate further target criteria when planning where to look next. For example, it might be interesting to maximize the coverage of previously unseen areas with each attentional shift [MFL<sup>+</sup>07], integrate top-down target information for visual search ([OMS08]; *cf.* Ch. 4), or implement a task-dependent spatial bias to specific regions of the environment ([DBZ07]; *cf.* Ch. 4). In our opinion, it is also beneficial to minimize ego-motion. This has several benefits such as, among others: First, it can reduce the time to focus the next and/or all selected objects. Second, it can save the energy that is required to move joints. Third, it can reduce wear-and-tear of mechanical parts due to an overall reduction of servo movement. It also has another beneficial side-effect, because it often leads to less erratic and – according to our subjective impression – more human-like head motion patterns compared to saliency-only exploration strategies.

Given the detected salient proto-objects, we can mathematically address the problem of where to look next as an optimization problem, *i.e.* to determine the order of proto-objects that minimizes a given target function. By adapting the target function toward different criteria, we easily can implement a diverse set of exploration strategies. In the following, we present our balanced exploration approach that realizes a tradeoff between rapid saliency maximization and ego-motion minimization. However, we hope that you will agree with us that given our problem formulation it is easily possible to integrate further target criteria.

### 3.5.1. Exploration Path

We define an exploration path  $EP \in S(\mathcal{O})$ , *i.e.* the order in which to attend the proto-objects, as a permutation of the proto-objects  $\{o_1, o_2, \dots, o_N\} = \mathcal{O}$  that are scheduled to be attended. Here,  $S(\mathcal{O})$  is the permutation group of  $\mathcal{O}$  with  $|\mathcal{O}| = N!$ . For example, the exploration path  $EP_{\text{example}} = (o_1, o_3, o_2, o_4)$  would first attend object  $o_1$ , then  $o_3$  followed by  $o_2$ , and finally  $o_4$ . In the following, we denote  $s_{o_i}$  as the saliency of object  $o_i$  and  $q_{o_i}$  represents the robot’s joint angle configuration needed to focus object  $o_i$ . Accordingly, the target function that determines the optimal exploration path has the form  $f_{\text{target}} : S(\mathcal{O}) \rightarrow \mathbb{R}$  and in the following we define  $f_{\text{target}}$  and try to solve for

$$EP_{\text{opt}} = \arg \min_{EP \in S(\mathcal{O})} f_{\text{target}}(EP) . \quad (3.86)$$

### 3.5.2. Exploration Strategies

**Saliency-based exploration path** Analog to saliency-only bottom-up exploration as presented in Sec. 3.4, we can sort all perceived proto-objects by their

saliency  $s_{o_i}$  in descending order and attend the proto-objects in the resulting order  $\text{EP}_{\text{saliency}}$ , *i.e.*

$$\text{EP}_{\text{saliency}} = (o_{i_1}, o_{i_2}, \dots, o_{i_N}) \text{ with } s_{o_{i_1}} \geq \dots \geq s_{o_{i_N}}. \quad (3.87)$$

**Distance-based exploration path** Alternatively, we can ignore the saliency and try to minimize the accumulated joint angle distances that are necessary to attend all selected proto-objects

$$\text{EP}_{\text{distance}} = \arg \min_{\text{EP} \in S(\mathcal{O})} \left\{ \sum_{k=1}^N \left\| q_{o_{i_k}} - q_{o_{i_{k-1}}} \right\| \right\}, \quad (3.88)$$

where  $q_{o_{i_k}}$  represents the joint angles needed to focus the  $k^{\text{th}}$  object and  $q_{o_{i_{k-1}}}$  is the joint angle configuration for the preceding object. Here,  $q_{o_{i_0}}$  is defined as being the initial joint angle configuration at which we start the exploration. We use the norm of the joint angle differences  $d_{m,n} = \|q_m - q_n\|$  as a measure for the amount of necessary ego-motion between two joint configurations. Unfortunately, to determine the minimal accumulated distance to attend all proto-objects is an NP-complete problem, because it equates to the traveling salesman problem [CLR90, Weg05]<sup>4</sup>. Consequently, we limit the computation to  $K$  local neighbors of the currently focused object that were not already attended. In our implementation, we use  $K = 10$ , which seems to provide good results at acceptable computational costs. This strategy leads to paths that minimize the required amount of ego-motion, but it does not take the saliency into account.

**Balanced exploration path** Considering the exploration path planning as a multiobjective optimization problem [Ehr05], we can combine the saliency-based and distance-based approach. To this end, we define a single aggregate objective function

$$\text{EP}_{\text{balance}} = \arg \min_{\text{EP} \in S(\mathcal{O})} \left\{ \sum_{k=1}^N f_d(\|q_{o_{i_k}} - q_{o_{i_{k-1}}}\|) \cdot f_s(s_{o_{i_k}}) \right\}, \quad (3.89)$$

where  $s_{o_{i_k}}$  is the saliency value of the proto-object  $o_{i_k}$ ,  $f_d$  is a distance transformation function, and  $f_s$  is a saliency transformation function. We define  $f_d$  as identity function and  $f_s(s; \alpha) = s^{-\alpha}$ , *i.e.*

$$\text{EP}_{\text{balance}}(\alpha) = \arg \min_{\text{EP} \in S(\mathcal{O})} \left\{ \sum_{k=1}^N \|q_{o_{i_k}} - q_{o_{i_{k-1}}}\| \cdot s_{o_{i_k}}^{-\alpha} \right\}. \quad (3.90)$$

This aggregate optimization function implements the tradeoff between attending far away proto-objects with a high saliency and nearby proto-objects with a

<sup>4</sup>Please note that the traveling salesman problem (TSP)'s additional requirement to return to the starting city does not change the computational complexity.

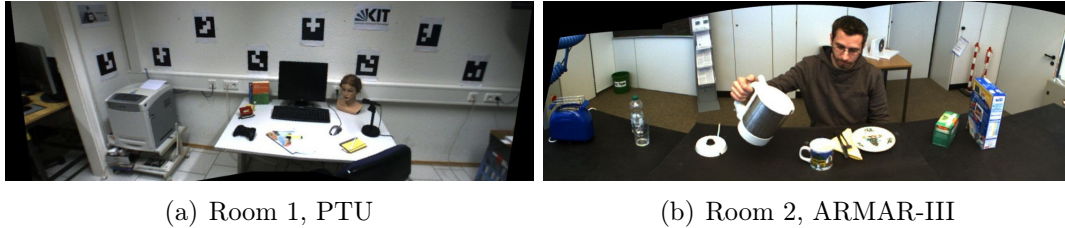


Figure 3.22.: Sample image stitches of the recordings with the stereo camera pan-tilt-unit (PTU) head (a) and with the ARMAR-III head (b).

scene	Number of Recordings		
	PTU/sensors	ARMAR-III	total
breakfast	15	15	30
office	10	10	20
neutral	5	5	10
total	30	30	60

Table 3.8.: Composition of the exploration path evaluation data.

lower saliency, where the choice of  $\alpha$  weights the target objective’s priorities. This optimization problem equates to an asymmetric TSP. It is asymmetric, because the aggregate function’s distance term depends on the object’s saliency, see Eq. 3.90, which leads to a different distance between two joint configurations depending on the end configuration. Accordingly, we search for an approximate solution and limit the search for the next best object to  $K$  local neighbors of the currently attended object.

### 3.5.3. Evaluation

Although it seems impossible to quantitatively evaluate the system behavior, see Sec. 3.4.5.B, we try to approach a quantitative evaluation of the exploration strategies in two steps: First, we record the whole environment in a scan sweep and calculate the locations of all salient proto-objects. This is similar to a person that takes a quick, initial glance around the room to get a first impression of an environment. Second, given a starting configuration, we can use the pre-calculated salient proto-object locations to plan the robot’s eye movement. This way, the first step enables us to analyze different methods to determine the salient regions and the second step makes it possible to analyze specific properties of the generated active behavior.

#### A. Data

We recorded a dataset that consists of 60 videos (30 seconds each) to evaluate our exploration strategies. The videos were recorded using two hardware platforms



in different environments, see Tab. 3.8 and Fig. 3.17. We re-enacted sequences in three scenarios: office scenes, breakfast scenes, and neutral scenes. Here, neutral scenes were recorded in the same environment, but with a reduced amount of salient objects.

## B. Evaluation Measures

Since a comparable evaluation has not been performed before, we had to develop novel evaluation measures that allow us to quantitatively compare the presented exploration strategies. We propose two evaluation measures that model different, competing goals.

We use the cumulated joint angle distance (CJAD) as measure of robot egomotion

$$\text{CJAD}(\text{EP}) = \frac{1}{N} \sum_{j=1}^N \|q_{\text{EP}_j} - q_{\text{EP}_{j-1}}\| , \quad (3.91)$$

where  $\text{EP}_j$  is the index of the  $j^{\text{th}}$  attended object of exploration path EP, and  $q_{o_i}$  represents the joint angle configuration that focuses object  $o_i$ , see Sec. 3.5.2. Since we want to reduce the amount of necessary head motion, we want to minimize the CJAD.

To investigate the influence of saliency on the exploration order, we use the cumulated saliency (CS) of already attended objects

$$\text{CS}(i; \text{EP}) = \sum_{j=1}^i s_{\text{EP}_j} , i \in \{1, 2, \dots, N\} . \quad (3.92)$$

We want to observe a steep growing curve, because this would mean that objects with higher saliency are attended first. Since the number of attended salient objects may vary depending on the saliency distribution in the scene, we denote the percentage of already attended objects as  $p$ , which makes it possible to integrate over the curves of different scenes. This way, we can calculate the area under the CS curve – we refer to it as integrated cumulated saliency (ICS) – as a compact evaluation measure, *i.e.*

$$\text{ICS}(\text{EP}) = \int \text{CS}(p; \text{EP}) dp . \quad (3.93)$$

Furthermore, we introduce NCJAD and NCS as normalized versions of CJAD and CS, respectively:

$$\text{NCJAD}(\text{EP}) = \frac{\text{CJAD}(\text{EP}) - \text{CJAD}(\text{EP}_{\text{distance}})}{\text{CJAD}(\text{EP}_{\text{saliency}}) - \text{CJAD}(\text{EP}_{\text{distance}})} \quad (3.94)$$

$$\text{NCS}(\text{EP}) = \frac{\text{ICS}(\text{EP}_{\text{saliency}}) - \text{ICS}(\text{EP})}{\text{ICS}(\text{EP}_{\text{saliency}}) - \text{ICS}(\text{EP}_{\text{distance}})} . \quad (3.95)$$

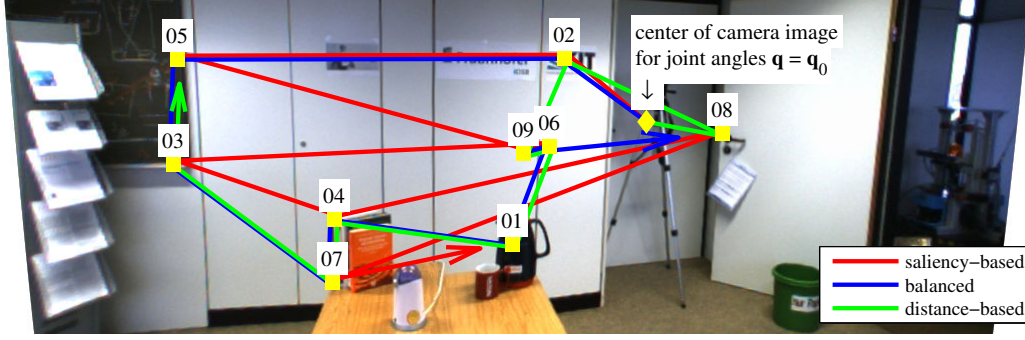


Figure 3.23.: An example to illustrate the different focus of attention selection strategies. The attention shifts for each path are illustrated in the stitched image (only the nine most salient locations are shown and yellow squares mark the positions of the objects). This illustration is best viewed in color.

The advantage of NCJAD and NCS is that they consider the spatial distribution of objects in the scene as well as their saliency distribution. This normalization terms

$$\text{CJAD}(\text{EP}_{\text{saliency}}) - \text{CJAD}(\text{EP}_{\text{distance}}) \quad \text{and} \quad (3.96)$$

$$\text{ICS}(\text{EP}_{\text{saliency}}) - \text{ICS}(\text{EP}_{\text{distance}}) \quad (3.97)$$

are the result of two considerations: First, the saliency-based exploration necessarily leads to the fastest growth of CS and thus highest ICS, but it is likely to have a high CJAD. Second, the distance-based strategy leads to the smallest CJAD, but is likely to exhibit a slow growth of CS.

To serve as a lower boundary for CS, we calculate  $\text{EP}_{\text{saliency}^*}$  which is the opposite strategy to  $\text{EP}_{\text{saliency}}$  that selects the least salient unattended object at each shift. Analogously, we calculate  $\text{EP}_{\text{distance}^*}$  which greedily selects the object with the highest distance at each step and is an approximate (greedy) strategy opposite to  $\text{EP}_{\text{distance}}$ .

### C. Results & Discussion

**Exploration path I – Saliency-based** First, we examine the saliency-based exploration approach (see Fig. 3.23, red; see Sec. 3.5.2) that is most widely found in related work and formed the basis for our qualitative experiments in Sec. 3.4. This strategy leads to a high amount of head movement (high CJAD), in fact the highest of all strategies, but it also leads to the highest growth of the cumulated saliency (high ICS; see Fig. 3.24). This leads to a slower exploration of all objects in the scene, but a fast analysis of the most salient objects.

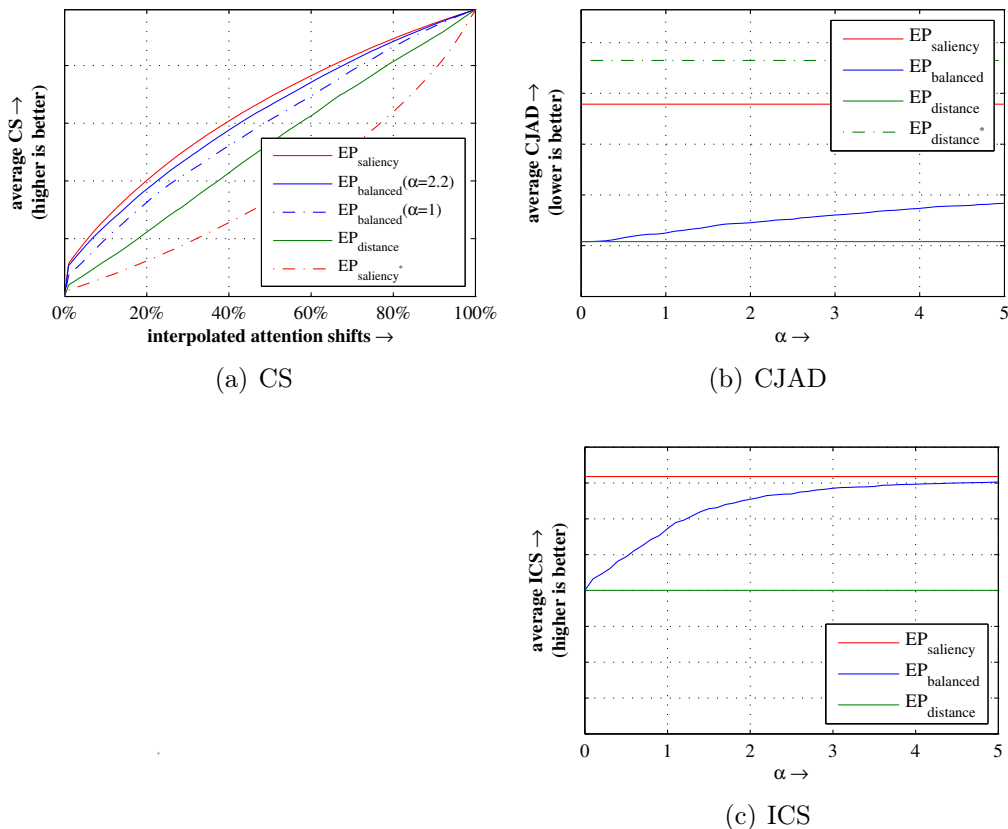


Figure 3.24.: The average cumulated saliency (a), the average cumulated joint angle distances (b), and the average area under the cumulated saliency curve (c) over all recordings in the database.

**Exploration path II – Distance-based** Second, we investigate the exploration strategy that minimizes the angular distances (see Fig. 3.23, green; see Sec. 3.5.2). Compared to the other strategies, the resulting exploration paths do not take into account saliency of proto-objects, which leads to the slowest cumulated saliency growth (low ICS; see Fig. 3.24). But, as can be seen in Fig. 3.24, the necessary angular distances and thus the time required for the full scene exploration is minimized (low CJAD). We would like to note that the computational limitation of using only  $K$  local neighbors for the TSP optimization (see Sec. 3.5.2) leads to a 25% longer distance in general [JM97].

**Exploration path III – Balanced** Finally, we consider the balanced strategy that implements a tradeoff between a small cumulated joint angle distance and steep growth of cumulated saliency (see Fig. 3.23, blue; see Sec. 3.5.2). We can adjust the priority of these two competing goals by changing the operating parameter  $\alpha$ . Interestingly, even a relatively high  $\alpha$  can already significantly reduce the CJAD while providing a high ICS, see Fig. 3.24.

When  $\alpha$  is set to 2.2, we achieve an average CJAD of 0.2972. For comparison the distance-based and saliency-based strategy achieve a CJAD of 0.2157 (72.6 %) and 0.7576 (254.9 %), respectively. At the same time, we achieve an average ICS of 156.5. Here, the distance-based and saliency-based strategy achieve an average ICS of 130.0 (83.1 %) and 161.8 (103.4 %), respectively. Thus, we provide an exploration strategy that effectively balances between favoring highly salient objects and efficient head movements.

## 3.6 Summary and Future Directions

---

We presented how we integrate saliency-driven, iterative scene exploration into a hierarchical, knowledge-driven audio-visual scene analysis approach. In principle, this follows the idea by Treisman *et al.*, see Sec. 2.1.1, and – to our best knowledge – has not been done to this extent by any other research group. To realize this system, we had to overcome several obstacles in different areas that we will recapitulate in the following.

When we started to work on audio-visual saliency-based exploration in 2010, we faced the situation that many methods that have been developed around the field of computational attention were not suited for use in real robotic systems. This was caused by the fact that most methods – including the saliency models themselves – were ill-suited for our use case, computationally too complex, or simply not state-of-the-art. Furthermore, only one comparable audio-visual robotic attention system existed [RLB<sup>+</sup>08], which relied on comparatively simple models and methods.

Although computational auditory attention seems to attract an increasing interest (*e.g.*, [NSK14, RMDB<sup>+</sup>13, SPG12]), still only few auditory saliency models exist (most importantly, ours and [KPLL05, Kal09]). And, the models that existed were computationally demanding and not suited for online processing. But, online processing was a necessary requirement for being able to immediately detect and react on interesting acoustic events (*e.g.*, a shattering glass or a person starting to speak). Having this goal in mind, we developed auditory surprise, which uses a Bayesian model to efficiently detect acoustic abnormalities.

For visual saliency detection, the situation was much better due to the multitude of visual saliency models. But, many computational models were too complex, requiring several seconds if not minutes to process a single video frame. We built on the work by Hou *et al.* [HZ07, HHK12] and derived quaternion-based models that are state-of-the-art in predicting human gaze patterns as well as computationally lightweight. Being able to calculate a saliency map in less than one millisecond, we developed the – to our best knowledge – fastest implementation of a state-of-the-art saliency model. Having seen that the color space can substantially influence the performance of spectral saliency models, we investigated color space decorrelation as a means to provide a more appropriate image-specific color space for low-level saliency models. This way, we were able to improve the performance of several, different visual saliency algorithms.

Equipped with applicable auditory and visual saliency models, we had to address how to represent the spatial saliency distribution in the 3D environment surrounding the robot. This was an essential aspect, because it would form the foundation for audio-visual saliency fusion and subsequent aspects such as, *e.g.*, implementing inhibition of return. Given our previous experience [SRP<sup>+</sup>09], we discarded the commonly found grid-like representations and tested a parametric Gaussian 3D model that implements the idea of salient 3D proto-objects. This

way, we can represent the spatial saliency distribution as a mixture of Gaussians, independent of the modality. This, of course, makes clustering and subsequent crossmodal fusion relatively easy to implement and computationally efficient. However, it is necessary to be able to efficiently transfer the visual and auditory saliency information into such a 3D proto-object model. While we simply adapted sound source localization toward auditory proto-objects, we proposed a novel method to efficiently detect and extract salient visual proto-objects based on the isophote curvature of the saliency map.

Being able to efficiently handle audio-visual saliency information in the 3D proto-object model, we could implement the actual scene exploration. This way, we could devise a balanced approach that combines the best aspects of two strategies that we encountered in the literature, *i.e.* try to minimize the ego-motion and investigate the most salient regions first. Furthermore, using the two strategies as baselines for good and bad behavior, we could derive evaluation measures to quantify the quality and tradeoffs made by the balanced approach.

**Future work** There remain many interesting directions for future work. With respect to audio-visual saliency detection, first, we see a lot of potential for better bottom-up as well as top-down auditory saliency models. An interesting development in this direction is the link between pupil dilation and auditory attention [WBM12], which might allow a quantitative evaluation methodology of bottom-up auditory saliency models that is not application oriented. In contrast, visual saliency detection is a very mature field, but there seems to be room for improvement when working with videos instead of images. Additionally, due to the rise of low-cost depth cameras such as, *e.g.*, Kinect the integration of depth information into visual saliency models is more important than ever before. It would also be very interesting to integrate high-level attentional modulation to incorporate task-based influences during, for example, visual search or human-robot interaction. However, we would like to note that all these aspects can be integrated seamlessly into our framework by adapting or replacing the auditory and visual attention models. With respect to our multiobjective exploration and the robot's overt attention, we think that it would be very interesting to investigate and evaluate how the generated head motion patterns can be made as human-like as possible.

# 4

## Multimodal Attention with Top-Down Guidance

In many situations, people want to guide our attention to specific objects or aspects in our environment. In fact, we have already seen an example of such attentional guidance as being part of advertisement design, see Fig. 4.1 or Fig. 2.10. However, such attentional guidance is not just a factor in effective advertisement. Instead, it is a natural process and part of everyday natural communication that we are not just frequently subjected to but often exercise ourselves – consciously as well as unconsciously. For example, when persons interact, interpreting non-verbal attentional signals such as, most importantly, pointing gestures [LB05] and gaze [BDT08] are essential to establish a joint focus of attention – *i.e.*, a common understanding of what we are talking and thinking about. Human infants develop the ability to interpret related non-verbal signals around the age of one year. This is a very important step in infant development, because it enables infants to associate verbal descriptions with the visual appearances of objects [Tom03, KH06]. This ability provides the means to acquire a common verbal dictionary and enable verbal communication with other humans, which is important to build strong social connections. As a consequence, there exists evidence “that joint attention reflects mental and behavioral processes in human learning and development” [MN07b].

In this chapter, we want to determine where other people want us to look at. In other words, we want to answer the question: Which object forms the intended

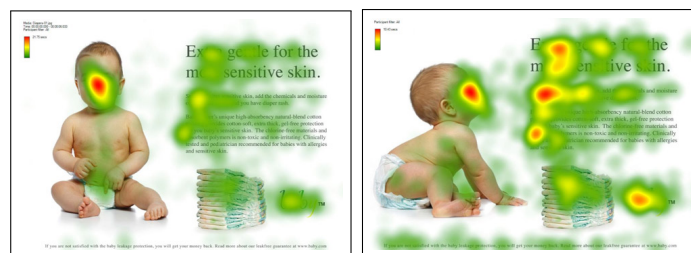


Figure 4.1.: Knowing how top-down cues can guide attention is an important aspect of advertisement design. Images from [Usa].

focus of the current scene or situation? We address two different domains: First, we investigate how we can model attentional guidance in human-robot interaction. During human-human and consequently human-robot interaction, we frequently use non-verbal (*e.g.*, pointing gestures, head nods, or eye gaze) and verbal (*e.g.*, object descriptions) to guide the attention of our conversation partner toward specific objects in the environment. Here, we focus on spoken object descriptions and pointing gestures. Second, we want to determine the object of interest in web images – in principle, in all forms of visual media that is created by human photographers, camera men, directors, etc. In such images and videos, it is not just the visible content such as, for example, the gaze of other people (see Fig. 4.1) that influences where we look, but the whole image is composed in such a way that the most important, relevant, and thus interesting object is highlighted. For example, photographer’s often place the object of interest in the image’s center or follow guidelines such as the rule of thirds to compose images.

In both domains, in contrast to our approach to evaluate visual saliency in Sec. 3.2, we do not try to predict where people will look. Instead, all images depict a specific object of interest and we try to determine this object. Therefore, we estimate a target object area and/or a sequence of target locations that are highly likely to be part of the intended object. This way, if we wanted to recognize the object, we only have to run our object classifiers on a very limited part of the image, thus saving computational resources. Furthermore, we can direct a robot’s sensors – *i.e.*, active vision – to focus on each of these target object hypotheses until we have seen the right object, which is similar to our approach in Sec. 3.4, but now includes top-down target information that we can acquire through interaction. If we want to learn – *e.g.*, from weakly labeled web images or through passive observation of human behavior –, then the limited set of image locations can serve as a prior to train better models<sup>1</sup>.

**Remainder** Complementary to our broad background discussion in Ch. 2, we provide an overview of related work that is relevant to understand the contributions presented in this chapter (Sec. 4.1). Then, we present how we adapt a state-of-the-art salient object detection algorithm to remove its implicit center bias (Sec. 4.2). Afterwards, we present how we can determine the target object in the presence of pointing gestures and linguistic descriptions of primitive visual attributes (Sec. 4.3). Finally, we conclude the technical part of this dissertation and present how we can use the introduced methods to determine objects that are being looked at in images that we collected from Flickr (Sec. 4.4).

**Acknowledgment** Part of the work described in this chapter has been done during my time in Gernot A. Fink’s “Pattern Recognition in Embedded Systems”

---

<sup>1</sup>We successfully implemented this idea to learn robust color term models from web images [SS12a]. For this purpose, we used salient object detection to serve as a spatial prior to weight the information at each image location.



---

computer science department and “Intelligent Systems” group at the Robotics Research Institute, TU Dortmund University. Furthermore, the pointing gesture detection implementation has been part of Jan Richarz’s dissertation and we would like to refer to his work for details on the method and implementation.

## 4.1 Related Work and Contributions

---

In the following, we first present the most important related work for each of the affected research topics. Then, after each topic’s overview, we discuss our contribution with respect to the state-of-the-art.

### 4.1.1. Joint Attention

To establish a joint focus of attention describes the human ability to verbally and non-verbally coordinate the focus of attention with interaction partners (see our introduction to Ch. 4). On one side this is achieved by directing the attention toward interesting objects, persons, or events, and on the other side by responding to these attention directing signals. Since this ability is one of the most important aspects of natural communication and social interaction, it has been addressed in various research areas such as, most importantly: psychology (*e.g.*, [Ban04, LB05, MN07a]), computational linguistics (*e.g.*, [SC09]), and robotics (*e.g.*, [Bro07, KH06, NHMA03, SC09, SKI<sup>+</sup>07, FBH<sup>+</sup>08, YSS10]). Especially for social robots the ability to initiate (*e.g.*, [DSS06, SC09, SKI<sup>+</sup>07]) and respond to (*e.g.*, [Bro07, SKI<sup>+</sup>07, TTDC06]) signals related to achieve joint attention are crucial aspects of natural and human-like interaction.

In the following, we address two specific aspects of responding to joint attention signals, *i.e.* how verbal object descriptions and pointing gestures can influence attention and guide visual search. In this context, we also address the influence of gaze, although we do not integrate or evaluate gaze as a feature in a human-robot interaction (HRI) domain.

#### A. Pointing

Pointing gestures are an important non-verbal signal to direct the attention toward a spatial region or direction and establish a joint focus of attention (*cf.* [Ban04, GRK07, HSK<sup>+</sup>10, LB05]). Accordingly, visually recognizing pointing gestures and inferring a referent or target direction has been addressed by several authors; *e.g.*, for interaction with smart environments (*e.g.*, [RPF08]), wearable visual interfaces (*e.g.*, [HRB<sup>+</sup>04]), and robots (*e.g.*, [HSK<sup>+</sup>10, KLP<sup>+</sup>06, NS07, SYH07, SHH<sup>+</sup>08, DSHB11]). Nickel and Stiefelhagen evaluated three different methods to estimate the indicated direction of a pointing gesture to determine the referent [NS07]. Most importantly, they achieved the best target object identification results with the line-of-sight model, *i.e.* the pointing ray originates from the hand and follows the head-hand direction. Interestingly, this model is even sometimes used in psychological literature (*e.g.*, [BO06]), where it has been found that the inherent inaccuracy of pointing gestures suggests that “pointing shifts attention into the visual periphery, rather than identifying referents” directly [Ban04]. Unfortunately, almost all technical systems require

that the objects present in the scene are already detected, segmented, recognized, categorized and/or their attributes identified (*e.g.*, [NS07, DSHB11]). For example, Droeschel *et al.* define that the pointed-at object is the object with the minimum distance to the pointing vector [DSHB11]. We would like to note at this point that we could simply implement Droeschel *et al.*'s approach even for unknown objects based on the proto-object model that we use for scene exploration in Sec. 3.4 and Sec. 3.5.

In principle, non-verbal signals such as pointing gestures circumscribe a referential domain to direct the attention toward an approximate spatial region (see [Ban04]). Naturally, this can clearly identify the referent in simple, non-ambiguous situations. However, as pointing gestures are inherently inaccurate in ambiguous situations (see [BI00, KLP<sup>+</sup>06]), context knowledge may be necessary to clearly identify the referent (see [LB05, SKI<sup>+</sup>07, SF10a]).

## B. Language

Language can provide contextual knowledge about the referent such as, *e.g.*, spatial relations and information about the object's visual appearance (see, *e.g.*, [KCP<sup>+</sup>13, SF10a]). Most importantly, verbal and non-verbal references can be seen to form composite signals, *i.e.* the speaker will compensate the inaccuracy or ambiguity of one signal with the other (see [Ban04, BC98, GRK07, KLP<sup>+</sup>06, LB05, Piw07, SKI<sup>+</sup>07]). Even without additional non-verbal signals, what we see in the environment or in a scene is often necessary to resolve ambiguities in otherwise ambiguous English sentences (*e.g.*, [KC06, HRM11]), while at the same time such sentences influence where we look in scenes, *i.e.* our gaze patterns.

When directly verbally referring to an object, most information about the referent is encoded in noun-phrases (see, *e.g.*, [MON08]), which consist of determiners (*e.g.*, "that"), modifiers (*e.g.*, "red") and a head-noun (*e.g.*, "book"). To analyze the structure of sentences and extract such information, tagging and shallow parsing can be applied. In corpus linguistics, part-of-speech (POS) tagging marks the words of a sentence with their grammatical function, *e.g.*, demonstrative, adjective, and noun. Based on these grammatical tags and the original sentence, shallow parsing determines the constituents of a sentence as, *e.g.*, noun-phrases. Commonly, machine learning methods are used to train taggers and shallow parsers on manually tagged linguistic corpora (*e.g.*, [Fra79, TB00]; *cf.* [Bri95]). The well-established Brill tagger uses a combination of defined and learned transformation rules for tagging [Bri95]. To apply the transformation rules it requires an initial tagging, which can be provided by stochastic n-gram or regular expression taggers (*cf.* [Bri95]).

## C. Color Terms

When verbally referring-to objects, relative and absolute features can be used to describe the referent (*cf.* [BC98]). Relative features require reference entities for

identification (*e.g.*, “the left cup”, or “the big cup”), whereas absolute features do not require comparative object entities (*e.g.*, “the red cup”). Possibly the most fundamental absolute properties of an object are its name, class, and color. When verbally referring to color, color terms (*e.g.*, “green”, “dark blue”, or “yellow-green”) are used to describe the perceived color (see [Moj05]). In [BK69], the cross-cultural concept of universal “basic color terms” is introduced, circumscribing that there exists a limited set of basic color terms in each language of which all other colors are considered to be variants (*e.g.*, the 11 basic color terms for English are: “black,” “white,” “red,” “green,” “yellow,” “blue,” “brown,” “orange,” “pink,” “purple,” and “gray”).

In order to relate the visual appearance of objects with appropriate color terms, color models for the color terms are required. Traditionally these models are either manually defined by experts or derived from collections of manually labeled color-chips (*cf.* [Moj05]). Alternatively, image search engines in the Internet can be used in order to collect huge weakly labeled datasets, which make it possible to use machine learning and train robust color naming models (*e.g.*, [vdWSV07]; [SS12a]).

#### D. Gaze

As we have seen in the introduction, where other people look at can and most likely will influence where a human observer will look. Like pointing gestures, gaze directions are a common signal to establish reference and infants show signs of following observed gaze directions of caregivers already at an age of 6 months [Hob05], well after infants have shown their first attraction to faces [CC03]. Again, like pointing gestures, an observed gaze direction steers the attention toward an approximate spatial region, along a corridor of attention, to establish a joint focus of attention (see, *e.g.*, [TTDC06, BDT08]).

Since where and at what people look at is an interesting information for many applications (*e.g.*, advertisement, driver assistance, and user interfaces), gaze estimation has been an active research area for over two decades (*e.g.*, [BP94, SYW97, TKA02, HJ10, VSG12]). But, despite all the research effort it has attracted, gaze estimation is still an unsolved problem; *e.g.*, even today there does not exist a gaze estimation method that can reliably estimate the gaze direction or – even more interestingly – the looked at object in an unconstrained domain such as web images. Most existing approaches focus on constrained scenarios such as, *e.g.*, limited head poses (*e.g.*, [SMSK08]) and/or rely on more reliable but only approximate estimates such as upper body orientation or head pose (*e.g.*, [VS08, RVES10]). Almost all approaches do not take into account the visible objects in the environment. Consequently, the estimation of a gaze direction and the subsequent deduction of the looked at object are treated as separate steps, where the latter is usually not even addressed by the authors or, similar to pointing gestures, requires that potential targets are already known.

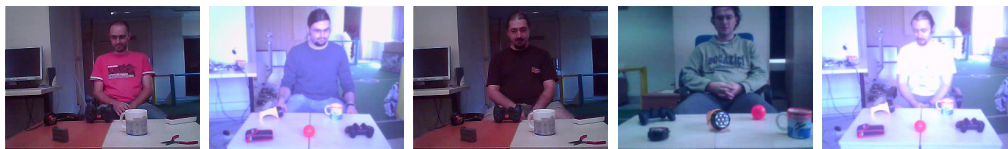


Figure 4.2.: Example images from Yücel *et al.*'s dataset [YScM<sup>+</sup>13].

Highly related to our work is the work by Yücel *et al.* [YScM<sup>+</sup>13], which combines visual saliency and an estimated gaze direction to highlight the object that is being looked at in a simple HRI scenario, see Fig. 4.2. Here, a person looks at objects on a table, where the objects are distributed apart from each other and have a high perceptual saliency.

**Contributions:** We model how verbal descriptions, pointing gestures, and visible eye gaze can influence visual attention and highlight intended target objects. We approach this topic with the intention of being able to interpret joint attention signals and identify the intended target object without or with only very limited information about the target object's visual appearance. For this purpose, our approach relies on saliency to direct the attention toward the referent. Accordingly, we use saliency as a kind of generalized object detector (see [ADF10]), which is related to the assumption that interesting objects are often visually salient [EI08]. Consequently, our work is very different to almost all work that tries to interpret pointing gestures and gaze signals (Yücel *et al.* [YScM<sup>+</sup>13] being a notable exception), because we do not require any a-priori knowledge about the objects in our environment. In principle, this enables us to use gestures and gaze to guide the attention and teach our system knowledge about previously unknown objects, which we demonstrated as part of the ReferAT dataset collection that will be presented in Sec. 4.3.2.E. This also means that, in contrast to most work on gaze and pointing gesture detection, recognition, and interpretation, we are not interested in a highly precise estimated gaze or pointing direction itself. Instead, our goal is to determine the image regions that most likely depict the looked-at or pointed-at target object.

Related to our scene exploration and analysis system (Sec. 3.4 and Sec. 3.5), this again means that by focusing on the most salient areas we can improve the perception through active vision – an aspect that we have demonstrated as part of our work [SRF10] and in Sec. 3.4 – and at the same time reduce the amount of data that has to be processed and analyzed. Here, it is interesting that our approach almost guarantees it that the correct target object will be focused after only a few focus of attention shifts.

As a sidenote, although we do not present the details of how we learn color term models in this thesis, we have improved the state-of-the-art in color term learning with two innovations: First, we developed a probabilistic color model to learn color models that better reflect natural color distributions despite the

fact they have been trained on web images that contain many post-processed or even entirely artificial images [SF10b]. We have shown that this way the learned models make more “human-like” errors, if they make errors. Second, we use salient object detection as a means to predict and weight the relevance and thus influence of each image pixel’s color information during training [SS12a].

### 4.1.2. Visual Attention

In principle, visual saliency models try to predict “interesting” image regions that are likely to attract human interest and, as a consequence, gaze. We have already discussed many aspects of bottom-up visual attention in earlier sections (Sec. 2.1.1, 3.1.1, and 3.2). Consequently, we do not address bottom-up saliency models and focus on two related but different types of models: First, salient object detection methods that try to identify and segment the most important or prominent object in an image. Second, visual attention models that allow to integrate knowledge for goal-directed adaptation of the visual saliency to support visual search.

#### A. Salient Object Detection

Most generally, “salient regions” in an image are likely to grab the attention of human observers. The task of “traditional” saliency detection is to predict where human observers look when presented with a scene, which can be recorded using eye tracking equipment, see Sec. 3.2. In 2007, Liu *et al.* adapted the traditional definition of visual saliency by incorporating the high level concept of a salient object into the process of visual attention computation [LSZ<sup>+</sup>07]. A “salient object” is defined as being the (most prominent) object in an image that attracts most of the user’s interest. Accordingly, Liu *et al.* [LSZ<sup>+</sup>07] defined the task of “salient object detection” as the binary labeling problem to separate the salient object from the background. Here, it is important to note that the selection of a salient object happens consciously by the user whereas the gaze trajectories that are recorded using eye trackers are the result of mostly unconscious processes. Consequently, also taking into account that salient objects attract human gaze (see, *e.g.*, [ESP08]), salient object detection and predicting where people look are very closely related yet substantially different tasks.

In 2009, Achanta *et al.* [AHES09, AS10] introduced a salient object detection approach that basically relies on the difference of pixels to the average color and intensity value. To evaluate their approach, they selected a subset of 1000 images of the image dataset that was collected from the web by Liu *et al.* [LSZ<sup>+</sup>07] and calculated segmentation masks of the salient objects that were marked by 9 participants using (rough) rectangle annotations [LSZ<sup>+</sup>07]. Since it was created, the salient object dataset by Achanta *et al.* serves as reference dataset to evaluate methods for salient object detection (see, *e.g.*, [AHES09, AS10, KF11, CZM<sup>+</sup>11]).

Since Liu *et al.* defined salient object detection as binary labeling (*i.e.*, binary segmentation) problem, it comes at no surprise that Liu *et al.* applied conditional random fields (CRFs) to detect salient objects, because CRFs have achieved state-of-the-art performance for several segmentation tasks such as, *e.g.*, semantic scene segmentation (*e.g.*, [LMP01, PPI09, VT08]). Here, semantic segmentation describes the task of labeling each pixel of an image with a semantic category (*e.g.*, “sky”, “car”, “street”). Closely related to Bayesian surprise (see Sec. 3.3.1), Klein *et al.* [KF11] use the Kullback-Leibler Divergence of the center and surround image patch histograms to calculate the saliency. Cheng *et al.* [CZM<sup>+</sup>11] use segmentation to define a regional contrast-based method, which simultaneously evaluates global contrast differences and spatial coherence. In general, we can differentiate between algorithms that rely on segmentation-based (*e.g.*, [CZM<sup>+</sup>11, ADF10]) and pixel-based contrast measures (*e.g.*, [AHES09, AS10, KF11]).

It has been observed in several eye tracking studies that human gaze fixation locations in natural scenes are biased toward the center of images and videos (see, *e.g.*, [Bus35, Tat07, PN03]). One possible bottom-up cause of the bias is intrinsic bottom-up visual saliency as predicted by computational saliency models. One possible top-down cause of the center bias is known as photographer bias (see, *e.g.*, [RZ99, PN03, Tat07]), which describes the natural tendency of photographers to place objects of interest in the center of their composition. In fact, what the photographer considers interesting may also be highly perceptually, bottom-up salient. Additionally, the photographer bias may lead to a viewing strategy bias [PLN02], which means that viewers may orient their attention more often toward the center of the scene, because they expect salient or interesting objects to be placed there. Thus, since in natural images and videos the distribution of objects of interest and thus saliency is usually biased toward the center, it is often unclear how much the saliency actually contributes in guiding attention. It is possible that people look at the center for reasons other than saliency, but their gaze happens to fall on salient locations. Therefore, this center bias may result in overestimating the influence of saliency computed by the model and contaminate the evaluation of how visual saliency may guide orienting behavior. Recently, Tseng *et al.* [TCC<sup>+</sup>09] were able to demonstrate quantitatively that center bias is correlated strongly with photographer bias and is influenced by viewing strategy at scene onset. Furthermore, *e.g.*, they were able to show that motor bias had almost no effect. Here, motor bias refers to a preference of short saccades over long saccades [Tat07, TCC<sup>+</sup>09]. This can affect the distribution of fixated image locations in eye tracking experiments, because in most free viewing experiments the participants are asked to start viewing from a central image location [TCC<sup>+</sup>09] (*e.g.*, for purpose of calibration or consistency).

Interestingly, although it is now a well-studied aspect of eye tracking experiments to such an extent that it has become an integral part of evaluation measures (see Sec. 3.2.1.F), the photographer bias has neither been thoroughly studied nor well modeled in the field of salient object detection. Most importantly, in Jiang *et al.*'s work [JWY<sup>+</sup>11] one of the criteria that characterize a salient object is

that “it is most probably placed near the center of the image”, which is justified with the “rule of thirds” (see, *e.g.*, [Pet03]). Most recently, Borji *et al.* [BSI12] evaluated several salient object detection models and also performed tests with an additive Gaussian center bias and conclude that the resulting “change in accuracy is not significant and does not alter model rankings”. But, this study neglected the possibility that well-performing models already have integrated, implicit biases.

**Contributions:** We provide an empirical justification why a Gaussian center bias is in fact beneficial for salient object detection in web images. Then, we show that implicit, undocumented biases are at least partially responsible for the performance of state-of-the-art algorithms and adapt the segmentation-based method by Cheng *et al.* [CZM<sup>+</sup>11] to remove its implicit center bias. This way, we achieve four goals: First, we could invalidate the statement that salient object detection is unaffected by a photographer or otherwise incurred center bias (see [BSI12]). Second, we could quantify the influence that an integrated center bias can have on salient object detection models. Third, we could improve the state-of-the-art in salient object detection on web images through the integration of an explicit, well-modeled center bias. Fourth, we derived the currently best performing unbiased algorithm. The latter aspect is especially interesting for many applications domains in which the image data is not biased by a photographer (*e.g.*, autonomous robots and cars, or surveillance).

Furthermore, with respect to our work on target object detection in the presence of top-down guidance, our task is substantially different to all prior art on salient object detection: First, we try to integrate top-down information such as pointing gestures, gaze, and language. Second, we do not limit ourselves to web images and thus the image data does not necessarily have a photographer bias. Third, our target objects are substantially smaller compared to typical salient objects in the most important datasets (see [LSZ<sup>+</sup>07, AHES09]).

## B. Visual Search

It has been shown that knowledge about the target object influences the saliency to speed-up the visual search (see [STET01, WHK<sup>+</sup>04]). However, not every piece of knowledge can influence the perceptual saliency. Instead, only specific information that refers to preattentive features allows such guidance [WHK<sup>+</sup>04]. For example, knowing the specific object or at least its color reduces the search slope, whereas categorical (*e.g.*, “animal” or “post card”) information typically does not provide top-down guidance (see [WHK<sup>+</sup>04]). Accordingly, in recent years, various computational saliency models have been developed that are able to integrate top-down knowledge in order to guide the attention in goal-directed search (*e.g.*, [TCW<sup>+</sup>95, IK01b, FBR05, Fri06, NI07, WAD09, Wel11]). However, the number of saliency models that have been designed for goal-directed search is small compared to the vast amount of bottom-up saliency models (see Sec. 2.1.1



and 3.1.1), which might be symptomatic for the fact that there does not exist any established dataset to evaluate top-down visual search algorithms.

Most importantly, Navalpakkam and Itti [NI07] introduced a saliency model that allows to predict the visual search pattern given knowledge about the visual appearance of the target and/or distractors. In principle, this can be seen as an implementation of Wolfe *et al.*'s guided search model (GSM) [WCF89, Wol94], see Sec. 2.1.1. Navalpakkam and Itti use the knowledge about the target's appearance to maximize the expected signal-to-noise ratio (SNR), *i.e.* target-to-distractor ratio, of the saliency combination across and within feature dimensions. For this purpose, every feature dimension (*e.g.*, orientation or intensity) is additionally subdivided by neurons with broadly overlapping Gaussian tuning curves to model varying neuron sensitivities to different value ranges within each feature dimension. Then, for each neuron's response the center-surround contrast is calculated to form each neuron's feature map. This way, it is possible to assign a higher or lower weight to salient aspects within value bands of each feature dimension; for example, to assign a higher importance to the response of neurons that encode very bright areas or roughly 45° edges.

**Contributions:** In contrast to prior art, we do not just focus on isolated non-verbal or verbal aspects. Instead, we integrate all available knowledge provided by different but complementing modalities in a computational model. Therefore, we use CRFs to integrate bottom-up saliency models, salient object detection methods, spatial corridors of attention given by gaze and pointing gestures, and – if available – spoken object descriptions. We also show that for our task the CRFs are able to significantly outperform neuron-based approaches such as Navalpakkam and Itti's model [NI07], which was adapted by Schauerte and Fink to use spectral saliency to calculate each neuron's feature map [SF10a]. We would like to note that Navalpakkam and Itti's neuron-based approach, albeit its age, is still the most established model in the field.

This way, we are often able to select the correct target object, even in complex situations. It is important to note that even for isolated aspects our latest datasets (*i.e.*, ReferAT and Gaze@Flickr) are substantially more complex than what has been used by other research groups (compare, *e.g.*, [Fri06, BK11] for visual search and [SKI<sup>+</sup>07, NS07] for pointed-at objects). An interesting aspect of our approach is that it is able to accurately segment the target object in most "simple" situations, although it can only predict salient regions of interest in complex situations.

## 4.2 Debiased Salient Object Detection

As we addressed in the introduction of Ch. 4, we are interested in two different data domains, see Fig. 4.3: First, human-robot interaction, in which the people try to direct a robot’s attention. Second, web images, in which a photographer tries to highlight and direct our attention to certain aspects of the scene. Here, the photographer might (*e.g.*, see the Gaze@Flickr dataset, Sec. Sec. 4.4.2) or might not (*e.g.*, see the MSRA dataset, Sec. Sec. 4.2.1) use persons and visible non-verbal cues to guide the attention. Naturally, the images from the two domains follow different biases such as, for example, that the target objects in photographs are often substantially larger compared to the target objects in our human-robot interaction scenes.

Web images and photographs in general form a domain that is substantially different from images that are not directly composed by humans (*e.g.*, surveillance footage, robot and unmanned aerial vehicle camera images, etc.). This is due to the fact that photographers follow image composition rules such as, for example, the rule of thirds (see, *e.g.*, [Pet03]), which leads to very specific biases. We have already seen that such image composition biases have an important influence on saliency models, because one of the reasons why the AUC evaluation measure (Sec. 3.2.1.F) is favored by many researchers is that it compensate for such biases; most importantly, the center-bias that is commonly found in eye tracking datasets is linked to the photographer bias [TCC<sup>+</sup>09].

Nowadays, most work on salient object detection focuses on web images and MSRA is the dominant dataset in that research area. As a consequence, this means that we can expect that the algorithms are (over-)adapted to that specific domain. However, current state-of-the-art salient object detection algorithms are very powerful and, consequently, we would like to build on that work and derive an unbiased salient object detection algorithm that can help us as a feature for other application domains such as, for example, surveillance footage or robotics.



Figure 4.3.: Example images of the MSRA, Gaze@Flickr, PointAT, and ReferAT datasets to illustrate the domain specific differences.

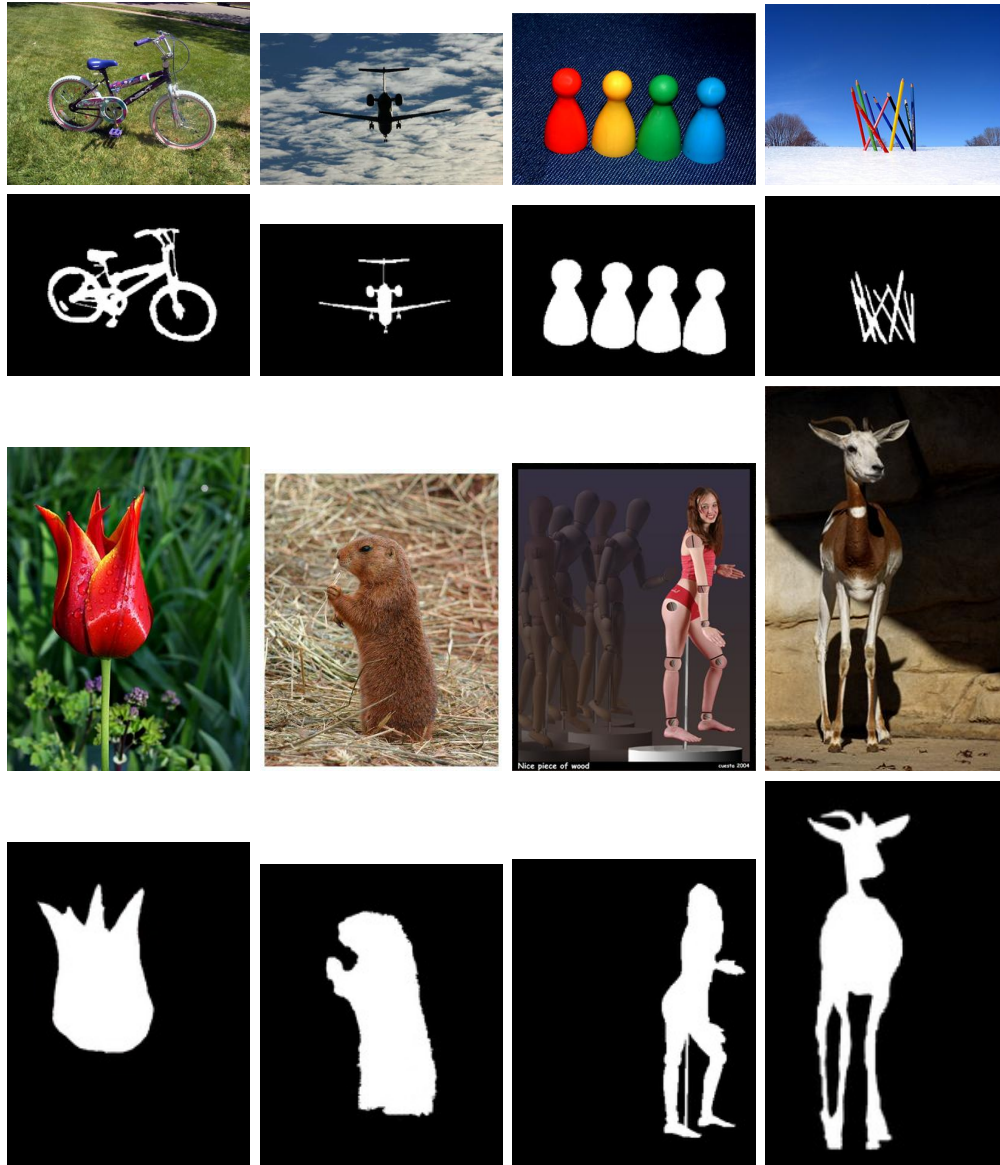


Figure 4.4.: Example images from the MSRA dataset. Depicted are selected images with their binary target segmentation masks.

### 4.2.1. The MSRA Dataset

The most important salient object detection dataset is the MSRA dataset that has been created by Achanta *et al.* and Liu *et al.* [AHES09, LSZ<sup>+</sup>07], see Fig. 4.4. The MSRA dataset is based on the salient object dataset by Liu *et al.* [LSZ<sup>+</sup>07] and consists of a subset of 1000 image for which Achanta *et al.* provide annotated segmentation masks [AHES09], see Fig. 4.4. The images in Liu *et al.*'s dataset have been collected from a variety of sources, mostly from image forums and image search engines. Liu *et al.* collected more than 60,000 images and subsequently selected an image subset in which all images contain a salient object or a distinctive foreground object [LSZ<sup>+</sup>07]. Then, 9 users marked the salient objects using (rough) bounding boxes and the salient objects in the image database have been defined based on the “majority agreement”. However, as a consequence of the selection process, the dataset does not include images without distinct salient objects and is potentially biased by the human selectors. This is an important aspect to consider when trying to generalize the results reported on Achanta *et al.*'s and Liu *et al.*'s dataset to other datasets or application areas.

**Dataset Properties** The 1000 images contain 1265 annotated target object regions. On average a target object region occupies 18.25% of the image area. Furthermore, as we will show in the following, the object locations in the dataset are strongly biased toward the center of the image.

### 4.2.2. MSRA's Photographer Bias

To investigate the spatial distribution of salient objects in photographs, we use the segmentation masks by Achanta *et al.* [AHES09, AS10]. More specifically, we use the segmentation masks to determine the centroids of all salient objects in the dataset and analyze the centroids' spatial distribution.

#### A. Salient Object Distribution Model

**The center** Our model is based on a polar coordinate system that has its pole at the image center. Since the images in Achanta's dataset have varying widths and heights, we use in the following normalized Cartesian image coordinates in the range  $[0, 1] \times [0, 1]$ . The mean salient object centroid location is  $[0.5021, 0.5024]^T$  and the corresponding covariance matrix is  $[0.0223, -0.0008; -0.0008, 0.0214]$ . Thus, we can motivate the use of a polar coordinate system that has its pole at  $[0.5, 0.5]^T$  to represent all locations relative to the expected distribution's mode.

**The angles are distributed uniformly** Our first model hypothesis is that the centroids' angles in the specified polar coordinate system are uniformly distributed in  $[-\pi, \pi]$ .

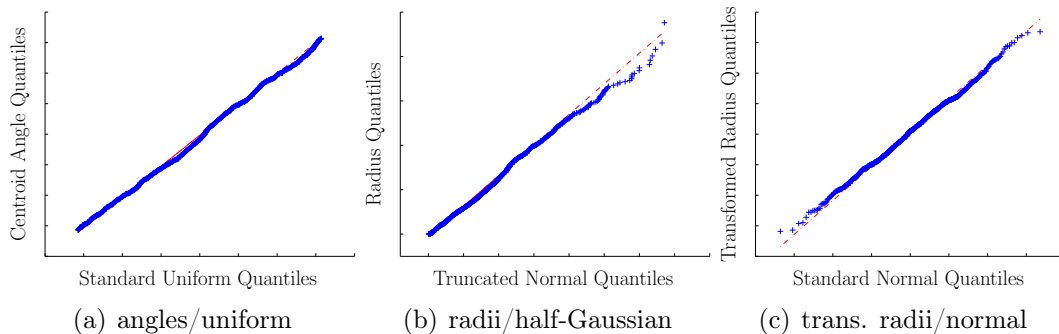


Figure 4.5.: Quantile-quantile (Q-Q) plots of the angles versus a uniform distribution (a), radii versus a half-Gaussian distribution (b), transformed radii (see Sec. 4.2.2.B) versus a normal distribution (c).

To investigate the hypothesis, we use a quantile-quantile (Q-Q) plot as a graphical method to compare probability distributions (see [NIS12]). In Q-Q plots the quantiles of the samples of two distributions are plotted against each other. Thus, the more similar the two compared distributions are, the better the points in the Q-Q plot will approximate the line  $f(x) = x$ . We calculate the Q-Q plot of the salient object location angles in our polar coordinate system versus uniformly drawn samples in  $[-\pi, \pi]$ , see Fig. 4.5(a). The apparent linearity of the plotted Q-Q points supports the hypothesis that the angles are distributed uniformly.

**The radii follow a half-Gaussian distribution** Our second model hypothesis is that the radii of the salient object locations follow a half-Gaussian distribution. We have to consider a truncated distribution in the interval  $[0, \infty)$ , because the radius – as a length – is by definition positive. If we consider the image borders, we could assume a two-sided truncated distribution, but we have three reasons to work with a one-sided model: The variance of the radii seems sufficiently small, the “true” centroid of the salient object may be outside the image borders (*i.e.*, parts of the salient object can be truncated by the image borders), and it facilitates the use of standard statistical tests (see Sec. 4.2.2.B).

We use a Q-Q plot against a half-Gaussian distribution to graphically assess the hypothesis, see Fig. 4.5(b). The linearity of the points suggests that the radii are distributed according to a half-Gaussian distribution. The visible outliers in the upper-right are caused by less than 30 centroids that are highly likely to be disturbed by the image borders. Please be aware of the fact that it is not necessary to know the half-Gaussian (or standard Gaussian) distribution’s model parameters when working with Q-Q plots (see [NIS12]).

## B. Empirical Hypothesis Analysis

We can quantify the observed linearity in the Q-Q plots, see Fig. 4.5, to analyze the correlation between the model distribution and the data samples using probability plot correlation coefficient (PPCC) [NIS12]. The PPCC is the correlation coefficient between the paired quantiles and measures the agreement of the fitted distribution with the observed data (*i.e.*, goodness-of-fit). The closer the correlation coefficient is to one, the higher the positive correlation and the more likely the distributions are shifted and/or scaled versions of each other. By comparing against critical values of the PPCC (see [VK89] and [NIS12]), we can use the PPCC as a statistical test that is able to reject the hypothesis that the observed samples come from identical distributions. This is closely related to the Shapiro-Wilk test [SW65]. Furthermore, we can use the correlation to test the hypothesis of no correlation by transforming the correlation to create a t-statistic.

Although often data analysts prefer to use graphical methods such as Q-Q plots to assess the feasibility of a model, formal statistical hypothesis tests remain the most important method to disprove hypotheses. The goal of statistical tests is to determine if the (null) hypothesis can be rejected. Consequently, statistical tests either reject (prove false) or fail to reject (fail to prove false) a null hypothesis. But, they can never prove it true (*i.e.*, failing to reject a null hypothesis does not prove it true). However, we can disprove alternate hypotheses and, additionally, we can use a set of statistical tests that are based on different principles. If all tests fails, we have – at least – an indicator that the hypothesis is potentially true.

**The angles are distributed uniformly** The obvious linearity of the Q-Q plot, see Fig. 4.5(a), is reflected by a PPCC of 0.9988<sup>2</sup>, which is substantially higher than the critical value of 0.8880 (see [VK89]) and thus the hypothesis of identical distributions can not be rejected. Furthermore, the hypothesis of no correlation is rejected at  $\alpha = 0.05$  ( $p = 0$ ).

We use Pearson's  $\chi^2$  test [Pea00] as a statistical hypothesis test against a uniform distribution. The test fails to reject the hypothesis at significance level  $\alpha = 0.05$  ( $p = 0.2498$ ). Considering the circular type of data, we use Rayleigh's and Rao's tests for circular uniformity and both tests fail to reject the hypothesis at  $\alpha = 0.05$  ( $p = 0.5525$  and  $p > 0.5$ , respectively; see [Bat81]). On the other hand, for example, we can reject the alternative hypotheses of a normal or exponential distribution using the Lilliefors test [Lil67] ( $p = 0$  for both distributions<sup>3</sup>).

---

<sup>2</sup>Mean of several runs with  $N = 1000$  uniform randomly selected samples.

<sup>3</sup>We report  $p = 0$ , if the tabulated values are 0 or the Monte Carlo approximation returns 0 or  $\epsilon$  (double-precision).



Figure 4.6.: An example of the influence of the center bias on segmentation-based salient object detection. Left-to-right: Image, region contrast without and with center bias (RC'10 and RC'10+CB, resp.), and locally debiased region contrast without and with center bias (LDRC and LDRC+CB, resp.).

**The radii follow a half-Gaussian distribution** In order to use standard statistical hypothesis tests, we transform the polar coordinates in such a way that they represent the same point with a combination of positive angles in  $[0, \pi]$  and radii in  $[-\infty, \infty]$ . According to our hypothesis, the distribution of the transformed radii should follow a normal distribution with its mode and mean at 0, see Fig. 4.5(c).

The correlation that is visible in the Q-Q plot, see Fig. 4.5(b) and 4.5(c), is reflected by a PPCC of 0.9987, which is above the critical value of 0.9984 (see [NIS12]). The hypothesis of no correlation is rejected at  $\alpha = 0.05$  ( $p = 0$ ).

Again we disprove exemplary alternate hypotheses: The uniform distribution is rejected by the test against the critical value of the PPCC as well as by Pearson's  $\chi^2$  test at  $\alpha = 0.05$  ( $p = 0$ ). The exponential distribution is rejected by Lilliefors test at  $\alpha = 0.05$  ( $p = 0$ ). We perform the Jarque-Bera, Lilliefors, Spiegelhalter's, and Shapiro-Wilk test (see [BJ80], [Lil67], [Spi83] and [SW65]) to test our null hypothesis that the radii have been sampled from a normal distribution (unknown mean and variance). Subsequently, we use a T-test to test our hypothesis that the mean of the radius distribution is 0. The Jarque-Bera, Lilliefors, Spiegelhalter's, and Shapiro-Wilks tests fail to reject the hypothesis at significance level  $\alpha = 0.05$  ( $p = 0.8746$ ,  $p = 0.2069$ ,  $p = 0.2238$ , and  $p = 0.1022$ , respectively). Furthermore, it is likely that the mode of the (transformed) radius is 0, because the corresponding T-test fails to reject the hypothesis at significance level  $\alpha = 0.05$  with  $p = 0.9635$ .

### 4.2.3. Salient Object Detection

#### A. Algorithm

We adapt the region contrast model by Cheng *et al.* [CZM<sup>+</sup>11]. Cheng *et al.*'s model is particularly interesting, because it already provides state-of-the-art performance, which is partially caused by an implicit center bias. Thus, we can observe how the model behaves if we remove the implicit center bias, which was neither motivated nor explained by the authors, and add an explicit Gaussian center bias. We extend the spatially weighted region contrast saliency equation

$S_{\text{RC}}$  (see Eq. 7 in [CZM<sup>+</sup>11]) and integrate an explicit, linearly weighted center bias:

$$S_{\text{RC+CB}}(r_k) = w_{\text{B}}S_{\text{RC}}(r_k) + w_{\text{C}}g(C(r_k); \sigma_x, \sigma_y) \quad \text{with} \quad (4.1)$$

$$S_{\text{RC}}(r_k) = \sum_{r_k \neq r_i} \hat{D}_s(r_k; r_i)w(r_i)D_r(r_k; r_i) \quad \text{and} \quad (4.2)$$

$$\hat{D}_s(r_k; r_i) = \exp(-D_s(r_k; r_i)/\sigma_s^2) . \quad (4.3)$$

Here, we use a convex combination<sup>4</sup> to control the strength of the influence of the center bias, *i.e.*  $w_{\text{B}} + w_{\text{C}} = 1$  ( $w_{\text{B}}, w_{\text{C}} \in \mathbf{R}_0^+$ ).  $\hat{D}_s(r_k; r_i)$  is the spatial distance between regions  $r_k$  and  $r_i$ , where  $\sigma_s$  controls the spatial weighting. Smaller values of  $\sigma_s$  influence the spatial weighting in such a way that the contrast to regions that are farther away contributes less to the saliency of the current region. The spatial distance between two regions is defined as the Euclidean distance between the centroids of the respective regions using pixel coordinates that are normalized to the range  $[0, 1] \times [0, 1]$ . Furthermore,  $w(r_i)$  is the weight of region  $r_i$  and  $D_r(\cdot; \cdot)$  is the color distance metric between the two regions (see [CZM<sup>+</sup>11] for more details). Here, the number of pixels in  $r_i$  is used as  $w(r_i) = |r_i|$  to emphasize color contrast to bigger regions.  $C(r_k)$  denotes the centroid of region  $r_k$  and  $g$  is defined as follows

$$g(x, y; \sigma_x, \sigma_y) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{1}{2}\frac{x^2}{\sigma_x^2}\right\} * \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left\{-\frac{1}{2}\frac{y^2}{\sigma_y^2}\right\} . \quad (4.4)$$

Interestingly, the unnormalized Gaussian weighted Euclidean distance used by Cheng *et al.* [CZM<sup>+</sup>11] causes an implicit Gaussian-like center bias, see Fig. 4.7, because it favors regions whose distances to the other neighbors are smaller. Unfortunately, this has not been motivated, discussed, or evaluated by Cheng *et al.* To remove this implicit bias, we introduce a normalized, *i.e.* locally debiased, distance function  $\check{D}_s(r_k; r_i)$  that still weights close-by regions higher than further away regions, but does not lead to an implicit center bias

$$\check{D}_s(r_k; r_i) = \frac{\hat{D}_s(r_k; r_i)}{\sum_{r_j} \hat{D}_s(r_k; r_j)} , \quad (4.5)$$

$$i.e. \quad \forall r_k : \sum_{r_i} \check{D}_s(r_k; r_i) = 1 \quad , \quad (4.6)$$

and define

$$S_{\text{LDRC}}(r_k) = \sum_{r_k \neq r_i} \check{D}_s(r_k; r_i)w(r_i)D_r(r_k; r_i) \quad \text{and} \quad (4.7)$$

$$S_{\text{LDRC+CB}}(r_k) = w_{\text{B}}S_{\text{LDRC}}(r_k) + w_{\text{C}}g(C(r_k); \sigma_x, \sigma_y) . \quad (4.8)$$

<sup>4</sup>We have considered different combination methods and provide a quantitative evaluation of different combination types in Appx. D. In short, the convex combination achieves the best overall performance of the evaluated combination methods.



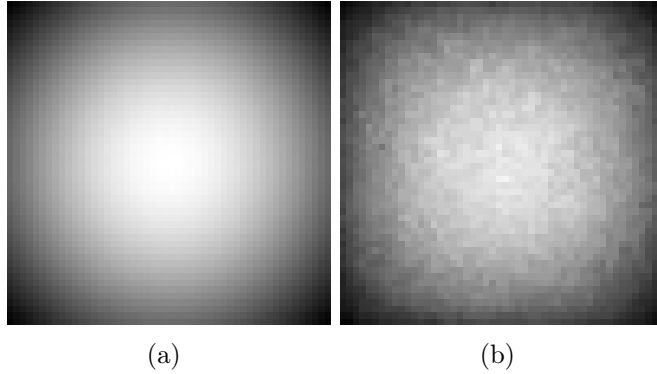


Figure 4.7.: Illustration of the implicit center bias in the method by Cheng *et al.* [CZM<sup>+</sup>11]. (a): Each pixel shows the distance weight sum, *i.e.*  $\sum_{r_i} \hat{D}_s(r_k; r_i)$ , to all other pixels in a regular grid. (b): The average weight sum depending on the centroid location calculated on the MSRA dataset.

## B. Evaluation

**Baseline algorithms** To compare our results, we use a set of state-of-the-art salient object detection algorithms: Achanta *et al.*'s frequency-tuned model (AC'09; [AHES09]), Achanta *et al.*'s maximum symmetric surround saliency model (MSSS'10; [AS10]), Klein *et al.*'s information-theoretic saliency model (BITS'11; [KF11]), and Cheng *et al.*'s region contrast model (RC'10; [CZM<sup>+</sup>11]) that uses Felzenszwalb's image segmentation method [FH04]. The latter is the original algorithm that we adapted in Sec. 4.2.3.A.

Additionally, we include the results of four algorithms that have not been designed for salient object detection: Itti and Koch's model [IKN98] as implemented by Itti's iLab Neuromorphic Vision Toolkit (iNVT'98), Harel *et al.*'s graph-based visual saliency (GBVS'07; [HKP07]), Goferman *et al.*'s context-aware saliency (CAS'12; [GZMT12]), and Guo *et al.*'s pure Fourier transform (PFT'07; [GMZ08, GZ10]; see Sec. 3.2.1.A).

To investigate the influence of the implicit center bias in the region contrast model, we calculate the performance of the locally debiased region contrast (LDRC) model without and with our explicit center bias (LDRC and LDRC+CB, respectively). For comparison, we also evaluate the region contrast model with the additional explicit center bias (RC'10+CB). As additional baseline, we provide the results for simple segment-based and pixel-based – *i.e.*, using Eq. 4.4 for each pixel with respect to the image center distance and constant variance – center bias models, *i.e.*  $w_C = 1$  (CB<sub>S</sub> and CB<sub>P</sub>, respectively).

**Measures** We can use the binary segmentation masks for saliency evaluation by treating the saliency maps as binary classifiers. At a specific threshold  $t$  we

Method	$F_\beta$	$F_1$	$\int\text{ROC}$	PHR
LDRC+CB	0.8183	0.8034	0.9624	0.9240
RC'10+CB	0.8120	0.7973	0.9620	0.9340
RC'10	0.7993	0.7855	0.9568	0.9140
LDRC	0.7675	0.7574	0.9430	0.8680
BITS'11	0.7582	0.7342	0.9316	0.7540
MSSS'10	0.7337	0.7165	0.9270	0.8420
GBVS'07	0.6242	0.6403	0.9088	0.8480
PFT'07	0.6009	0.5995	0.8392	0.7100
CB <sub>S</sub>	0.5764	0.5793	0.8623	0.6980
CAS'12	0.5615	0.5857	0.8741	0.6920
CB <sub>P</sub>	0.5452	0.5604	0.8673	0.7120
iNVT'98	0.4012	0.3383	0.5768	0.6870

Table 4.1.: The maximum  $F_1$  score, maximum  $F_\beta$  score, ROC AUC ( $\int\text{ROC}$ ), and PHR of the evaluated algorithms (sorted by descending  $F_\beta$ ).

regard all pixels that have a saliency value above the thresholds as positives and all pixels with values below the thresholds as negatives. By sweeping over all thresholds  $\min(S) \leq t \leq \max(S)$ , we can evaluate the performance using common binary classifier evaluation measures.

We use four evaluation measures to quantify the performance of the evaluated algorithms. We calculate the area under the curve (AUC) of the ROC curve ( $\int\text{ROC}$ ). Complementary to the  $\int\text{ROC}$ , we calculate the maximum  $F_1$  and  $F_{\sqrt{0.3}}$  scores (see [AHES09] and Sec. 3.3.2.A).  $F_\beta$  with  $\beta = \sqrt{0.3}$  has been proposed by Achanta *et al.* to weight precision higher than recall for salient object detection [AHES09]. Additionally, we calculate the PHR, see Sec. 4.3.1.E, which measures how often the pixel with the maximum saliency belongs to a part of the target object.

**Results** The performance of RC'10 drops substantially if we remove the implicit center bias as is done by LDRC, see Tab. 4.1. However, if we add our explicit center bias model to the unbiased model, the performance is substantially increased with respect to all evaluation measures. Furthermore, with the exception of pixel hit rate (PHR), the performance of LDRC+CB and RC'10+CB is nearly identical with a slight advantage for LDRC+CB. This indicates that we did not lose important information by debiasing the distance metric (LDRC+CB vs RC'10+CB) and that the explicit Gaussian center bias model is advantageous compared to the implicit weight bias (LDRC+CB and RC'10+CB vs RC'10).

Most interestingly, LDRC is the best model without center bias, which makes it interesting for applications in which the image data can not be expected to have a photographer's center bias (*e.g.*, image data of surveillance cameras or autonomous robots).

(a) PointAT				
Algorithm	PHR	FHR	$\int$ FHR	NTOS
	none			
RC'10	0.94%	1.88%	0.029	0.659
LDRC	2.83%	4.24%	0.045	0.792
	automatic			
RC'10	53.77%	67.92%	0.817	5.901
LDRC	58.49%	74.53%	0.828	5.831
	annotated			
RC'10	61.79%	78.30%	0.843	6.596
LDRC	60.85%	80.66%	0.852	6.431
(b) ReferAT				
Algorithm	PHR	FHR	$\int$ FHR	NTOS
	none			
RC'10	0.00%	1.65%	0.025	0.226
LDRC	0.00%	2.47%	0.026	0.427
	automatic			
RC'10	2.48%	19.42%	0.421	3.190
LDRC	4.96%	28.51%	0.551	3.333
	annotated			
RC'10	4.13%	23.97%	0.502	3.778
LDRC	7.02%	30.99%	0.621	3.916

Table 4.2.: Target object prediction results of RC'10 and its debiased counterpart LDRC on the PointAT and ReferAT datasets. A description of the evaluation measures can be found in Sec. 4.2.3.B.

#### 4.2.4. Debiased Salient Object Detection and Pointing

Although it is interesting that we were able to improve the state-of-the-art in salient object detection by analyzing and explicitly modeling task specific biases, see Sec. 4.2.3.B, our initial motivation has been to improve the ability of salient object detection algorithms to perform well in other domains such as, *e.g.*, our PointAT and ReferAT pointing datasets. We can see in Tab. 4.2 that our debiased LDRC algorithm leads to substantial performance improvements in predicting the right target objects in three conditions: First, without any directional information from the pointing gestures, see Tab. 4.2 “none”. Second, in combination with a heuristic model – the heuristic method will be described in the subsequent Sec. 4.3 – to integrate automatically detected pointing gestures, see Tab. 4.2 “automatic”. Third, in combination with the heuristic method and manually annotated pointing information, see Tab. 4.2 “annotated”.

However, we will see in the following section that LDRC itself is not the best predictor for this task and its performance is surpassed by our spectral saliency models (see Sec. 3.2). This is most likely caused by the small target object sizes that are atypical for common salient object detection tasks. Nonetheless, LDRC will turn out to be a useful feature that helps us to highlight the right target object area. Furthermore, being able to control the center bias allows us to seamlessly disable the center bias for PointAT and ReferAT while adding a center bias when processing Gaze@Flickr in Sec. 4.4.

## 4.3 Focusing Computational Attention in Human-Robot Interaction

---

As briefly mentioned in the introduction to Ch. 4, verbal and non-verbal signals that guide our attention are an essential aspect of natural interaction and help to establish a joint focus of attention (*e.g.*, [Ban04, BI00, GRK07, Piw07, Roy05, SC09]). In other words, the ability to generate and respond to attention directing signals allows to establish a common point of reference or conversational domain with an interaction partner, which is fundamental for “learning, language, and sophisticated social competencies” [MN07a].

When talking about the focus of attention (FoA) in interaction, we have to distinguish between the FoA within the conversation domain (*i.e.*, what people are talking about), and the perceptual focus of attention (*e.g.*, where people are looking at). In many situations, the conversational FoA and the perceptual FoA can and will be distinct. However, when persons are referring to specific objects within a shared spatial environment, multimodal – here, non-verbal and verbal – references are an important part of natural communication to direct the perceptual FoA toward the “referent”, *i.e.* the referred-to object, and achieve a shared conversational FoA. Accordingly, we have to distinguish between the saliency of objects in the context of the conversation domain at some point during the interaction and the inherent, perceptual saliency of objects present in the scene (see [BC98]). Although the conversational domain is most important when identifying the referent – especially when considering object relations –, the perceptual saliency can influence the generation and interpretation of multimodal referring acts to such extent that in some situations “listeners [...] identify objects on the basis of ambiguous references by choosing the object that was perceptually most salient” [BC98, CSB83].

In this section, we focus on a situation in which an interacting person uses non-verbal (*i.e.*, pointing gestures) and verbal (*i.e.*, specific object descriptions) signals to direct the attention of an interaction partner toward a target object. Consequently, our task is to highlight the intended target object. One of the challenges in such a situation is that we may know nothing about the target object’s appearance. In fact, it might be the actual goal of the multimodal reference to teach something about the referent; for example, imagine a pointing gesture that is accompanied by an utterances like “Look at that! I bet you have never seen Razzmatazz<sup>5</sup> before!”.

Interestingly, exactly in situations in which we know nothing about the target object’s appearance, we can use visual saliency models (see Sec. 3.2 and 4.2) to determine image regions that are highly likely to render potential objects of interest. And, as we have mentioned before, the actual multimodal reference itself might even be influenced by the perceptual saliency. Furthermore, some

---

<sup>5</sup>Razzmatazz is a color name and describes a shade of rose or crimson.

target information (*e.g.*, information about the target’s color) is known to subconsciously influence our visual attention system and accordingly we can try to integrate such information as well, if it is available. For both situations, *i.e.* pointing gestures in the absence (Sec. 4.3.1) or presence (Sec. 4.3.2) of verbal object descriptions, we present two approaches: First, purpose-built heuristic and neuron-based models that were specifically proposed for this purpose by Schauerte and Fink. Second, we use machine learning with conditional random fields, which has two advantages: This approach leads to a better, more robust predictive performance and it is simpler to integrate additional features and target information.

### 4.3.1. Pointing Gestures

Like gaze, pointing gestures direct the attention into the visual periphery, which is indicated by the pointing direction (see Sec. 4.1). The pointing gesture’s directional information is defined by the origin  $o$  – usually the hand or finger – and an estimation of the direction  $d$ . The referent, *i.e.* the object that is being pointed at, can then be expected to be located in the corridor of attention alongside the direction. However, the accuracy of the estimated pointing direction depends on multiple factors: the inherent accuracy of the performed gesture (see [BO06, BI00, KLP<sup>+</sup>06]), the method to infer the pointing direction (see [NS07]), and the automatic pointing gesture detection itself.

#### A. Pointing Gesture Detection

In the following, we briefly describe how we detect pointing gestures, calculate the indicated pointing direction, assess the detected pointing gesture’s inaccuracy, and model how likely the pointing gesture directs the attention to specific image locations. As has been done by Nickel and Stiefelhagen [NS07], we use the line-of-sight model to calculate the indicated pointing direction. In this model, the pointing direction is equivalent to the line-of-sight defined by the position of the eyes  $h$  and the pointing hand  $o$ , and accounts for “the fact that [in many situations; A/N] people point by aligning the tip of their pointing finger with their dominant eye” [BO06]. To recognize pointing gestures, we adapted Richarz *et al.*’s approach [RPF08]<sup>6</sup>. Without going into any detail, we replaced the face detector with a head shoulder detector based on histogram of oriented gradients (HOG) features to improve the robustness. We detect the occurrence of a pointing gesture by trying to detect the inherent holding phase of the pointing hand. Therefore, we group the origin  $o_t$  and direction  $d_t$  hypotheses over time  $t$  and select sufficiently large temporal clusters to detect pointing occurrences.

---

<sup>6</sup>Please note that the approaches presented in this section are independent of the actual pointing gesture recognition method – *e.g.*, we are currently working toward using Microsoft Kinect – as long as it is possible to derive an angular (in-)accuracy measure.

We consider three sources of pointing inaccuracy: Due to image noise and algorithmic discontinuities the detected head-shoulder rectangles exhibit a position and scaling jitter. In order to model the uncertainty caused by estimating the eye position from that detection rectangle  $r_t$  (at time  $t$ ), we use a Normal distribution around the detection center  $\bar{r}_t$  to model the uncertainty of the estimated eye position

$$p_e(x|r_t) = \mathcal{N}(\bar{r}_t, \sigma_e^2) \quad . \quad (4.9)$$

$\sigma_e$  is chosen so that one quarter of  $\bar{s}$  is covered by  $2\sigma_e$ , *i.e.*

$$\sigma_e = \bar{s}/8 \quad , \quad (4.10)$$

where  $\bar{s}$  is the mean of the detection rectangle's size over the last image frames. Furthermore, we consider the variation in size of the head-shoulder detection rectangle, and the uncertainty of the estimated pointing direction  $d$ , which is caused by shifts in the head and hand detection centers. We treat them as independent Gaussian noise components and estimate their variances  $\sigma_s^2$  and  $\sigma_d^2$ . As  $\sigma_e^2$  and  $\sigma_s^2$  are variances over positions, we approximately transfer them into an angular form by normalizing with the length  $r = \|d\|$

$$\tilde{\sigma}_e^2 = \frac{\sigma_e^2}{r^2} \quad \text{and} \quad \tilde{\sigma}_s^2 = \frac{\sigma_s^2}{r^2} \quad , \quad (4.11)$$

respectively. This approximation has the additional benefit to reflect that the accuracy increases when the distance to the pointer decreases and the arm is outstretched.

**Spatial pointing target probability** Due to Eq. 4.11, the combined accuracy distribution has become a distribution over angles

$$p_G(x) = p(\alpha(x; o, d)|d, o) = \mathcal{N}(0, \sigma_c^2) \quad , \quad (4.12)$$

with  $\alpha(x; o, d)$  being the angle between the vector from the pointing origin  $o$  to the image point  $x$  given the pointing direction  $d$ , and

$$\sigma_c^2 \approx \tilde{\sigma}_e^2/r^2 + \tilde{\sigma}_s^2/r^2 + \sigma_d^2 \quad . \quad (4.13)$$

This equation models the probability  $p_G(x)$  that a point  $x$  in the image plane was referred-to by the pointing gesture given the current head-shoulder detection  $d$  and the pointing direction  $o$ , and thus defines our corridor of attention. To account for the findings by Kranstedt *et al.* [KLP<sup>+</sup>06], we enforce a lower bound of  $3^\circ$ , *i.e.*

$$\hat{\sigma}_c = \max(3^\circ, \sigma_c) \quad , \quad (4.14)$$

so that 99.7%, which corresponds to  $3\sigma$ , of the distribution's probability mass covers at least a corridor of  $9^\circ$ .

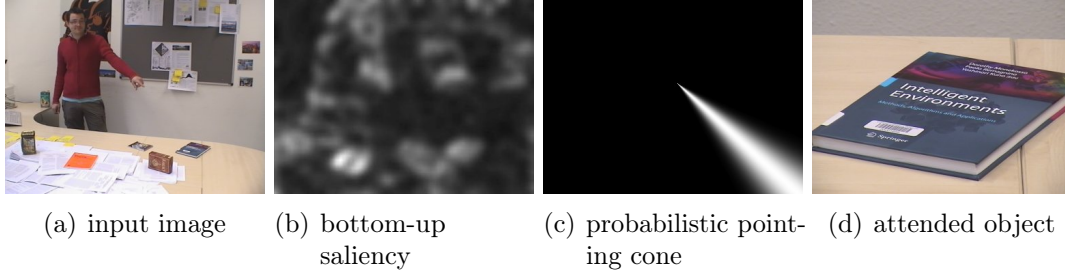


Figure 4.8.: Heuristic detection of the referent.

## B. Heuristic Integration

Given a detected pointing gesture and the spatial target probability  $p_G(x)$ , see Eq. 4.12, we can calculate a top-down feature map

$$S_t(a, b) = p_G(x) \quad (4.15)$$

with  $x = (a, b)$ . This, in effect, defines a blurred cone of Gaussian probabilities – which encode how likely it is that the referent is located at the pixel’s position – emitted from the hand along the pointing direction in the image plane, see Fig. 4.8(c). Due to its form, we refer to this map as being the probabilistic pointing cone (PPC).

We can now use visual saliency to act as a form of generalized object detector. For this purpose, we use QDCT to calculate the visual saliency map  $S_b$  based on the PCA decorrelated CIE Lab color space, see Sec. 3.2. The saliency map  $S_b$  is normalized to  $[0, 1]$  and each pixel’s value is interpreted as being the probability that this pixel is part of the target object.

The final saliency map  $S$  is then obtained by calculating the joint probability

$$S = S_b \circ S_t, \quad (4.16)$$

where  $\circ$  represents the Hadamard product. In other words, we highlight image regions alongside the pointing direction that are highly likely to contain a salient (proto-)object.

## C. Conditional Random Field

**Structure, Learning, and Prediction** In general, a CRF models the conditional probabilities of  $x$  (here, “does this pixel belong to a target object?”), given the observation  $y$  (*i.e.*, features), *i.e.*

$$p(x|y) = \frac{1}{Z(y)} \prod_{c \in C} \psi(x_c, y) \prod_{i \in V} \psi(x_i, y) \quad , \quad (4.17)$$

where  $C$  is the set of cliques in the CRF’s graph and  $i$  represent individual nodes. Here,  $\psi$  indicates that the value for a particular configuration  $x_c$  depends on the input  $y$ .



Naturally, our problem is a binary segmentation task, since the location depicted by a pixel can either belong to the target object or not, *i.e.*  $x_i$  can either be “target” or “background”. We use a pairwise, 4-connected grid CRF structure. We linearly parametrize the CRF parameter vector  $\Theta$  in unary node  $u(y, i)$  (*i.e.*, information at an image location) and edge features  $v(y, i, j)$  (*e.g.*, relating neighbored image locations). Here, it is important to consider that the cliques in a 4-connected, grid-structured graph are the sets of connected nodes, which are represented by the edges. Thus, we fit two matrices  $F$  and  $G$  such that

$$\Theta(x_i) = Fu(y, i) \quad (4.18)$$

$$\Theta(x_i, x_j) = Gv(y, i, j) . \quad (4.19)$$

Here,  $y$  is the observed image and  $\Theta(x_i)$  represents the parameter values for all values of  $x_i$ . Similarly,  $\Theta(x_i, x_j)$  represents the parameter values for all  $x_i, x_j$ . Then, we can calculate

$$p(x; \Theta) = \exp \left[ \sum_i \Theta(x_i) + \sum_j \Theta(x_i, x_j) - A(\Theta) \right] , \quad (4.20)$$

where  $A(\Theta)$  is the log-partition function that ensures normalization.

We use tree-reweighted belief propagation (TRW) to perform approximate marginal inference, see [WJ08]. TRW addresses the problem that it is computationally intractable to compute the log-partition function  $A(\Theta)$  exactly and approximates  $A(\Theta)$  with

$$\hat{A}(\Theta) = \max_{\mu \in \mathcal{L}} \Theta \cdot \mu + \hat{H}(\mu) , \quad (4.21)$$

where  $\hat{H}$  is TRW’s entropy approximation [WJ08]. Here,  $\mathcal{L}$  denotes the valid set of marginal vectors

$$\mathcal{L} = \left\{ \mu : \sum_{x_c \setminus i} \mu(x_c) = \mu(x_i) \wedge \sum_{x_i} \mu(x_i) = 1 \right\} , \quad (4.22)$$

where  $\mu$  describes a mean vector, which equals a gradient of the log-partition function. Then, the approximate marginals  $\hat{\mu}$  are the maximizing vector

$$\hat{\mu} = \arg \max_{\mu \in \mathcal{L}} \Theta \cdot \mu + \hat{H}(\mu) . \quad (4.23)$$

This can be approached iteratively until convergence or a maximum number of updates [Dom13].

To train the CRF, we rely on the clique loss function, see [WJ08],

$$L(\Theta, x) = - \sum_c \log \hat{\mu}(x_c; \Theta) . \quad (4.24)$$

Here,  $\hat{\mu}$  indicates that the loss is implicitly defined with respect to marginal predictions – again, in our implementation these are determined by TRW – and not the true marginals. This loss can be interpreted as empirical risk minimization of the mean Kullback-Leibler divergence of the true clique marginals to the predicted ones.

**Features** As unary image-based features, we include the following information at each CRF grid point: First, we include each pixel’s normalized horizontal and vertical image position in the feature vector. Second, we directly use the pixel’s intensity value after scaling the image to the CRF’s grid size. Third, we include the scaled probabilistic pointing cone (PPC), see Sec. 4.3.1.A. Then, after scaling each saliency map to the appropriate grid size, we append QDCT image signature saliency maps based on the PCA decorrelated Lab color space (see Sec. 3.2.1) at three scales:  $96 \times 64$  px,  $168 \times 128$  px, and  $256 \times 192$  px. Optionally, we include the LDRC salient object prediction, see Sec. 4.2.

As CRF edge features, first, we use a constant of one that allows to model general neighborhood relations. Second, we use 10 thresholds to discretize the  $L^2$  norm of the color difference and thus contrast of neighboring pixels. Then, we multiply the existing features by an indicator function for each edge type (*i.e.*, vertical and horizontal), effectively doubling the number of features and encoding conjunctions of features and edge type. This way, we parametrize vertical and horizontal edges separately [Dom13].

#### D. The PointAT Dataset

To assess the ability of systems to identify arbitrary target objects in the presence of a pointing gesture, we collected a dataset that contains 220 images of 3 persons pointing at various objects [SRF10]. The dataset was recorded in two environments (an office space and a conference room) with a large set of objects of different category, shape, and texture. The dataset focuses on the object detection and recognition capabilities and was not supposed to be used to evaluate the performance of pointing gesture detection, which explains the limited number of pointing persons.

**Hardware Setup** The dataset was recorded using a monocular Sony EVI-D70P pan-tilt-zoom camera. The camera provides images in approximately PAL resolution ( $762 \times 568$  px), and offers an optical zoom of up to  $18\times$ . Its wide horizontal opening angle is  $48^\circ$ . To reflect a human or humanoid point of view, we mounted the camera on eye height of an averagely tall human [MFOF08].

**Procedure** Each person performed several pointing sequences, with varying numbers and types of objects present in the scene. We neither restricted the body posture of the subjects in which pointing gestures had to be performed, nor

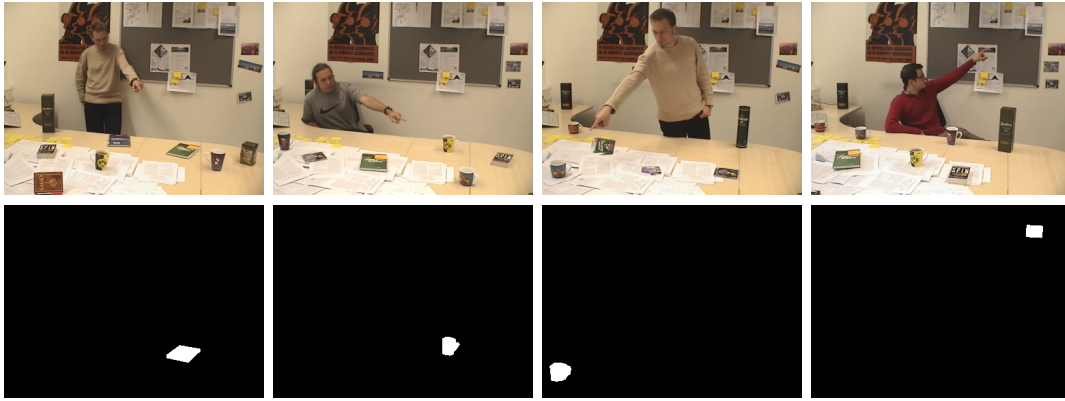


Figure 4.9.: Examples of pointing gestures performed in the evaluation. Depicted are some exemplary images with the corresponding binary target object masks that we generate based on the annotated target object boundaries.

did we define fixed positions for the objects and persons. The only restriction imposed was that the subjects were instructed to point with their arms extruded, so that pointing gestures would comply with the line-of-sight model employed. To evaluate the ability of the iterative shift of attention to focus the correct object in the presence of distractors, we occasionally arranged clusters of objects so that the object reference would be ambiguous. Accordingly, the dataset contains a wide variety of pointing references, see Fig. 4.9. Since we do not specifically evaluate the pointing detector (see [RPF08]), we discard cases with erroneous pointing gesture detections. Thus, in total, our evaluation set contains 220 object references.

For each object reference, we manually annotated the target object’s boundaries.<sup>7</sup> Furthermore, we annotated the dominant eye, the pointing finger, and the resulting pointing direction. This makes it possible to assess the influence that the automatic pointing gesture recognition’s quality has on the identification of the pointed-at object.

**Properties** On average the target object occupies only 0.58% of the image area. In other words, at random we would require roughly 200 trials to expect to select one pixel that is part of the target object. The average differences between the annotated and automatically determined pointing origin and direction are 15.30 px and 2.95°, respectively. The former is mostly caused by the fact that the system detects the center of the hand, instead of the finger. The latter is due to the fact that the eye positions are estimated given the head-shoulder detection (see [RPF08]), and that the bias introduced by the dominant eye is

<sup>7</sup>Please note that we re-annotated the original dataset to calculate the results presented in this dissertation, because the original annotations only consisted of bounding boxes. Accordingly, our results are not directly comparable to previously reported results [SRF10].

unaccounted for (see [BO06]). In some cases, the ray that originates from the finger tip (*i.e.*, the pointing origin) and follows the pointed direction does not intersect the target object’s boundaries, *i.e.* it “misses” the object. The rate of how often the object’s annotated boundary polygon is missed by the pointing ray is 5.66% for the annotated pointing information and 19.34% for the automatic detection.

### E. Evaluation Measures

The fovea is responsible for detailed, sharp central vision. As such, it is essential for all tasks that require visual details such as, most importantly, many recognition tasks. However, the fovea itself comprises less than 1% of the retinal area and can only perceive the central 2° of the visual field. In the following, we define that an object has been perceived or “focused”, if it or a part of it has been projected onto the fovea. To make our evaluation independent of the recording equipment and image resolution – an important aspect for our evaluation on the Gaze@Flickr dataset (Sec. 4.4.2) –, we assume that a (hypothetical) human observer sits in front of a display on which the image is shown in full screen mode. This way, we can estimate the extent of the display’s – and thus the image’s – area that would be projected onto the model observer’s fovea. Here, we assume a circular fovea area and, thus, a circular model FoA. We approximate the radius  $r_{\text{FoA}}$  of the FoA on the display as follows

$$r_{\text{FoA}} = \tan(a_{\text{FoA}}/2) \frac{\sqrt{w^2 + h^2}}{D} d, \quad (4.25)$$

where  $a_{\text{FoA}}$  is the angle perceived by the fovea’s visual field,  $d$  is the distance between the viewer’s eyes and the screen, and  $D$  is the display’s diameter. For example, this results in an FoA radius of 7.5 px for a fovea angle of 2°, a viewing distance of 65 cm – not untypical for office environments –, a 60.96 cm (24 inch) display diagonal, and an image resolution of 320 × 240 px. In the following, we define a fovea angle of 2°, a viewing distance of 65 cm, and a 24 inch display diagonal for all evaluations.

We derive two related evaluation measures: The pixel hit rate (PHR) measures how often the most salient pixel lies within the object boundaries (see [SS13a, SF10a, SRF10]). The focus of attention hit rate (FHR) measures how often the object is covered – and thus perceived – at least partially by the FoA (see [SF10a, SRF10]). To compute the FoA hit rate (FHR), we calculate whether the radial FoA and the annotated object’s boundary polygon collide. Here, the assumption of an FoA area (*i.e.*, FHR) instead of a simple FoA point (*i.e.*, PHR) has an important benefit: Since saliency models tend to highlight edges, the most salient point is often related to the object’s boundaries and as a consequence can be located just a bit outside of the actual object, very close to the boundary.

Additionally, we can calculate the FHR after shifting the focus of attention to the next most salient region in the image. To this end, we inhibit the location

Algorithm	PHR	FHR	$\int$ FHR	NTOS
	no pointing			
Heuristic	7.54%	10.84%	0.238	2.094
CRF, no LDRC	9.43%	16.04%	0.257	2.110
CRF, w/ LDRC	10.38%	17.92%	0.246	2.045
	pointing detected			
Heuristic	59.91%	82.08%	0.838	7.425
CRF, no LDRC	80.66%	92.45%	0.879	9.727
CRF, w/ LDRC	81.60%	92.45%	0.874	9.734
	pointing annotated			
Heuristic	77.83%	88.68%	0.859	8.143
CRF, no LDRC	84.91%	91.98%	0.877	10.189
CRF, w/ LDRC	85.38%	92.92%	0.878	10.083

Table 4.3.: Target object detection performance on the PointAT dataset. A description of the evaluation measures can be found in Sec. 4.3.1.E.

that has already been attended, *i.e.* we set the saliency of all pixels within the current FoA to zero. Let  $FHR^{+k}$  denote the FoA hit rate after  $k$  attentional shifts, *i.e.* how frequently the target is perceived within the first  $k$  shifts of attention. Then, we can integrate over the  $FHR^{+k}$  until a given  $k \leq n$  (in the following, we set  $n = 10$ ). We refer to this measure as  $\int$ FHR and it has the advantage that it also reflects the cases in which the target object has not been found after  $n$  shifts.

Furthermore, similar to the normalized scanpath saliency (NSS) saliency measure (see Appx. C; [PIIK05, PLN02]), we calculate the mean saliency of the pixels that are part of the object area to compare the target area’s saliency to the background saliency. This measure has a different purpose than PHR and FHR, because it does not directly evaluate the ability to focus the target object. Instead, it measures how strong the target object is highlighted against – *i.e.*, separated from – the background. We call this measure normalized target object saliency (NTOS). For this purpose, the saliency map is normalized to have zero mean and unit standard deviation [PIIK05, PLN02], *i.e.* a NTOS of 1 means that the saliency in the target object area is one standard deviation above average. Consequently, an  $NTOS \geq 1$  indicates that saliency map has significantly higher saliency values at target object locations. An  $NTOS \leq 0$  means that the saliency does not predict a target object location better than picking random image locations.

## F. Evaluation Results

**Procedure** To train and evaluate our models, we use a leave-one-person-out training procedure. Furthermore, we mirror the samples along the vertical axis

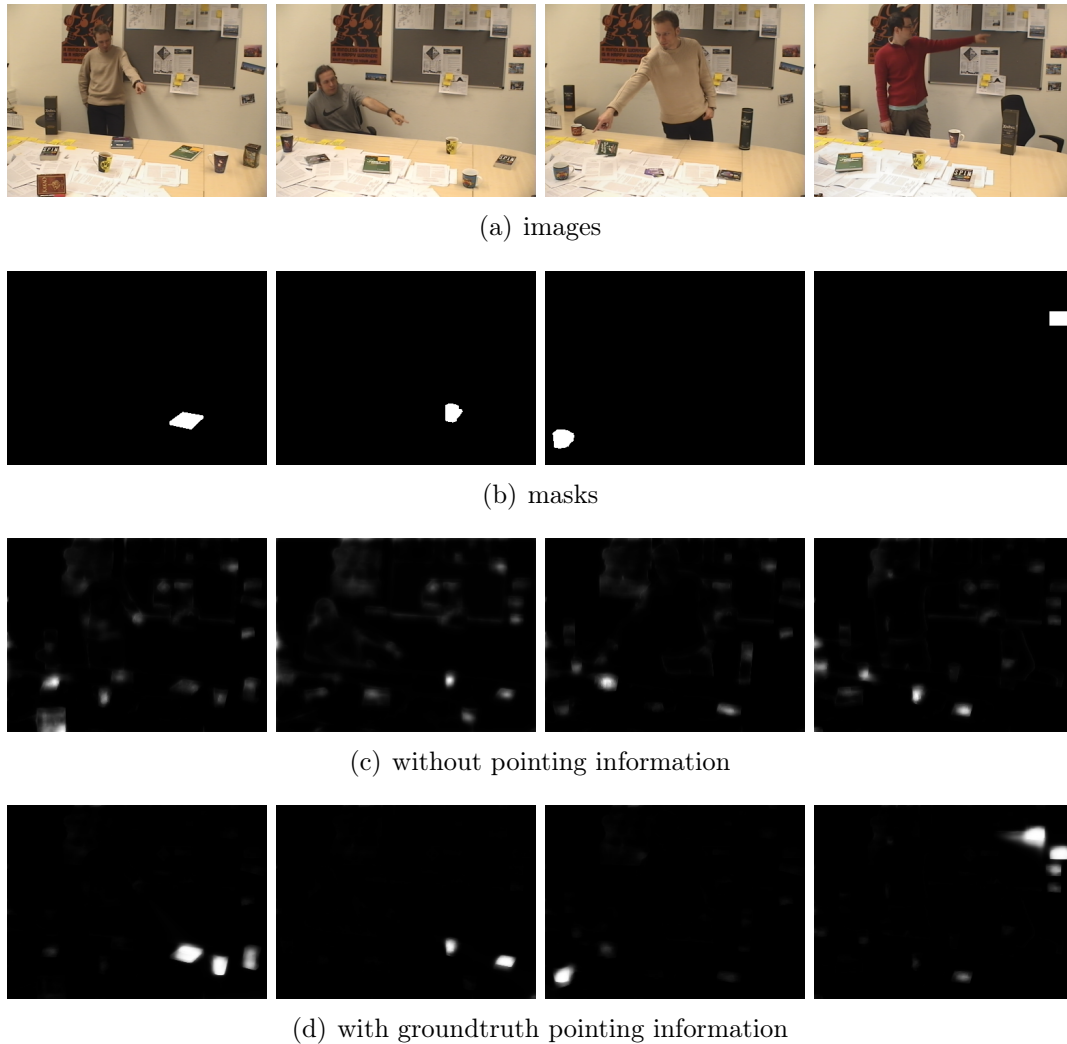


Figure 4.10.: Examples of pointing gestures performed in the PointAT dataset and CRF target region predictions.

to double the available image data. The CRF is trained with a grid resolution of  $381 \times 284$  and a 4-connected neighborhood.

**Results** As can be seen in Tab. 4.3, CRFs provide a better predictive performance than the heuristic baseline method (see Sec. 4.3.1.B). Most interestingly, the CRF that we trained and tested with automatic gesture detections is able to outperform the heuristic method even if the latter relies on groundtruth pointing information. This shows that the CRF model is better capable to compensate for the inaccuracy of automatically detected pointing gestures. Accordingly, if we compare the performance of both methods with detected and annotated pointing information, we can see that the relative performance difference of the heuristic model is much higher than the performance difference for the CRF.

Furthermore, we can see that LDRC in addition to our spectral features helps us to improve the results of the overall approach.

To serve as a baseline, we asked human observers to guess the pointed-at object and they were able to estimate the correct object for about 87% of the images [SRF10]. Accordingly, we can see that our model is able to come close to this baseline in terms of PHR. However, we can also see that the FHR is in fact higher than those 87%. How can that be? Most importantly, in ambiguous situations in which two potential target objects stand close to each other, the predicted target object location tends to be between both objects or just on the point of a border of one object that is closest to the other object. Thus, the target object might not have been selected by the most salient pixel, but at least a part of it is visible in the assumed FoA.

### 4.3.2. Language

Although pointing gestures are an important aspect of natural communication, there exist ambiguous situations in which it is not possible to identify the correct target object based on the pointing gesture alone. Instead, we commonly require additional knowledge or a shared context with our interaction partner to make the right assumptions about the referent. Although the combined use of gestures and language depends on the referring persons [Piw07], linguistic and gestural references can be seen to form composite signals, *i.e.* as one signal becomes more ambiguous the speaker will less rely on it and compensate with the other (see, *e.g.*, [Ban04, BC98, GRK07, KLP<sup>+</sup>06, LB05, Piw07, SKI<sup>+</sup>07]). Accordingly, we now want to go a step further and not just react to pointing gestures, but also integrate spoken information about a target object’s visual appearance. For this purpose, we have to make it possible to use the available information to guide the attention and highlight the intended target object.

#### A. Language Processing

To automatically determine spoken target information, we have to process and analyze the spoken utterance to determine references to object attributes or known objects. Before we proceed, please let us note that our intention is not to implement perfect language processing capabilities. Instead, we only want to assess how automatic – and thus sometimes faulty – extraction of target object information can influence our model’s performance.

**Target information** Language often provides the discriminating context to identify the referent amidst other potential target objects. Most importantly, it is used to specify objects (*e.g.*, “my Ardbeg whisky package”), classes (*e.g.*, “whisky package”), visually deducible attributes (*e.g.*, “red”, or “big”), and/or relations (*e.g.*, “the cup on that table”). When directly referring to an object,

this information is encoded in noun-phrases as pre-modifiers, if placed before the head-noun, and/or as post-modifiers after the head-noun [BC98].

We focus on noun-phrases with adjectives and nouns acting as pre-modifiers (*e.g.*, “the *yellow* cup” and “the *office* door”, respectively). We do not address verb phrases acting as pre-modifiers (*e.g.*, “the *swiftly opening* door”), because these refer to activities or events which cannot be handled by our attention models. Furthermore, to avoid in-depth semantic analysis, we ignore post-modifiers which typically are formed by clauses and preposition phrases in noun phrases (*e.g.*, “the author *whose* paper is reviewed” and “the cup *on* the table”, respectively).

**Parsing** We determine the noun-phrases and their constituents with a shallow parser which is based on regular expressions and was tested on the CoNLL-2000 Corpus [TB00]. Therefore, we trained a Brill tagger, which is backed-off by an n-gram and regular expression tagger, on the Brown corpus [Fra79].

**Extraction** Once we have identified the referring noun-phrase and its constituents, we determine the linguistic descriptions that can influence our attention model. First, we match the adjectives against a set of known attributes and their respective linguistic descriptions. Here, we focus on the 11 English basic color terms [BK69]<sup>8</sup>. Furthermore, we try to identify references to known object entities. Therefore, we match the object specification (consisting of the pre-modifiers and the head-noun) with a database that stores known object entities and their (exemplary) specifications or names, respectively. We also include adjectives in this matching process, because otherwise semantic analysis is required to handle ambiguous expressions (*e.g.*, “the *Intelligent* Systems Book” or “the *Red* Bull Can”). However, usually either attributes or exact object specification are used, because their combined use is redundant. A major difficulty is that the use of object specifiers varies depending on the user, the conversational context, and the environment. Thus, we have to regard partial specifier matches, *e.g.* “the Hobbits” equals “the Hobbits cookies package”. Obviously, the interpretation of these references depends on the shared conversational context. Given a set of known, possible, or plausible objects (depending on the degree of available knowledge), we can treat this problem with string and tree matching methods by interpreting each specifier as node in a tree. Consequently, we use an edit distance to measure the similarity, see [EIV07], and apply a modified version of the Levenshtein distance that is normalized by the number of directly matching words. Then, we determine the best matching nodes in the tree of known specifications. An object reference is detected, if all nodes in the subtree defined by the best matching node belong to the same object and there do not exist multiple nodes with equal minimum distance that belong to different objects.

---

<sup>8</sup>Please note that we can easily train color term models for other color term sets (*e.g.*, to integrate additional color terms or to work with different languages) [SF10b, SS12a].



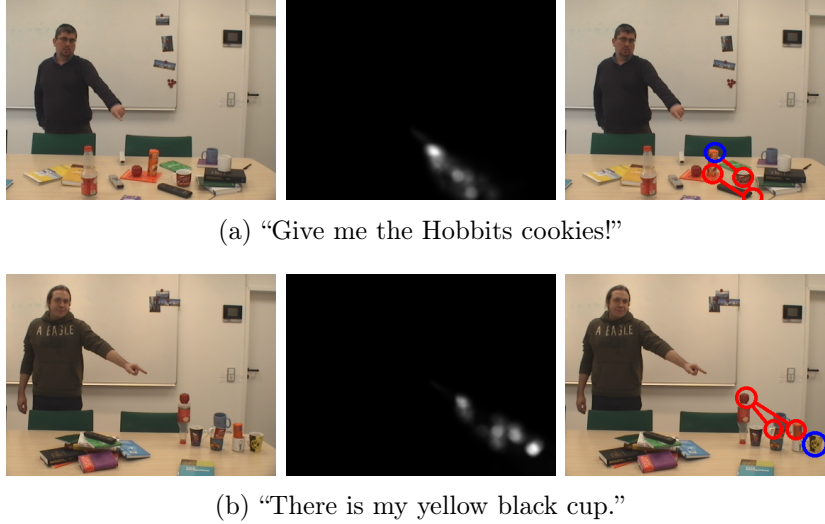


Figure 4.11.: Example of neuron-based top-down modulation, left-to-right: the images, their multimodal saliency maps, and the FoA shifts (the initial FoA is marked blue). The presented approach reflects how pointing gestures and verbal object references guide the perceptual focus of attention toward the referred-to object of interest.

## B. Neuron-based Saliency Model

To integrate visual saliency, pointing gestures, and spoken target object descriptions, Schauerte and Fink introduced a top-down modulatable saliency model [SF10a]. The model combines Navalpakkam’s idea of a modulatable neuron-based model [NI07], which itself is based on ideas of Wolfe *et al.*’s GSM (see Sec. 2.1.1), with the use of spectral saliency (see Sec. 3.2.1) to calculate the contrast of each neuron’s response, see Fig. 4.11. In this model, each feature dimension  $j$  – *e.g.* color, orientation, and lightness – is encoded by a population of  $N_j$  neurons with overlapping Gaussian tuning curves (cf. Fig. 2.2) and for each neuron  $n_{ij}$  a multi-scale saliency map  $s_{ij}$  is calculated. Therefore, we calculate the response  $n_{ij}(I^m)$  of each neuron for each scale  $m$  of the input image  $I$  and use spectral whitening, see Sec. 3.2.1, to calculate the feature maps

$$s_{ij}^m = g * \mathcal{F}^{-1} \left\{ e^{i\Phi(\mathcal{F}\{n_{ij}(I^m)\})} \right\} \quad (4.26)$$

with the Fourier-Transform  $\mathcal{F}$ , the Phase-Spectrum  $\Phi$ , and an additional 2D Gaussian filter  $g$ . Then, we normalize these single-scale feature maps and use a convex combination in order to obtain the cross-scale saliency map  $s_{ij}$

$$s_{ij} = \sum_{m \in M} w_{ij}^m \mathcal{N}(s_{ij}^m) \quad (4.27)$$

with the weights  $w_{ij}^m$  and the normalization operator  $\mathcal{N}$ . The latter performs a cross-scale normalization of the feature map range, attenuates salient activation

spots that are caused by local minima of  $n_{ij}(I^m)$ , and finally amplifies feature maps with prominent activation spots (*cf.* [IK01b]). However, since we do not incorporate knowledge about the size of the target, we define the weights  $w_{ij}^m$  as being uniform, *i.e.*

$$\sum_{m \in \mathcal{M}} w_{ij}^m = 1 . \quad (4.28)$$

The multi-scale saliency maps  $s_{ij}$  of each individual neuron are then combined to obtain the conspicuity maps  $s_j$  and the final saliency map  $S_B$

$$s_j = \sum_{i=1}^{N_j} w_{ij} s_{ij} \quad \text{and} \quad S_B = \sum_{i=1}^N w_j s_j , \quad (4.29)$$

given the weights  $w_j$  and  $w_{ij}$ .

These weights are chosen in order to maximize the signal-to-noise ratio (SNR) between the expected target and distractor saliency ( $S_T$  and  $S_D$ )

$$\text{SNR} = \frac{\mathbb{E}_{\theta \| T}[S_T]}{\mathbb{E}_{\theta \| D}[S_D]} , \quad (4.30)$$

given known models of the target and distractor features (*i.e.*,  $\theta \| T$  and  $\theta \| D$ ). Therefore, we need to predict the SNR for each neuron (*i.e.*,  $\text{SNR}_{ij}$ ) and feature dimension (*i.e.*,  $\text{SNR}_j$ ) to obtain the optimal weights  $w_{ij}$  and  $w_j$  according to

$$w_{ij} = \frac{\text{SNR}_{ij}}{\frac{1}{n} \sum_{k=1}^n \text{SNR}_{kj}} \quad \text{and} \quad w_j = \frac{\text{SNR}_j}{\frac{1}{N} \sum_{k=1}^N \text{SNR}_k} , \quad (4.31)$$

as has been proposed by Navalpakkam and Itti [NI06].

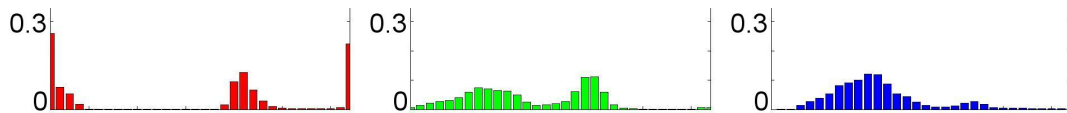
The SNR calculation is critical for this model, especially since we aim at using general models for saliency modulation that can also be applied for recognition and naming of objects. This stands in contrast to most prior art, in which saliency modulation was directly learned from target image samples (*e.g.* [Fri06, IK01b, NI07]; *cf.* [FRC10]). In our implementation, we use probabilistic target and distractor feature models (*i.e.*,  $p(\theta \| T)$  and  $p(\theta \| D)$ , respectively) and calculate  $\text{SNR}_{ij}$  and  $\text{SNR}_j$  according to

$$\text{SNR}_{ij} = \left[ \frac{\mathbb{E}_{\theta \| T, I}[s_{ij}]}{\mathbb{E}_{\theta \| D, I}[s_{ij}]} \right]^\alpha \quad \text{and} \quad \text{SNR}_j = \frac{\mathbb{E}_{\theta \| T, I}[s_j]}{\mathbb{E}_{\theta \| D, I}[s_j]} , \quad (4.32)$$

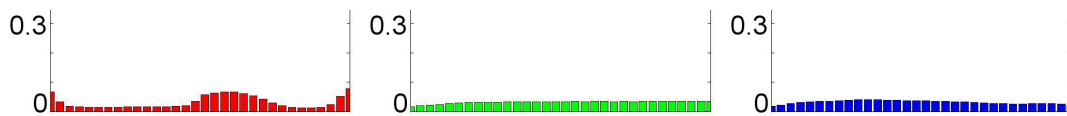
where  $\mathbb{E}_{\theta \| T, I}[s_{ij}]$  and  $\mathbb{E}_{\theta \| D, I}[s_{ij}]$  is the expected saliency, according to the calculated neuron saliency map  $s_{ij}$ , of the respective feature model in image  $I$ . Here the constant exponent  $\alpha$  is an additional parameter that influences the modulation strength. This is especially useful when dealing with smooth feature models, *e.g.* color term models (see, *e.g.*, Fig. 4.12), to force a stronger modulation. Analogously,  $\mathbb{E}_{\theta \| T, I}[s_j]$  and  $\mathbb{E}_{\theta \| D, I}[s_j]$  is the expected target and distractor saliency, respectively, for the calculated conspicuity map  $s_j$ , see Eq. 4.29.



(a) Known target object’s model view and its segmentation mask obtained with color spatial variance.



(b) Marginal distributions of the target object’s HSL color model.



(c) Uniform combination of the “red” and “blue” color term models.

Figure 4.12.: An example of the automatic acquisition of a target object’s color models from a model view, see (a) and (b). For comparison, a combined color term target model for “red” and “blue”, see (c).

### C. Target Information and Models

Since our neuron-based saliency model requires to calculate an expected SNR to highlight the target object, we require the target feature information in a probabilistic model. In the following, we explain our three sources of target feature information. First, target color terms (*e.g.*, “red”) that we learn from web images. Second, target objects that are generated from images in a target object database, *i.e.* a set of known objects. Third, information about distracting features such as the color of the background or potential clutter.

**Colors** The color models  $p(\theta||T_{\text{color}})$ , see *e.g.* Fig. 4.12, are learned using the Google-512 dataset [SF10b], which was gathered from the Internet for the 11 English basic color terms (see Sec. 4.1.1.B). Therefore, we use probabilistic latent semantic analysis with a global background topic and a probabilistic HSL color model [SF10b]. The latter reflects the different characteristics of real-world images and images retrieved from the Internet. Here, we use HSL as color space, because the color channels are decoupled and thus support the use of independent neurons for each channel. However, since color term models are as general as possible, we can in general not expect as strong modulation gains as with specific target object models.

**Objects** If we have access to an image of a target object (*e.g.*, the close-up object views that are part of the ReferAT dataset’s object database), we can calculate object-specific target feature models  $p(\theta|T_{\text{obj}})$ . For this purpose, we exploit that the target objects are usually well-centered in the model views and use the color spatial variance – *i.e.*, a known salient object detection feature [LSZ<sup>+</sup>07] – to perform a foreground/background separation, see Fig. 4.12. Additionally, the acquired segmentation mask is dilated to suppress noise and omit background pixels around the object boundaries. Then, we calculate  $p(\theta|T_{\text{obj}})$  as the feature distribution of the foreground image pixels. If we have access to multiple views of the same object, we use a uniform combination to combine the models.

**Distractors** In the absence of a pointing gesture, the model of distracting objects and background of each image  $p(\theta|D_I)$  is estimated using the feature distribution of the whole image. Thus, we roughly approximate a background distribution and favor objects with infrequent features. In the presence of a pointing gesture, it is beneficial to incorporate that pointing gestures narrow the spatial domain in which the target object is to be expected. Consequently, we focus the calculation of the distractor feature distribution  $p(\theta|D_I)$  on the spatial region that was indicated by the pointing gesture. Therefore, we calculate a probabilistic map – similarly to the pointing cone, see Eq. 4.12, but with an increased variance  $\sigma_c^2$  – to weight the histogram entries when calculating the feature distribution  $p(\theta|D_I)$ . However, since in both cases the target object is part of the considered spatial domain, the distractor feature models are smoothed to avoid suppressing useful target features during the modulation.

#### D. Conditional Random Field

We rely on the same conditional random field structure that has been explained in Sec. 4.3.1.C and just adapt the features. Here, we rely on the same target information as the neuron-based approach, see Sec. 4.3.2.C.

**Features** One of the major advantages of our machine learning based approach with CRFs is that it is very simple to include additional features and thus target information. To incorporate information about the target object’s appearance, we rely on the probabilistic target models that we introduced for the neuron-based approach (see Sec. 4.3.2.C). Therefore, after scaling the image to the CRF’s grid size, we calculate the target probability, *i.e.*,  $p(\theta|T_{\text{color}})$  or  $p(\theta|T_{\text{obj}})$ , at each image pixel and append the probability to the unary feature vector at the corresponding CRF grid location.

#### E. The ReferAT Dataset

To evaluate how well multimodal – here, pointing gestures and spoken language – references guide computational attention models, we collected a dataset which



Figure 4.13.: Representative object references in our ReferAT evaluation dataset. Depicted are some exemplary images with the corresponding binary target object masks that we can generate based on the annotated target object boundaries.

contains 242 multimodal referring acts that were performed by 5 persons referring-to a set of 28 objects in a meeting room [SF10a], see Fig. 4.13. This limited set of objects defines a shared context of objects that are plausible in the scene and can be addressed. We chose the objects from a limited set of classes (most importantly: books, cups, packages, and office utensils) with similar intra-class attributes, *i.e.* size and shape. Thus, in most situations, object names and colors are the most discriminant verbal cues for referring-to the referent. Please note that the limited number of classes further forces the participant to use specifiers to address the objects, because the object class information alone would often lead to ambiguous references.

**Hardware Setup** The hardware setup is identical to the PointAT dataset’s hardware setup, see Sec. 4.3.1.D. The dataset was recorded using a monocular Sony EVI-D70P pan-tilt-zoom camera. The camera provides images in approximately PAL resolution ( $762 \times 568$  px), and offers an optical zoom of up to  $\times 18$ . Its wide horizontal opening angle is  $48^\circ$ . The camera was mounted at eye height of an averagely tall human to reflect a human-like point of view [MFOF08].

**Procedure** We intended to obtain a challenging dataset. Thus, we allowed the participants at every moment to freely change their own position as well as select and arrange the objects that are visible in the scene, see Fig. 4.13. Furthermore, after we explained that our goal is to identify the referent, we even encouraged them to create complex situations. However, naturally the limited field of view of the camera limits the spatial domain, because we did not allow references to objects outside the field of view. Furthermore, we asked the participants to point with their arms extruded, because we use the line-of-sight to estimate pointing direction [RPF08, Ban04] and do not evaluate different methods to determine the pointing direction (*cf.* [NS07]). In order to verbally refer to an object, the participants were allowed to use arbitrary sentences. But, since the participants often addressed the object directly, in some cases only a noun phrase was used in order to verbally specify the referent.

We manually transcribed the occurring linguistic references to avoid the influence of speech recognition errors. Furthermore, we annotated the dominant eye, the pointing finger, and the resulting pointing direction. Accordingly, we are able to assess the quality of the automatically recognized pointing gesture and its influence on the detection of the referent. Additionally, for each linguistic reference, we annotated the attributes, target object, and whether the specific target object can be recognized without the visual context of the complementary pointing gesture (*e.g.* “the cup” vs. “the Christmas elk cup”).

**Object Database** To make it possible to highlight a known target object as well as to enable object recognition, we collected a database that contains images of all objects that have been referenced in the dataset. For this purpose,



Figure 4.14.: Exemplary acquisition of an object model for the ReferAT dataset’s object database. The trainer points to an objects that is then automatically identified (a). Then, the system automatically estimates the target object’s size and uses the camera’s zoom functionality to obtain a close-up image (b).

we used our automatic pointing reference resolution system in a learning mode (see Sec. 4.3.1.B). We placed each object that had to be learned at a position where the pointing reference was unambiguous. Then, we referred to the object via a pointing gesture and a verbal specification. Our system then used the pointing gesture to identify the referred-to object, applied segmentation based on maximally stable extremal regions (MSER) [MCUP04] to roughly estimate the object boundaries, and zoom toward the object to obtain a close-up view for learning. These close-up views acquired are stored in a database, in which they are linked with the verbal specification.

**Dataset Properties** On average the target object occupies 0.70% of the image area. Naturally, due to the experimental environment, the target object locations are concentrated in the lower half of the image, *i.e.* the table area. The average differences between the annotated and automatically determined pointing origin and direction are 12.50 px and 3.92°, respectively. Again, the former is caused by the fact that the system is based on the hand’s center and not the finger. The latter, again, is due to the fact that the eye positions are estimated given the head-shoulder detection. In some cases, the ray that originates from the finger tip (*i.e.*, pointing origin) and follows the pointed direction does not intersect the target object’s boundaries and misses the object. The rate of how often the object’s annotated boundary polygon is missed by the pointing ray is 7.85% for the annotated pointing information and 26.03% for the automatic detection.

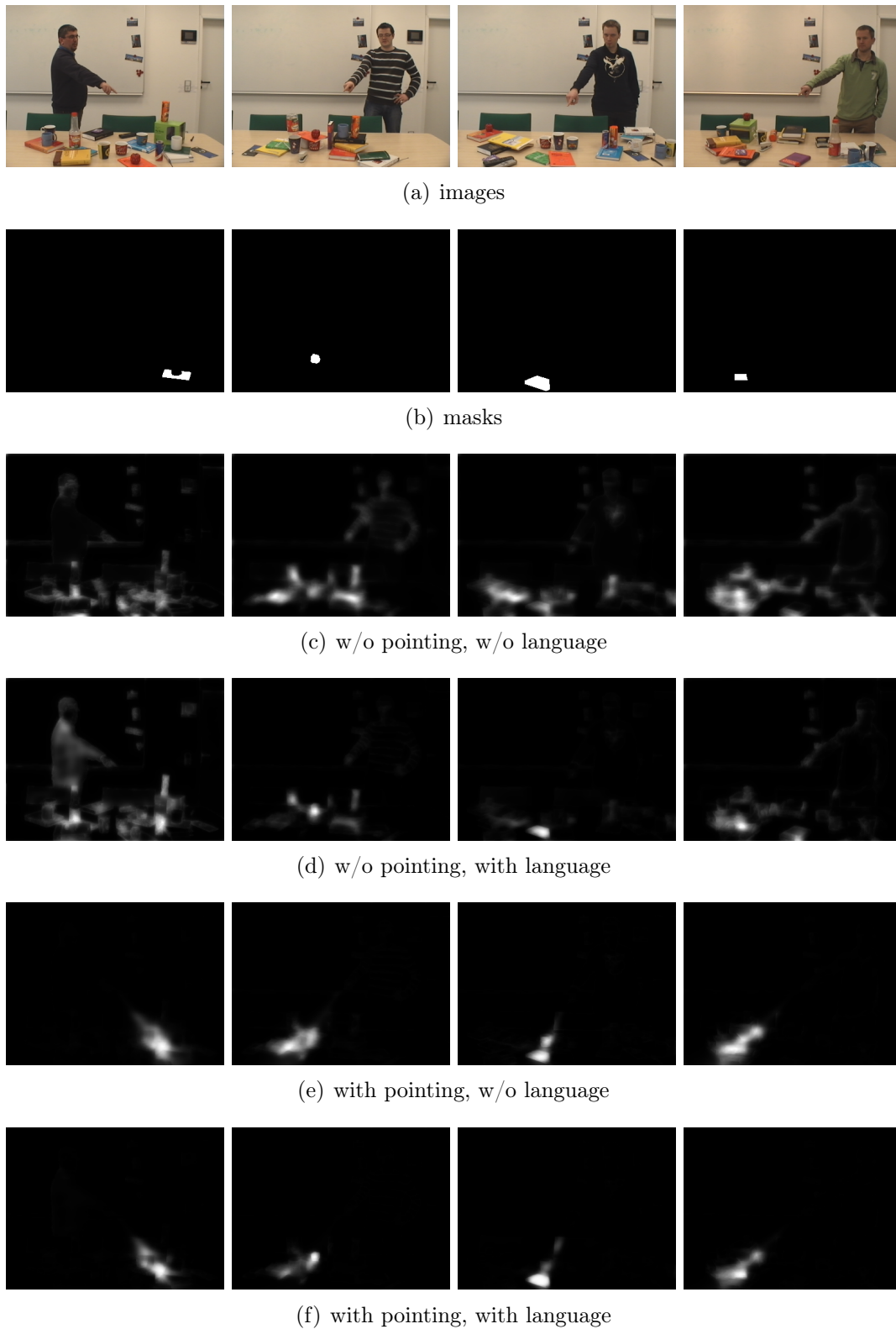


Figure 4.15.: Representative object references in the ReferAT dataset and CRF target region predictions (with groundtruth pointing information and LDRC as additional feature).



Algorithm	No language			
	PHR	FHR	$\int$ FHR	NTOS
	no pointing			
Neuron-based	–	9.90%	–	–
CRF, no LDRC	6.61%	18.60%	0.307	1.767
CRF, w/ LDRC	5.37%	19.83%	0.339	1.808
	pointing detected			
Neuron-based	–	46.30%	–	–
CRF, no LDRC	28.51%	59.92%	0.708	4.565
CRF, w/ LDRC	27.69%	64.05%	0.722	4.627
	pointing annotated			
Neuron-based	–	51.20%	–	–
CRF, no LDRC	33.47%	69.42%	0.769	5.511
CRF, w/ LDRC	33.06%	69.83%	0.779	5.535

Table 4.4.: Target object detection performance on the ReferAT dataset without spoken target object information. A description of the evaluation measures can be found in Sec. 4.3.1.E.

## F. Evaluation Results

**Procedure and parameters** To train and evaluate our models, we use a leave-one-person-out training procedure. The CRF is trained with a grid resolution of  $381 \times 284$  and a 4-connected neighborhood.

As a reference for the neuron-based model, which we have described in detail in Sec. 4.3.2.B, we use the results reported by Schauerte and Fink [SF10a]. Since the performance of the CRF is a substantially better than the neuron-based model (the CRF’s PHR is often higher than the neuron-based model’s FHR), we refrain from (re-)evaluating the neuron-based model to calculate PHR and  $\int$ FHR. Schauerte and Fink’s model was based on the hue, saturation, lightness, and orientation as feature dimensions. Each feature dimension had a sparse population of 8 neurons. The SNR exponent  $\alpha$  was set to 2.

The language processing correctly detected 123 of 123 color references and 123 of 143 references to specific objects (*e.g.*, as negative example: “the tasty Hobbits” as reference to the “the Hobbits cookies package” was not detected; for comparison, as non-trivial positive matching samples, “valensina juice bottle” and “ambient intelligence algorithms book” have been matched to “valensina orange juice package” and “algorithms in ambient intelligence book” in the data base, respectively). Most importantly, the specifier matching of object descriptions made only one critical mismatch (“the statistical elements book” has been matched to “the statistical learning book” instead of “the elements of statistical learning book”). Since wrong target information will lead us to

Algorithm	Detected language			
	PHR	FHR	$\int$ FHR	NTOS
	no pointing			
Neuron-based	–	15.70%	–	–
CRF, no LDRC	24.38%	33.88%	0.467	3.270
CRF, w/ LDRC	24.79%	37.19%	0.463	3.318
	pointing detected			
Neuron-based	–	50.00%	–	–
CRF, no LDRC	55.37%	72.73%	0.783	6.551
CRF, w/ LDRC	52.48%	75.21%	0.801	6.497
	pointing annotated			
Neuron-based	–	63.20%	–	–
CRF, no LDRC	66.12%	80.17%	0.830	7.377
CRF, w/ LDRC	65.29%	81.41%	0.834	7.355

Table 4.5.: Target object detection performance on the ReferAT dataset with automatically determined spoken target object information. A description of the evaluation measures can be found in Sec. 4.3.1.E.

highlight the wrong image areas, this is an important aspect and the reason why we chose such a cautious matching method as described in Sec. 4.3.2.A.

**Results** We present the results for different target information conditions in separate tables. Tab. 4.4 shows the results without any linguistic target information, Tab. 4.5 contains the results obtained with automatically determined target object information, and Tab. 4.6 provides the results achieved with groundtruth target information. Each table presents the results achieved without pointing information, with detected pointing information, and with groundtruth pointing information.

First of all, we can notice that the CRFs accurately segmented the target object in most “simple” object arrangements such as in the PointAT dataset, see Fig. 4.10, but they can only predict rough salient regions of interest in complex situations, see Fig. 4.15.

If we use FHR as key evaluation measure – as has been done by Schauerte and Fink [SF10a] –, then the integration of LDRC as a CRF feature clearly improves the results. However, in contrast, the performance as quantified by PHR decreases when LDRC is integrated, at least if pointing information is integrated. This stands in contrast to our results on PointAT and can most likely be explained by the fact that LDRC seems to be unable to highlight a single object in a dense cluster of distractors, which is not surprising given its definition. But, it often highlights small clusters of spatially close objects in which the target object is contained, which increases in FHR while it decreases PHR.

Algorithm	Groundtruth language			
	PHR	FHR	$\int$ FHR	NTOS
	no pointing			
Neuron-based	–	16.50%	–	–
CRF, no LDRC	19.83%	30.58%	0.424	2.951
CRF, w/ LDRC	19.42%	34.71%	0.452	3.022
	pointing detected			
Neuron-based	–	54.10%	–	–
CRF, no LDRC	54.13%	72.31%	0.780	6.351
CRF, w/ LDRC	47.93%	73.55%	0.794	6.181
	pointing annotated			
Neuron-based	–	59.90%	–	–
CRF, no LDRC	63.63%	82.23%	0.837	7.206
CRF, w/ LDRC	60.74%	84.71%	0.842	7.069

Table 4.6.: Target object detection performance on the ReferAT dataset with groundtruth spoken target object information. A description of the evaluation measures can be found in Sec. 4.3.1.E.

In any case, CRFs clearly outperform the neuron-based model, often by more than a 20% higher FHR. In fact, the performance of the CRFs with detected pointing information often even outperform the performance of the neuron-based model with groundtruth pointing information, although the use of detected pointing information leads to a substantial drop in the overall performance. The performance difference that is caused by the use groundtruth and detected pointing information can be explained by the imprecision of the detected pointing origin and pointing direction – see the dataset property discussion in Sec. 4.3.2.E –, which often causes the pointing ray to miss the intended target object.

Since pointing gestures substantially limit the spatial area in which we expect target objects, it is intuitively clear that the integration of pointing gestures substantially improves the performance under all three target information conditions (*i.e.*, no language, automatically extracted spoken target object information, and annotated target information, see Tab. 4.4, 4.5, and 4.6, respectively; compare “no pointing” to “pointing detected” and “pointing annotated”).

The integration of language also substantially improves the performance on its own, *i.e.* without accompanying pointing gestures (compare Tab. 4.4 to Tab. 4.5 and 4.6). Here, we have to differentiate between our two types of target information, *i.e.* the knowledge of the target object itself or just a color description. To further investigate this aspect, we labeled the color attributes for all objects in the ReferAT dataset. This way, we can train and test CRFs that are always given the correct target object or the correct target object’s color term models. As we can see in Tab. 4.7, if we use the exact target object model, we achieve a better performance compared to the object’s color attribute

(a) target object				
Algorithm	<b>Groundtruth object</b>			
	PHR	FHR	$\int$ FHR	NTOS
	no pointing			
CRF, no LDRC	25.62%	39.26%	0.480	3.265
CRF, w/ LDRC	24.79%	39.67%	0.477	3.317
	pointing annotated			
CRF, no LDRC	71.49%	88.84%	0.860	7.614
CRF, w/ LDRC	70.66%	90.50%	0.864	7.490
(b) target color				
Algorithm	<b>Groundtruth color</b>			
	PHR	FHR	$\int$ FHR	NTOS
	no pointing			
CRF, no LDRC	21.07%	31.82%	0.425	3.039
CRF, w/ LDRC	19.42%	34.30%	0.436	3.043
	pointing annotated			
CRF, no LDRC	60.33%	83.88%	0.844	6.997
CRF, w/ LDRC	59.09%	82.23%	0.838	6.930

Table 4.7.: Target object detection performance on the ReferAT dataset given the groundtruth target object model or the appropriate color term model, see Fig. 4.12. A description of the evaluation measures can be found in Sec. 4.3.1.E.

description. This could have been expected, because the color term models are general and not as specific and discriminative as the visual target object models, see Fig. 4.12. Nevertheless, target color term models can guide the attention and lead to substantially better results than models without any verbal target object information (compare Tab. 4.7 to Tab. 4.4).

Finally, the combination of both modalities leads to further improvements compared to each unimodal result (see Tab. 4.4, Tab. 4.5, Tab. 4.6, and Tab. 4.7). This observation confirms that the guidance provided by the individual modalities, *i.e.* pointing gestures and verbal descriptions, complement each other. This could be expected, because in our scenario both modalities provide different types of information: The pointing gestures guide the attention toward spatial areas of interest, along their spatial corridor of attention. The verbal description provide information about the target object’s visual appearance that help to discriminate the object from the background and surrounding clutter.

## 4.4 Gaze Following in Web Images

In the previous Sec. 4.3, we have focused on a typical HRI task with an evaluation in a laboratory environment. In the final technical section of this thesis, we will address a topic that we are not even close to solving. However, it represents the logical consolidation of the work that we presented in previous sections. We will show how we can use the methods that we developed to interpret attentional signals in HRI to implement gaze following in web images. In other words, we try to identify the object that is being looked at in images that have been composed by human photographers. Accordingly, we transfer and combine our approaches from previous sections and use conditional random fields with features such as, most importantly, a probabilistic corridor of interest that encodes the gaze direction, spectral saliency detection, and locally debiased region contrast saliency with an explicit center bias.

Internet image collections and datasets pose many challenges compared to data that is acquired in controlled laboratory environments. For example, we have to cope with an extreme variety in the depicted objects and environments, image compositions, and lighting conditions, see Fig. 4.16. The challenges that arise with such data are also responsible for the fact that even today there does not exist a computer vision algorithm that is able to reliably estimate a gaze direction on the dataset that we present in this section. However, especially the variety of target objects makes it particularly interesting to test our concepts on web images, because our saliency-based approach does not require hundreds or thousands of object detectors to detect all kinds of objects. Furthermore, we are convinced that any approach that we develop on this particularly challenging data



Figure 4.16.: Example Gaze@Flickr images to illustrate the complexity and variance in the dataset.

will perform even better in simpler scenarios. For example, depth data would undoubtedly assist the identification and segmentation of the target objects, and information about the type and location of potential target objects could serve as valuable prior.

### 4.4.1. Approach

In principle, we rely on the same methodology as for pointing gestures, see Sec. 4.3.1.

#### A. Spatial Gaze Target Probability

In principle, gaze serves the same purpose as pointing, *i.e.* to direct the attention to certain parts of the image. Analogue to Eq. 4.12, we represent the observed gaze direction's attention corridor as

$$p_G(x) = p(\alpha(x, o)|d, o) = \mathcal{N}(0, \sigma_c^2) . \quad (4.33)$$

Here,  $\alpha(x, o)$  is the angle between the vector from the eyes  $o$  to the image point  $x$  given the gaze direction  $d$ , and  $\sigma$  encodes the assumed gaze direction inaccuracy or uncertainty. This equation represents the probability  $p_G(x)$  that the object at point  $x$  in the image is being looked-at and defines our probabilistic corridor of attention.

#### B. Heuristic Integration

Again, see Sec. 4.3.1.B, we implement a heuristic approach to serve as a baseline for our CRFs. Given  $p_G(x)$ , see Eq. 4.33, we calculate the top-down gaze map

$$S_t(a, b) = p_G(x) \quad (4.34)$$

with  $x = (a, b)$ . Then, we use QDCT to calculate the visual saliency map  $S_b$  based on the PCA decorrelated CIE Lab color space, see Sec. 3.2. The saliency map  $S_b$  is normalized to  $[0, 1]$ . The final heuristic saliency map  $S$  is defined as

$$S = S_b \circ S_t , \quad (4.35)$$

where  $\circ$  represents the Hadarmard product.

#### C. Conditional Random Field

The CRF relies on the same structure, learning method, and prediction method as in the previous section (Sec. 4.3).

As unary image-based features, we include the following information at each CRF grid point: First, we include each pixel's normalized horizontal and vertical image position in the feature vector. Second, we directly use the pixel's intensity

value after scaling the image to the CRF’s grid size. Third, we include the scaled probabilistic gaze cone, see Sec. 4.3.1.A. Then, after scaling each saliency map to the appropriate grid size, we append QDCT saliency maps based on the PCA decorrelated Lab color space (see Sec. 3.2.1) at three scales:  $96 \times 64$  px,  $168 \times 128$  px, and  $256 \times 192$  px. Furthermore, we either include the RC’10, LDRC, or LDRC+CB saliency map, see Sec. 4.2.

Again, as CRF edge features, we use a 1-constant and 10 thresholds to encode the color difference of neighboring pixels. Then, we multiply the existing features by an indicator function for each edge type (*i.e.*, vertical and horizontal), which allows to parametrize vertical and horizontal edges separately.

#### 4.4.2. The Gaze@Flickr Dataset

To evaluate the ability to identify at which object a person is looking in web images, we collected a novel dataset that we call Gaze@Flickr. Our dataset itself is based on the MIRFLICKR-1M dataset<sup>9</sup> [HTL10], which consists of 1 million Flickr images under the Creative Commons license. We collected and annotated our dataset in several, subsequent steps:

1. First, we inspected all MIRFLICKR-1M images and selected the images that show at least one person who gazes at something<sup>10</sup>.
2. Then, we selected a subset of 1000 images, for which we outlined the heads/faces of up to three persons.
3. For each annotated head/face region, we annotated the gaze direction under two viewing conditions:
  - a) First, we annotated the gaze direction while being able to see the whole image. We call this the “full” condition.
  - b) Second, we annotated the gaze direction while only being shown the face/head region. For this purpose, the other parts of the image were shown as being black. We call this the “blank” condition.
4. Then, for each face/head with an annotated gaze direction, we annotated the boundary of the object at which the person is looking, see Fig. 4.17. In some cases the target object was either “ambiguous” (*i.e.*, the target was not visible or there were several equally plausible target objects) or – most likely – “outside” the image (*i.e.*, not depicted in the image). In both cases, we were unable to label the target object and instead just tagged the images accordingly.

---

<sup>9</sup><http://press.liacs.nl/mirflickr/>

<sup>10</sup>During this process, we also collected all images that depict persons pointing at something. However, we could not find a sufficient number of such images to build a “pointing gestures in the wild dataset”.

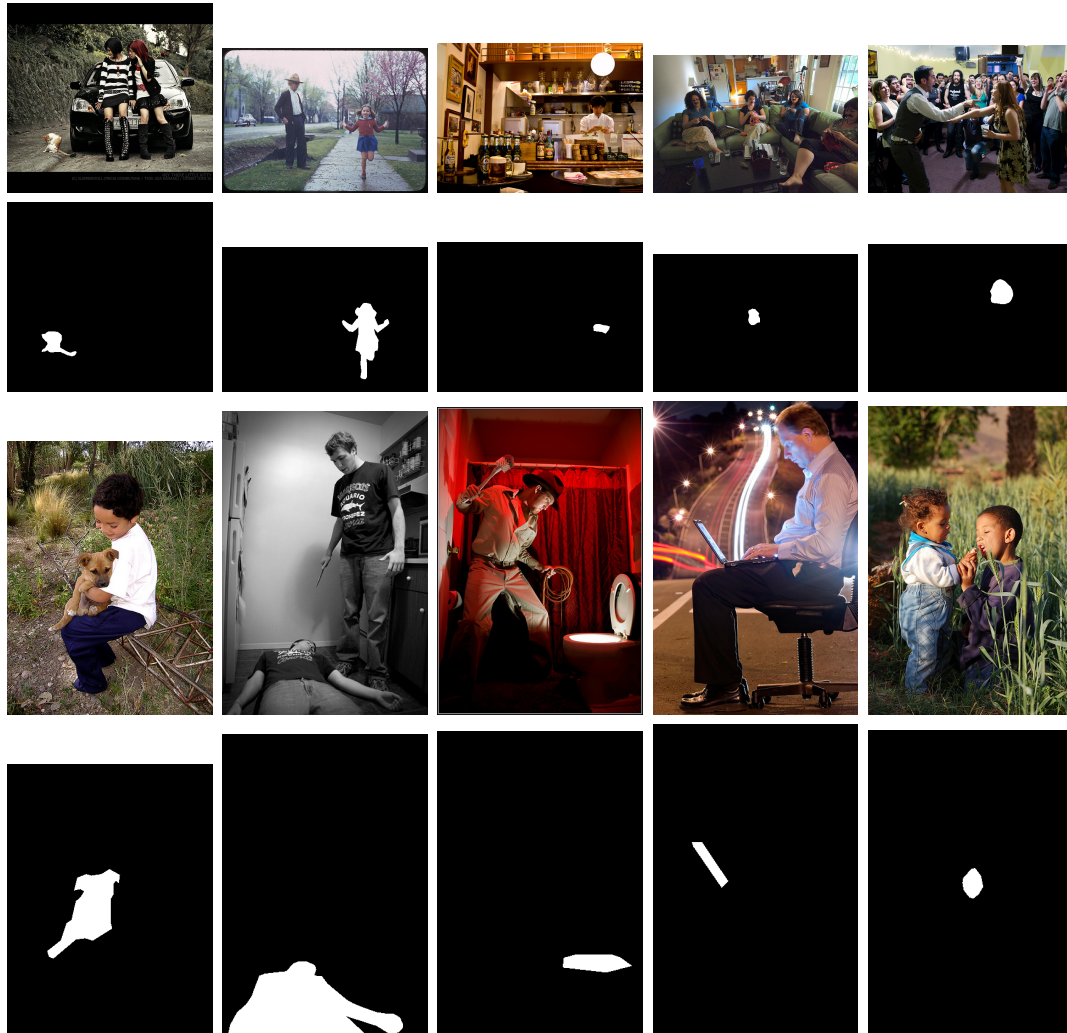


Figure 4.17.: Representative object references in our Gaze@Flickr dataset. Depicted are some exemplary images with their corresponding binary target object masks that we can generate based on the annotated target object boundaries.



In total, our dataset contains 863 images that depict 1221 annotated gaze references, excluding gaze samples with unclear or out-of-sight targets.

**Properties** On average the target object occupies 4.33% of the image area, which makes the objects substantially larger compared to the target objects in the PointAT (0.58%) and ReferAT (0.70%) datasets. However, at the same time, the objects are also considerably smaller compared to the MSRA dataset (18.25%). As could be expected, the gaze annotation viewing condition (*i.e.*, full or blank) influences the annotated gaze directions. The average difference between the annotated directions is  $12.10^\circ$ . Similar to pointing gestures, the ray that originates from the eyes (*i.e.*, gaze origin) and follows the gaze direction can miss the target object’s polygon. The rate of how often the object’s annotated target polygon is missed depends substantially on the viewing condition: The rays that were annotated under the full condition miss only 4.67% of the target objects, while the rays under the blank conditions miss 26.94% of the targets, *i.e.*  $5.76\times$  more often. This, in combination with the substantial deviation of annotated gaze directions, demonstrates the important influence that context information – *i.e.*, the information about potential target objects – can have on gaze estimates made by humans.

### 4.4.3. Evaluation

#### A. Procedure and Parameters

To train and evaluate our CRFs, we follow a 5-fold cross-validation procedure. Accordingly, we have about 976 training images and 244 test images for each fold. The CRF is trained with a grid resolution of  $300 \times 300$  and a 4-connected neighborhood.

We have to rely on the two annotated gaze directions in the evaluation. This is due to the fact that there does not exist a computer vision method that is able to reliably produce gaze estimates on our Gaze@Flickr dataset. Accordingly, since we can not estimate a gaze direction uncertainty or inaccuracy, we use a fixed probabilistic gaze cone  $\sigma$  of approximately  $14^\circ$ , *i.e.* 0.25 rad.

#### B. Measures

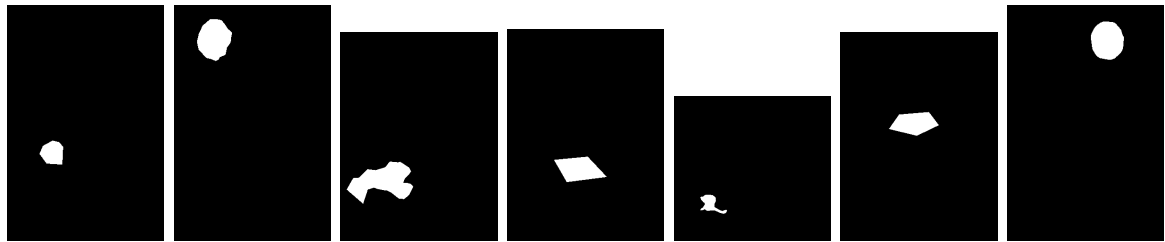
We use all evaluation measures for salient object detection and focus of attention selection that we have used in Sec. 4.2.3.B and Sec. 4.3.1.E, respectively.

#### C. Data Analysis

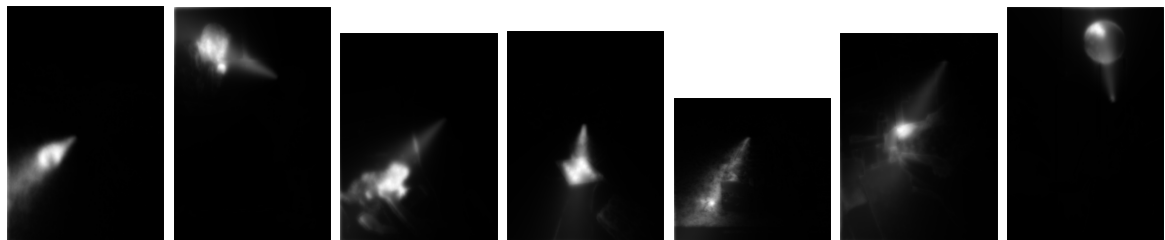
**Do people look at salient things?** An interesting question is whether or not the objects that people are looking at are already perceptually salient (the opposite question would be “Do people in images frequently look at objects



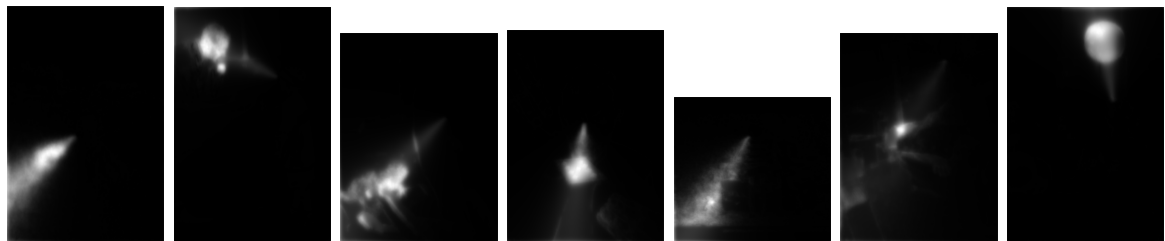
(a) images



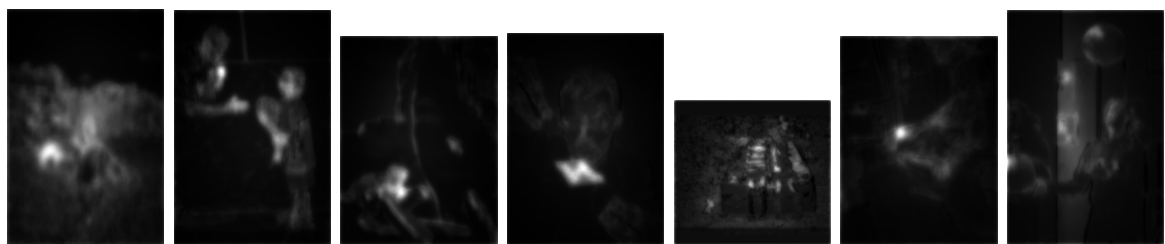
(b) masks



(c) CRF with QDCT & LDRC+CB



(d) CRF with QDCT & LDRC



(e) CRF with QDCT & LDRC, no gaze information

Figure 4.18.: Example predictions of our CRFs on the Gaze@Flickr dataset.

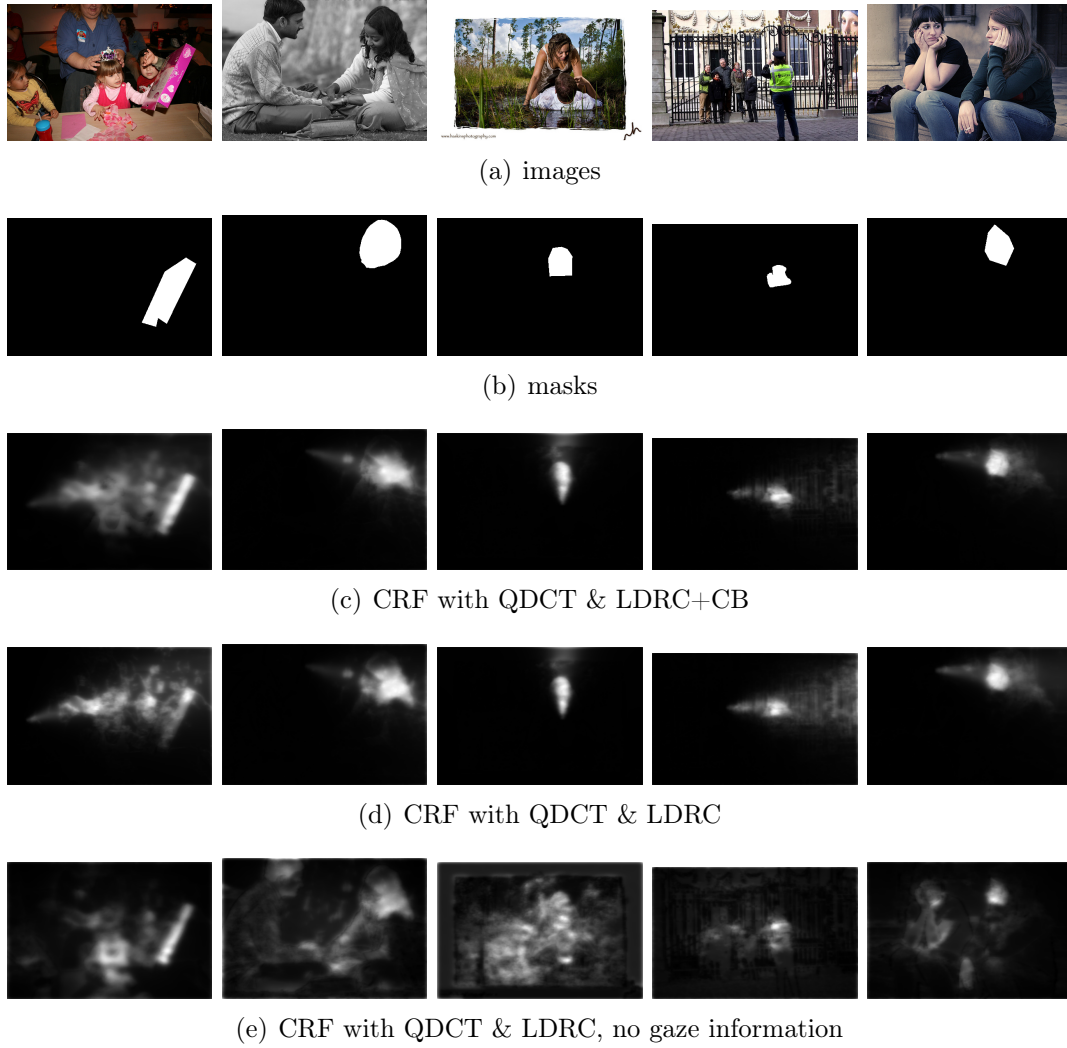


Figure 4.19.: Example predictions of our CRFs on the Gaze@Flicker dataset.

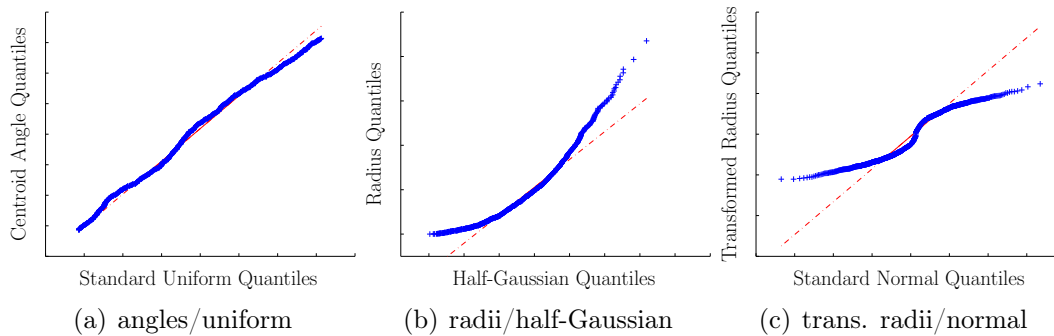


Figure 4.20.: Quantile-quantile (Q-Q) plots of the angles versus a uniform distribution (a), radii versus a half-Gaussian distribution (b), transformed radii (see Sec. 4.2.2.B) versus a normal distribution (c). Compare to the Q-Q plots on MSRA in Fig. 4.5.

that do not pop-out from the background?”). We can use the NTOS evaluation measure to investigate this question, see Sec. 4.3.1.E. NTOS compares the mean saliency at target object location to the mean saliency of the background. The measure is normalized in terms of the standard deviation of the saliency values. Thus,  $\text{NTOS} \leq 0$  means that the saliency at the target object’s location is not higher compared to the background, whereas an NTOS of 1 means that the saliency in the target object area is one standard deviation above average.

For this purpose, we calculated the visual saliency of several algorithms without gaze integration, see Tab. 4.8. Apparently, the NTOS is substantially higher than 0 for almost all evaluated visual saliency algorithms. This means that people in photographs often look at objects that are perceptually salient. It is in fact very interesting that visual saliency algorithms seem so capable to highlight the target objects, because it indicates that the images are composed in a way that the gaze of persons that view the image is likely to be attracted by the target image regions (see Sec. 3.2). The fact that AC’09 has a negative NTOS can be explained by the fact that the target objects are too small and non-target areas too heterogeneous for Achanta’s simple approach [AHES09].

**Are the target locations center biased?** Since Gaze@Flickr is composed of web images, we can expect that the data is influenced by photographer biases (see Sec. 4.2). If we look at the results in Tab. 4.8, we can answer this question without an empirical analysis. Without integrated gaze information, LDRC+CB performs better than LDRC with respect to almost all evaluation measures. This indicates that the data is center biased. However, the target object location radii do not follow a Gaussian distribution, see Fig. 4.20. In fact, we can reject the hypothesis of a Gaussian distribution with separate tests such as, *e.g.*, the Jarque-Bera test ( $p = 0.001$ ). This could be expected, because in a considerable number of images the gazing person is in the image’s center and not the target

## 4.4. GAZE FOLLOWING IN WEB IMAGES

Method	PHR	FHR	$\int$ FHR	NTOS	$F_\beta$	$F_1$	$\int$ ROC
	no gaze						
center bias	8.44%	12.37%	0.1531	0.6231	0.0664	0.0948	0.5959
heuristic, CCH'12	12.20%	15.48%	0.3201	0.7212	0.0844	0.1063	0.6812
heuristic, GBVS'07	13.92%	19.08%	0.2209	0.8983	0.0642	0.0920	0.5775
heuristic, IK'98	13.35%	18.26%	0.2140	0.8711	0.0684	0.0974	0.6065
heuristic, PFT'07	11.71%	15.15%	0.3136	0.8240	0.0856	0.1059	0.7061
heuristic, DCT'11	12.29%	15.89%	0.3167	0.8440	0.0802	0.1035	0.6782
heuristic, EPQFT	11.06%	14.17%	0.3223	0.8369	0.0869	0.1090	0.7079
heuristic, QDCT	11.63%	15.23%	0.3256	0.8625	0.0833	0.1066	0.6905
heuristic, AC'09	2.54%	3.44%	0.0973	-0.1437	0.0480	0.0656	0.4575
heuristic, MSSS'10	10.48%	13.60%	0.2534	0.5087	0.0762	0.0969	0.6215
heuristic, RC'10	12.29%	17.12%	0.2434	0.6270	0.0774	0.1002	0.6603
heuristic, LDRC	11.06%	14.91%	0.2248	0.5354	0.0706	0.0930	0.6463
heuristic, LDRC+CB	13.35%	17.94%	0.2501	0.7180	0.0775	0.1063	0.6569
heuristic, QDCT + LDRC+CB	11.47%	14.99%	0.3321	0.9291	0.0848	0.1097	0.6993
CRF, QDCT	8.51%	14.94%	0.2079	0.7330	0.0815	0.0890	0.5583
CRF, QDCT & RC'10	11.08%	19.98%	0.2741	0.9204	0.1019	0.1188	0.6215
CRF, QDCT & LDRC	9.79%	16.12%	0.2319	0.8158	0.0898	0.1045	0.5945
CRF, QDCT & LDRC+CB	14.34%	22.55%	0.3116	0.9851	0.1116	0.1285	0.6608
	gaze "blank"						
heuristic, QDCT	28.91%	38.17%	0.5634	1.9352	0.1234	0.1561	0.7590
heuristic, RC'10	24.24%	34.23%	0.5028	1.6816	0.1212	0.1526	0.7580
heuristic, LDRC	24.08%	33.66%	0.4868	1.5990	0.1181	0.1514	0.7668
heuristic, LDRC+CB	24.16%	34.64%	0.5013	1.6915	0.1265	0.1632	0.7573
heuristic, QDCT + LDRC+CB	29.89%	39.48%	0.5956	2.0258	0.1264	0.1603	0.7619
CRF, QDCT	29.48%	45.80%	0.5228	1.8771	0.2053	0.2314	0.8068
CRF, QDCT & RC'10	33.63%	49.75%	0.5709	1.9805	0.2108	0.2387	0.8216
CRF, QDCT & LDRC	31.75%	48.76%	0.5644	1.9703	0.2121	0.2381	0.8180
CRF, QDCT & LDRC+CB	34.72%	52.62%	0.6053	2.0509	0.2226	0.2466	0.8303
	gaze "full"						
heuristic, QDCT	32.76%	42.10%	0.6303	2.2066	0.1314	0.1655	0.7715
heuristic, RC'10	31.20%	42.75%	0.5971	1.9711	0.1317	0.1650	0.7739
heuristic, LDRC	31.29%	42.26%	0.5817	1.8760	0.1308	0.1618	0.7858
heuristic, LDRC+CB	30.96%	42.59%	0.5816	1.9951	0.1388	0.1782	0.7741
heuristic, QDCT + LDRC+CB	34.64%	45.37%	0.6715	2.3198	0.1350	0.1705	0.7741
CRF, QDCT	40.06%	57.67%	0.6572	2.5321	0.2579	0.2923	0.8887
CRF, QDCT & RC'10	43.62%	65.18%	0.7114	2.6589	0.2652	0.2970	0.8932
CRF, QDCT & LDRC	43.62%	62.61%	0.6957	2.5936	0.2628	0.2963	0.8903
CRF, QDCT & LDRC+CB	44.02%	66.17%	0.7263	2.7071	0.2712	0.3024	0.8981

Table 4.8.: Target object detection performance on the Gaze@Flickr dataset. A description of the evaluation measures can be found in Sec. 4.3.1.E.

object. Nonetheless, our Gaussian center bias achieves a better performance than RC’10’s intrinsic bias, see Tab. 4.8.

## D. Results

Our quantitative evaluation results are shown in Tab. 4.8 and example CRF predictions are depicted in Fig. 4.18 (portraits) and Fig. 4.19 (landscapes).

We would like to start with a short reminder that we rely on two classes of evaluation measures: First, measures that evaluate the ability to focus the target object (PHR, FHR, and  $\int$ FHR). Second, measures that mainly evaluate the saliency map’s ability to separate the target object from the background (NTOS,  $F_\beta$ ,  $F_1$ , and  $\int$ ROC). At this point, we would like to explain one of the reasons, why we did not employ the  $F_\beta$  and  $F_1$  measure in Sec. 4.3. As we can see in Tab. 4.8, the  $F_\beta$  and  $F_1$  values are substantially smaller compared to the values that we observed on the MSRA dataset, see Tab. 4.1. This is related to the smaller target object size, because the errors made for the background pixels have a stronger influence on the evaluation measure for smaller target object sizes. If we consider that the target objects in the PointAT and ReferAT dataset are even smaller, it is understandable that the evaluation measures lose their informative value on these datasets.

Without integrated gaze information and without CRFs, we can see that QDCT exhibits a better salient object detection performance (*i.e.*,  $F_\beta$ ,  $F_1$ , and  $\int$ ROC) than the region contrast algorithms (*i.e.*, LDRC, LDRC+CB, and RC’10), whereas LDRC+CB and RC’10 provide a better performance in terms of PHR and FHR. Here, we can observe the influence of the center bias, because LDRC exhibits a lower performance than RC’10, LDRC+CB and, as a sidenote, also QDCT. If we integrate gaze, it is evident that QDCT is superior in terms of all evaluation measures that are related to the FoA on the basis of the “blank” gaze annotation. Furthermore, LDRC+CB provides a better performance with the “full” gaze annotation in terms of salient object detection evaluation measures. However, this situation is not perfectly consistent over both gaze annotations.

Similar to our experience with pointing gestures and spoken target information (see Sec. 4.3.1 and Sec. 4.3.2; “detected” vs “annotated”), we can observe that the evaluated CRFs are often able to provide a better performance with the “blank” gaze annotation than the heuristic baselines with the “full” gaze direction, especially in terms of salient object detection (*i.e.*,  $F_\beta$ ,  $F_1$ , and  $\int$ ROC). This comes at no surprise, since CRF are well known for their good performance in various image segmentation tasks, see Sec. 4.1.2.A. If we compare the CRFs to the heuristic method with the same gaze annotation, we can see that the CRFs outperform the heuristic gaze integration by a considerable margin, see Tab. 4.8.

Since we want to integrate our work on salient object detection and visual saliency, we investigate the performance that we can achieve when we let the CRFs combine QDCT with LDRC, LDRC+CB, or RC’10. To serve as a heuristic baseline, we present the results that we can achieve with a linear integration of the

saliency maps of QDCT and LDRC+CB (*i.e.*, “QDCT + LDRC+CB”). As can be seen in Tab. 4.8, the CRFs that use the saliency maps of QDCT and LDRC+CB as features are able to achieve a considerably higher performance than the heuristic linear integration scheme. If we compare the quantitative results of the evaluated CRFs, then LDRC+CB is the basis for the best performance, followed by RC’10, and then LDRC; almost perfectly consistent over all evaluation measures. Here, the rightmost image of Fig. 4.18 is an interesting example that illustrates the differences between LDRC and LDRC+CB, *i.e.* that LDRC is not biased to assign a higher saliency to segments at the image’s center. Thus, the fact that LDRC+CB and RC’10 lead to a better performance than LDRC is not surprising, because the target objects in the Gaze@Flickr dataset appear to be biased toward the image center, as has been discussed in Sec. 4.4.3.C. However, the radii do not follow half-Gaussian distribution, as is apparent in Fig. 4.20. Nonetheless, our explicit Gaussian center bias model<sup>11</sup> in the LDRC+CB algorithm leads to a better performance compared to RC’10’s intrinsic bias. Accordingly, we can assume that the Gaussian bias better reflects the actual target object location distribution than the intrinsic bias of RC’10.

## E. Future Work

Unlike in the other technical sections of this dissertation, we would like to end with an outlook on future work, because we described exploratory work that leaves many aspects open. In many images in the Gaze@Flickr dataset, people look at people or faces. Accordingly, we would like to integrate face and person detection as another feature for the CRF. We actually refrained from doing so in the presented evaluation, because of the evaluation’s focus on our integration of the aspects described in Sec. 4.2 and 4.3. Furthermore, we have performed experiments with additional features such as, most importantly, histogram of oriented gradients (HOG) and local binary patterns (LBP). These features efficiently encode information about edges and segments around each image pixel and thus can further improve the performance. However, we are currently unable to train the CRFs with these additional features for the complete Gaze@Flickr dataset, because the training would require more random access memory (RAM) than is available on our servers. As a sidenote, we require roughly 110 GB of memory to train the models that achieve the results presented in Tab. 4.8. Finally and most importantly, we would like to replace the groundtruth that currently forms the foundation of our experiments with automatically estimated gaze directions. However, since gaze estimation in the wild is still an unsolved problem, there currently does not exist any computer vision algorithm that is able to provide gaze estimates on our Gaze@Flickr dataset. Consequently, we are also interested in using the Gaze@Flickr dataset to develop novel gaze estimation

<sup>11</sup>Please note that we could have replaced the Gaussian center bias model with a model specifically adapted to the Gaze@Flickr dataset. We refrained from doing so, because we prefer not to overadapt to the Gaze@Flickr dataset.

methods that work on images that we can find in the web and other largely unconstrained image sources.



## 4.5 Summary and Future Directions

---

We presented how we can identify an object of interest that another person wants us to look at and analyze. We addressed two domains: First, attentional signals in human-robot interaction (HRI) that can guide the perceptual saliency. Second, images that have been composed by photographers; without and with the display of additional attentional signals such as people looking at things. For this purpose we came in contact with a wide range of research fields such as salient object detection, photographic image composition, perceptual saliency, image segmentation, HRI, pointing gestures, natural language processing and dialogue, and joint attention.

We initially addressed visual saliency models for HRI in 2010. At this point, only individual relevant aspects have been addressed, most importantly: First, the inherent inaccuracy of pointing gestures and the concept of a corridor of attention. Second, that the perceptual saliency can influence the generation and resolution of multimodal referring acts. Third, how linguistic references and knowledge about a target object’s visual appearance can influence visual search patterns. However, there did not exist any computer vision or robotic system that systematically integrated these aspects. Furthermore, all systems that tried to identify pointed-at objects relied on the assumption that all objects and their locations in the environment are known beforehand. Among other aspects, this also contradicted one important goal in robotics: the ability to being able to intuitively teach a robot about unknown things. Accordingly, we integrated several previously isolated ideas, methods, and models to guide the visual saliency and this way help to establish joint attention in multimodal HRI. Here, we are able to efficiently guide the focus of attention in multimodal HRI toward image locations that are highly likely to depict the referent; often being able to directly identify the intended target object.

We knew from our preceding work on eye fixation prediction about the importance of dataset biases such as and most importantly the photographer bias (Sec. 3.2). Thus, we were surprised to find that this aspect has not been addressed for salient object detection, although the bias was apparent in the datasets. Since we were interested to apply salient object detection for our HRI tasks, we started to analyze, model, and remove the center bias in salient object detection methods to facilitate the transfer of such methods to other data domains. This way, we improved the state-of-the-art in salient object detection and derived the currently best unbiased salient object detection algorithm.

After having studied how photographers compose web images and people use attentional signals in interaction to direct our attention, we became interested in the combination of both research directions. Therefore, we collected a novel dataset to learn to identify the object that is being looked at in web images. In this type of images the photographer’s placement of objects as well as visible attentional signals such as gaze direct our attention toward specific objects that

form a central element of these images. Thus, similar to pointing gestures, we address a different problem than almost all work on gaze estimation, *i.e.* we are not interested in an exact gaze direction estimates and instead focus on the identification and segmentation of the image area that is being looked at. Again, with no knowledge about the target object’s visual appearance, class, type, size, or any other identifying information – except that it is being looked at. Here, we have to deal with several challenges, ranging from the huge image data variance to the sheer non-existence of reliable gaze estimation methods for this type of unconstrained image data. We have achieved promising results with the methods that we initially developed for HRI data. Yet, we consider this work as being mainly exploratory to help us assess the potential of our methods in unconstrained environments with huge deviations in, for example, illumination, depicted content, and image composition. Consequently, there remain many challenges and aspects of future work.

**Future work** Related to identifying the salient object in web images, we see a lot of potential in the integration of descriptors that encode the coarse layout, gist, or global context of the scene, because this information is typically related to the general image composition (*e.g.*, does the image focus on one object? Is there a visible horizon? Is it indoors or outdoors?). Since eye tracking experiments have already been an important aspect of our work (see Sec. 3.2), a logical next step would be to evaluate how good our models predict human gaze patterns for images and videos in which persons use verbal and/or non-verbal attentional signals. Since such datasets do not exist yet, an important aspect would be to either extend our existing datasets with eye tracking data or to create a new dataset specifically for eye tracking studies. With respect to gaze estimation in the wild, we have observed that human gaze direction estimates depend on the image content and context. Accordingly, as an important aspect of our future work, we want to jointly estimate the gaze direction and the image region that is being looked at.

# 5

## Conclusion

In addition to the discussion and presentation of future work in Sec. 3.6 and Sec. 4.5, let us briefly summarize our contributions and provide an outlook on ongoing and future work to conclude this thesis.

### 5.1 Summary

---

We derived several novel quaternion-based spectral visual saliency models (QDCT, ESR, ESW, and EPQFT), all of which perform state-of-the-art on three well-known eye tracking datasets. Furthermore, we proposed to decorrelate each image’s color information as a preprocessing step for a wide variety of visual saliency models. We have shown that color space decorrelation can improve the performance by about 4% (normalized) for eight visual saliency algorithms on three established datasets with respect to three complementing evaluation measures. Although an improvement of 4% is far from drastic, it is nevertheless a considerable achievement, because we are not aware of any other method or preprocessing step that is able to consistently and significantly improve the performance of such a wide range of algorithms. Furthermore, we improved the state-of-the-art in predicting where people look when human faces are visible in the image. Compared to Cerf *et al.*’s approach, we were able to improve the performance by 8% (*i.e.*, 25.2% normalized by the ideal AUC) with automatic face detections.

quaternion-  
based spectral  
saliency

color  
decorrelation

To realize auditory attention, we introduced a novel auditory saliency model that is based on the Bayesian surprise of each frequency. To allow for real-time computation on a robotic platform, we derived Gaussian surprise, which is efficient to calculate due to its simple closed form solution. Since we addressed a novel problem domain, we had to introduce a novel quantitative, application-oriented evaluation methodology and evaluated our model’s ability to detect arbitrary salient auditory events. Our results show that Bayesian surprise can efficiently and reliably detect salient acoustic events, which is shown by  $F_1$ ,  $F_2$ , and  $F_4$  scores of 0.767, 0.892, and 0.967.

auditory  
surprise

We combined auditory and visual saliency in a biologically-plausible model based on crossmodal proto-objects to implement overt attention on a humanoid robot’s head. We performed a series of behavioral experiments, which showed

audio-visual  
proto-objects

multiobjective  
exploration

that our model exhibits the desired behaviors. Based on a formalization as multiobjective optimization problem, we introduced ego motion as a further criterion to plan which proto-object to attend next. This way, we were able to substantially reduce the amount of head ego motion while still preferring to attend the most salient proto-objects first. Our solution exhibits a low normalized cumulated joint angle distance (NCJAD) of 15.0%, which represents that the chosen exploration order requires a low amount of ego motion to attend all proto-objects, and a high normalized cumulated saliency (NCS) of 83.3%, which indicates that highly salient proto-objects are attended early.

salient object  
distribution

We investigated how the spatial distribution of objects in images influences salient object detection. Here, we provided the first empirical justification for a Gaussian center bias. This is shown by a probability plot correlation coefficient (PPCC) of 0.9988 between a uniform distribution and the angular distribution of salient objects around the image center, and a PPCC of 0.9987 between a half-Gaussian distribution and the distribution of distances of salient objects to the image center. Then, we demonstrated that the performance of salient object detection algorithms can be substantially influenced by undocumented spatial biases. We debiased the region contrast algorithm and subsequently integrated a well-modeled Gaussian bias. This way, we achieved two goals: First, through integration of our explicit Gaussian bias, we improved the state-of-the-art in salient object detection for web images and at the same time quantified the influence of the center bias. Second, we derived the currently best unbiased salient object detection algorithm, which is advantageous for other application domains such as, *e.g.*, surveillance and robotics.

debiased salient  
object  
detection

attention for  
multimodal  
interaction

We presented saliency models that are able to integrate multimodal signals such as pointing and spoken object descriptions to guide the attention in human-robot interaction. We started with an initial heuristic model that combines our spectral saliency detection with a probabilistic corridor of attention, *i.e.* the “probabilistic pointing cone”, to reflect the spatial information given by pointing references. Additionally, we discussed a biologically-inspired neuron-based saliency model that is able to integrate knowledge about the target object’s appearance into visual search. We outperform both models by training conditional random fields that integrate features such as, most importantly, our locally debiased region contrast, multi-scale spectral visual saliency with decorrelated color space, the probabilistic pointing cone, and target color models. This way, we are able to focus the correct target object in the initial focus of attention for 92.45% of the images in the PointAT dataset, which does not provide spoken target descriptions, and 75.21% for the ReferAT dataset, which includes spoken target references. This translates to an improvement of +10.37% and +25.21% compared to the heuristic and neuron-based saliency models, respectively.

gaze following  
in web images

Finally, we learn to determine objects or object parts that are being looked-at by persons in web images. This can be interpreted as a form of gaze following in web images. For this purpose, we integrated our work on salient object detection in web images and the interpretation of attentional signals in human-robot

interaction. Consequently, we transferred our methods and train conditional random fields to integrate features such as, most importantly, spectral visual saliency, region contrast saliency, and a probabilistic corridor of interest that represents the observed gaze direction. This way, the looked-at target object is focused in the initial focus of attention for 66.17% of images in a dataset that we collected from Flickr.

To quantify the performance of our approaches, we had to collect several datasets and even propose novel evaluation procedures, because we often addressed novel tasks, problems, and domains. We derived novel evaluation procedures for these tasks: First, we quantified the ability of our auditory saliency model to determine arbitrary salient acoustic events. Therefore, we relied on measures that are commonly used to evaluate salient object detection algorithms. Second, we introduced several novel evaluation measures to evaluate tradeoffs made by our multiobjective exploration path strategies. Furthermore, we proposed several measures to quantify the ability of saliency models to highlight target objects and focus the objects after a minimum amount of focus of attention shifts. We collected novel datasets for the following tasks: First, we created a dataset that consists of 60 videos to evaluate multiobjective exploration strategies. Second and third, we recorded two new datasets to evaluate how well we are able to guide our saliency model in human-robot interaction; in the absence (PointAT) and presence (ReferAT) of spoken target object information. For this purpose, fourth, we also gathered the Google-512 dataset to train our color term models. Fifth, to evaluate the identification and segmentation of gazed-at objects in web images, we collected the Gaze@Flickr dataset that we selected out of one million Flickr images.

evaluation  
procedures

datasets

## 5.2 Future Work

There are many aspects of high-level influences on visual saliency, search, and attention that represent interesting research directions such as, for example: How does coarse contextual information about the scene prime visual attention mechanisms and influence eye gaze patterns? Or, how do depicted pointing gestures and gaze directions in images influence gaze patterns of human observers? Here, it would be very interesting to compare the predictions of our top-down guided saliency models against human gaze behavior.

Furthermore, we have seen that human gaze estimates seem to depend on the visible image content. Accordingly, it seems that a very interesting research direction is to fuse gaze estimation and the detection of potential looked-at regions. In our opinion, an integrated approach should be able to jointly improve both estimates, *i.e.* the estimated gaze direction and the predicted looked-at image region. Naturally, the same assumption and approach could also benefit pointing gestures and other directed non-verbal signals such as, *e.g.*, head nods.

In our opinion, the currently most important open question for auditory and also audio-visual saliency models is a quantitative evaluation methodology based on human behavior. For example, it would be interesting to investigate the potential use of eye pupil dilation to evaluate bottom-up auditory saliency models. Furthermore, eye tracking experiments and datasets to investigate audio-visual saliency models and integration are still lacking. Consequently, the collection of a public benchmark dataset for audio-visual saliency models would represent a very valuable contribution to the field that could accelerate future research.

Our crossmodal proto-object model provides us with great flexibility to implement overt attention and saliency-based exploration mechanisms. However, since our hardware platform was stationary, we were only able to plan and evaluate head motion. Accordingly, one major open task is the implementation and evaluation of a parametric proto-object model as a basis for exploration mechanisms of non-stationary platforms. However, we are very confident that our model will prove to be viable in that scenario, the biggest challenge being sufficiently accurate robot self-localization that is necessary to update the proto-objects in the world model.



# Applications

In the following, we briefly present further applications that use saliency models that we developed as part of this thesis. First, described in Sec. A.1, Martinez uses our Gaussian surprise model, see Sec. 3.3.1.B, to detect patient agitation in intensive care units. Second, described in Sec. A.2, Rybok relies on our quaternion image signature saliency model, see Sec. 3.2.1, and a notion of visual proto-objects, see Sec. 3.4, to improve the accuracy of activity recognition.

## A.1 Patient Agitation

---

Appropriate patient sedation is a complex problem in intensive care units (ICUs), because excessive sedation can threaten the patient’s life while insufficient sedation can lead to excessive patient anxiety and agitation. The appropriate sedation protocol varies between patient and it does not just depend on easily measurable vital signs (*e.g.*, heart rate), but also on behavioral cues that indicate signs for agitation, which are usually recorded by the nursing staff. In order to automate and improve the incorporation of such behavioral cues, it has been suggested to use computer vision systems to continuously monitor the patient’s body and face for signs of stress, discomfort, or abnormalities. Here, the fact that such an automated system provides quantified and more objective measurements is a welcomed side effect. The most common behavioral cues are patient agitation patterns, because they are meaningful, robust to occlusions, and relatively easy to measure. For this purpose, Martinez proposed to apply surprise to detect and quantify agitation patterns that become apparent in a patient’s face.

### A.1.1. Method

The first step in Martinez’s framework is to determine the bed position, which is assumed to be roughly centered in the sensor setups field of view, see Fig. A.1. Then, the bed plane is estimated by a segmentation via region growing based on the depth map. Then two features are extracted from each image frame: First, the the depth camera’s information is used to calculate the bed occupancy feature, which measures the occupied volume over the bed plane. This feature is suited to detect events such as when the patient enters and exits the bed.

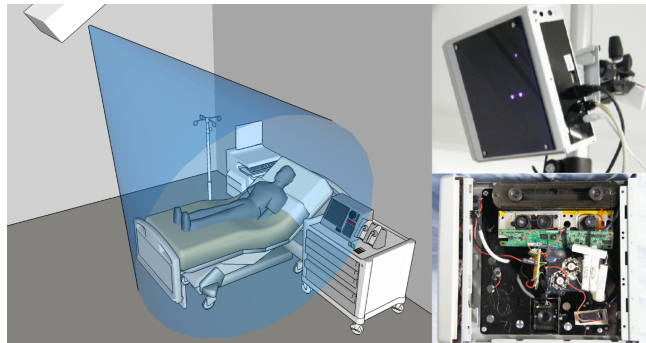


Figure A.1.: An illustration of Martinez’s medical recording device (MRD) that is used to monitor patients in ICUs. The device uses stereo and depth cameras to record the entire body, while an additional high-resolution camera focuses on the face region. Image from [MS13].

Furthermore, it can also be used as a feature for body agitation and – given sufficient accuracy of the depth sensor – breath patterns. Second, the face camera is used to calculate a measure for agitation signals that are visible in the patient face. For this purpose, each image is resized to  $32 \times 32$  px and Gaussian surprise is calculated for each pixel with a history length of 25 frames which is equivalent to 500 ms. This feature is suited to detect facial agitation patterns that are evident when the patient shows signs of discomfort.

### A.1.2. Qualitative Evaluation

Due to the subjective behavior of the measurements and the lack of a public database or even a common evaluation methodology, it is impossible to quantitatively compare Martinez’s results to alternative approaches. Instead, a qualitative behavioral experimental evaluation was performed. For this purpose, Martinez enacted and simulated a series of scenarios and compared the observed system behavior with the desired behavior, see Fig. A.2. It was shown that the system provides reasonable behavioral descriptions of scenarios. In contrast to prior art [BHCS07, GCAS<sup>+</sup>04, BHCS07], the system is able to achieve this without relying special markers, invasive measures, or the need to control the illumination conditions.

## A.2 Activity Recognition

---

Action and activity recognition is an important computer vision task with many potential application areas such as, for example, human-robot interaction, surveillance and multimedia retrieval. It is important to understand what differentiates the concepts of actions and activities. While the first describe simple motion events (*e.g.*, “person stands up”), the latter describe complex



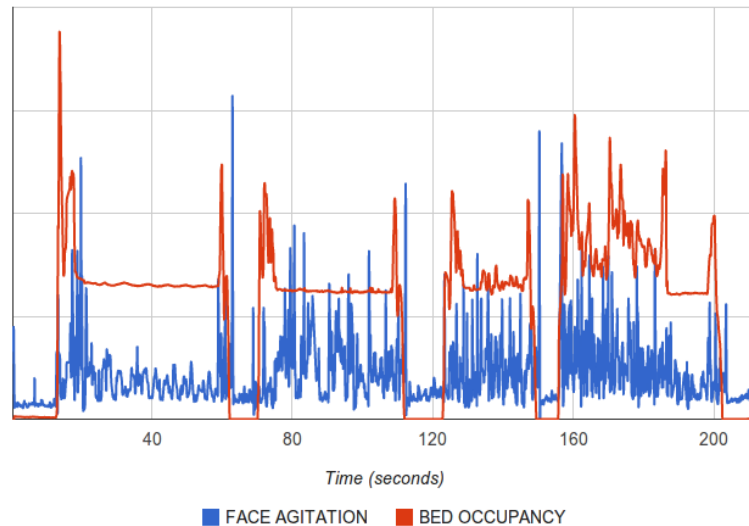


Figure A.2.: Surprise-based face agitation estimation (blue) in a simulation of 4 scenarios. From second 10 to 60: Sleeping relaxed shows an almost flat bed occupancy indicator and low agitation levels in the face. From second 70 to 110: Sleeping with pain expressions is not reflected in the volumetric information, but it is detected by the face agitation levels. From second 120 to 145: Being restless in bed is reflected by a clear response in both indicators. From second 145 to 200: Strong compulsions ending with an accident and sudden loss of consciousness. This illustration is best viewed in color. Image from [MS13].

action sequences (*e.g.*, “person cleans kitchen”) that form an activity. According to action identification theory, actions and, as a consequence, activities are not just defined by motion patterns but derive their meaning from context [VW87]. For example, the motion patterns for “wiping” and “waving” can look very similar and hard to distinguish without the context in which they are performed. Consequently, it can be necessary to incorporate – among other contextual cues – the location where an action is performed or which objects are manipulated in the activity classification process.

Most work on activity recognition does not integrate contextual knowledge or requires specifically trained detectors. However, such detectors require a considerable amount of manually annotated training data, which is costly to acquire and makes it hard to transfer the activity recognition systems to new application areas and domains. As an alternative, Rybok proposes to use salient proto-objects to detect candidate objects, object parts, or groups of objects (see Sec. 3.4) that are potentially relevant for the activity. This approach makes it possible to integrate contextual object knowledge into the activity recognition

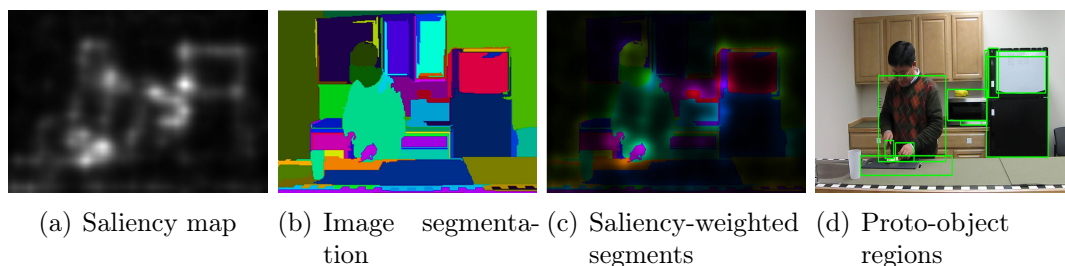


Figure A.3.: Example of the proto-object extraction approach. This illustration is best viewed in color. Image from [RSAHS14].

based on unsupervised methods, *i.e.* without the need for accurate object labels or detectors.

### A.2.1. Method

Rybok relies on the QDCT image signatures, see Sec. 3.2.1, to calculate the visual saliency of each frame in a video sequence, see Fig. A.3. To determine the image’s proto-object regions, Felzenszwalb’s graph-based algorithm [FH04] is used to segment each frame. Following the classical winner-take-all method for attentional shifts and inhibition of return (see Sec. 2.1.1), Rybok iteratively extracts the image segment that contains the most salient peak and then inhibits the saliency at the segment’s location. This is repeated, until the saliency map’s maximum saliency value either falls below 70% of its initial maximum or the 30 most salient segments have been extracted. The extracted segments form the set of proto-object regions that serves as context (*i.e.*, as object or object part candidates) for the activity recognition. To this end, the appearance of each extracted proto-object region is encoded by Dalal and Trigg’s histogram of oriented gradients (HOG) [DT05]. Given the activity recognition training sequences, the proto-object HOG feature vectors are clustered with k-means to generate a proto-object codebook.

The classification of motion patterns is based on Laptev *et al.*’s space time interest points (STIP) [LMSR08]. Laptev *et al.*’s Harris 3D interest point detection is used to determine interesting points in space and time. Either the histogram of optical flow (HOF) alone or a combination of HOF and HOG is used as feature vector to describe each interest point. The HOG descriptor in this context is different from Dalal and Trigg’s HOG descriptor [DT05], because it accumulates the gradients within the spatio-temporal STIP region.

Each image sequence is represented by a bag-of-words feature vector with a 1000-element codebook for motion features (HOG or HOG-HOF) and a 200-element codebook for proto-objects. Given these features, a linear SVM is trained to classify video sequences. To boost the performance, the feature vector is normalized and then each element is raised to the power of  $\alpha = 0.3$  (*cf.* [RR13]).

### A.2.2. Results

Rybok evaluates the approach on three activity recognition benchmark datasets: URADL [MPK09], CAD-120 [KS13], and KIT Robo-Kitchen [RFHS11]. As can be seen in Tab. A.1, the saliency-driven approach is able improve the state-of-the-art on all three datasets, although the employed motion features are relatively common and simple. To demonstrate the benefit of saliency-driven object candidate extraction, Rybok compares to an alternative approach in which all image segments are used as contextual information (“all segments”, Tab. A.1). Although the performance achieved with all image segments is better than the model without context information, it is clear that the saliency-driven image region selection provides a substantial performance benefit, see Tab. A.1.

(a) URADL	
Method	Accuracy (%)
HOF	79.3
HOF & all segments	86.7
HOF & proto-objects	97.7
HOGHOF	94.0
HOGHOF & all segments	94.7
HOGHOF & proto-objects	<b>100.0</b>
Matikainen <i>et al.</i> [MHS10], 2010	70.0
Messing <i>et al.</i> [MPK09], 2009	89.0
Prest <i>et al.</i> [PFS12], 2012	92.0
Wang <i>et al.</i> [WCW11], 2011	96.0
Yi and Lin [YL13], 2013	98.0
(b) CAD-120	
Method	Accuracy (%)
HOF	72.6
HOF & all segments	75.0
HOF & proto-objects	<b>79.0</b>
HOGHOF	70.0
HOGHOF & all segments	72.6
HOGHOF & proto-objects	77.4
Sung <i>et al.</i> [SPSS12], 2012	26.4
Koppula <i>et al.</i> [KS13], 2013	75.0
(c) KIT	
Method	Accuracy (%)
HOF	85.6
HOF & proto-objects	<b>88.7</b>
HOGHOF	86.6
HOGHOF & proto-objects	88.5
Rybok <i>et al.</i> [RFHS11], 2011	84.9
Onofri <i>et al.</i> [OSI13], 2013	88.3

Table A.1.: Activity recognition results on the (a) URADL, (b) CAD-120, and (c) KIT datasets. As can be seen, the combination of contextual proto-object information and simple HOG and HOG-HOF features provides state-of-the-art performance.

# B

## Dataset Overview

Throughout this thesis, we rely on several datasets to evaluate our approaches. In the following, we provide a short overview of these datasets.

### Main datasets (Chapter 3):

- 1. Bruce/Toronto: Eye tracking (Sec. 3.2.1.F)**  
This dataset [BT09] contains 120 color images depicting indoor and outdoor scenes. The dataset contains eye tracking data of 20 subjects (4 seconds, free-viewing).
- 2. Judd/MIT: Eye tracking (Sec. 3.2.1.F)**  
This dataset contains 1003 images of varying resolutions [JEDT09] that were collected from Flickr and the LabelMe database. Eye tracking data was recorded for 15 subjects (3 seconds, free-viewing).
- 3. Kootstra: Eye tracking (Sec. 3.2.1.F)**  
This dataset [KNd08] contains 100 images that were collected from the McGill calibrated color image database [OK04]. The images were shown to 31 subjects (free-viewing).
- 4. Cerf/FIFA: Eye tracking (Sec. 3.2.3.C)**  
To evaluate the influence of faces on human visual attention, this dataset [CFK09] consists of eye tracking data (2 seconds, free-viewing) of 9 subjects for 200 images of which 157 contain one or more faces.
- 5. CLEAR2007: Acoustic events (Sec. 3.3.2.B)**  
The CLEAR2007 acoustic event detection dataset [CLE, TMZ<sup>+</sup>07]) contains recordings of meetings in a smart room. A human analyst marked and classified occurring acoustic events that were remarkable enough to “pop-out”. 14 acoustic event classes were identified and tagged (*e.g.*, “laughter”, “door knocks”, “phone ringing” and “key jingling”). Events that could not be identified by the human analyst were tagged as “unknown”.

**Main datasets (Chapter 4):**

- 6. IROS2012: Exploration strategies (Sec. 3.5)**

To evaluate scene exploration strategies [KSKS12], this dataset consists of 60 videos (30 seconds each), in which specific sequences were re-enacted in three scenarios: office scenes, breakfast scenes, and neutral scenes.
- 7. PointAT: Pointing (Sec. 4.3.1.D)**

This dataset contains 220 instances of 3 persons pointing at objects in an office environment and conference room [SRF10]. Pointed-at objects were predicted online while recording the dataset and used to automatically zoom on the target object, which additionally makes it possible to evaluate the influence of foveation on object recognition.
- 8. ReferAT: Pointing & language (Sec. 4.3.2.E)**

This dataset contains 242 multimodal referring acts (composed of pointing gestures and spoken object descriptions) that were performed by 5 persons referring to a set of 28 objects in a meeting room [SF10a]. The objects were chosen in such a way that, in most situations, object names and colors are the most discriminant verbal cues for referring-to the referent.
- 9. Gaze@Flickr: Gaze (Sec. 4.4.2)**

Our Gaze@Flickr dataset contains 863 Flickr images that contain 1221 gaze references, *i.e.* persons gazing at a target object. The dataset provides annotated head regions of the gazing persons as well as two different gaze directions that were annotated under different viewing conditions.
- 10. MSRA: Salient objects (Sec. 4.2.1)**

MSRA is the most widely used dataset to evaluate salient object detection. It has been created by Achanta *et al.* and Liu *et al.* [AHES09, LSZ<sup>+</sup>07] and consists of 1000 images with binary segmentation masks of the target object.

**Additional datasets:**

- A. Google-512: Color terms (Sec. 4.3.2.C)**

We use our Google-512 dataset [SF10b] to learn color term models. The dataset consists of 512 images for each of the eleven basic English color terms. The images were collected using Google’s image search. The learned color term models are commonly evaluated on another dataset: Weijer *et al.*’s e-Bay dataset [vdWSV07].
- B. Brown: Language (Sec. 4.3.2.A)**

We use the Brown corpus [Fra79] and its annotated part-of-speech (POS) tags to train a Brill tagger [Bri95], which we use to determine noun-phrases and their constituents with a shallow parser which is based on regular

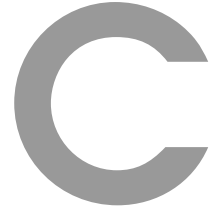
---

expressions. The Brown corpus is a general text collection, *i.e.* general corpus, that contains 500 samples of English text, compiled from works that were published in the United States in 1961.

Additionally, we performed behavioral experiments to evaluate our audio-visual overt attention system, see Sec. 3.4. Furthermore, we collected some datasets that are not relevant to the content in this thesis such as, *e.g.*, the Flower Box dataset for visual obstacle detection and avoidance [KSS13].







# Color Space Decorrelation: Full Evaluation

In the following, we present further color space decorrelation evaluation results that complement our evaluation and discussion in Sec. 3.2.2.B. For this purpose, we provide the evaluation results of our baseline algorithms for three evaluation measures, see Tab. C.1. We provide results for the following color spaces: RGB, CIE Lab, CIE XYZ, ICOPP (*e.g.*, [GMZ08]), LMS [SG31], and Gauss [GvdBSG01]. Furthermore, we use statistical tests (see Sec. 3.2.2.B) to test the performance of each algorithm on the original color space against the performance based on the decorrelated color space (“better”, “better or equal”, “probably equal”, “equal or worse”, and “worse”).

## Evaluation measures

In this dissertation, we focused on the AUC evaluation measure in the main document, since it is the most established measure. However, in our extended evaluation results for color space decorrelation (Appx. C), we use the CC and NSS as complementary evaluation measures to show that color space decorrelation is beneficial as quantified by all three evaluation measure classes [RDM<sup>+</sup>13], see Sec. 3.2.1.F. Let us present the three evaluation measures in more detail:

The shuffled, bias-correcting AUROC or shorter AUC measure (see, *e.g.*, [HHK12]) tries to compensate for biases such as, *e.g.*, the center-bias that is commonly found in eye tracking datasets. To this end, it defines a positive and a negative set of eye fixations for each image. The positive sample set contains the fixation points of all subjects on that image. The negative sample set contains the union of all eye fixation points across all other images from the same dataset. To calculate the AUROC, each saliency map is thresholded and the resulting binary map can be seen as being a binary classifier that tries to classify positive and negative samples. Sweeping over all thresholds leads to the ROC curves and defines the the area under the ROC curve. When using the AUROC as a measure, the chance level is 0.5 (random classifier), values  $< 0.5$  indicate negative correlation, values  $> 0.5$  represent positive correlation, and a AUROC of 1 means perfect classification.

AUC

- CC The linear correlation coefficient (CC) is a measure for the strength of a linear relationship between two variables. Let  $G$  denote the groundtruth saliency map that is generated by adding a Gaussian blur to the recorded eye fixations and  $S$  the algorithm’s saliency map [JOvW<sup>+</sup>05, RBC06], then  $CC(G, S) = \frac{cov(G, S)}{\sigma_G \sigma_S}$ , where  $\sigma_G$  and  $\sigma_S$  are the standard deviations of  $G$  and  $S$ , respectively. A CC close to +1 or −1 indicates an almost perfectly linear relationship between the prediction  $S$  and groundtruth  $G$ . As the CC approaches 0 there is less of a relationship, *i.e.* it is closer to being uncorrelated.
- NSS The normalized scanpath saliency (NSS) is the average saliency at human eye fixations in an algorithm’s saliency map. To make the values comparable, the saliency map is normalized to have zero mean and unit standard deviation [PIIK05, PLN02], *i.e.* a NSS of 1 means that the predicted saliency at recorded eye fixations is one standard deviation above average. Consequently, an NSS  $\geq 1$  indicates that the saliency map has significantly higher saliency values at locations that were fixated by the human subjects than at other locations. An NSS  $\leq 0$  means that the predicted saliency does not predict eye fixations better than picking random image locations, *i.e.* chance.

Method	Bruce/Toronto			Kootstra			Judd/MIT		
	ROC	CC	NSS	ROC	CC	NSS	ROC	CC	NSS
CAS’12	0.692	0.370	1.255	0.603	0.246	0.544	0.662	0.235	0.948
CCH’12	0.666	0.268	0.905	0.583	0.219	0.478	0.648	0.218	0.873
JEDA’09	0.624	0.420	1.379	0.549	0.307	0.651	0.665	0.342	1.351
AIM’09	0.666	0.261	0.898	0.574	0.176	0.383	0.637	0.184	0.747
GBVS’07	0.660	0.420	1.381	0.558	0.220	0.458	0.584	0.174	0.693
COH’06	0.650	0.310	0.990	0.547	0.263	0.510	0.697	0.210	0.990
IK’98	0.645	0.393	1.293	0.574	0.279	0.585	0.636	0.261	1.039
iNVT’98	0.544	0.155	0.553	0.518	0.092	0.210	0.536	0.099	0.418
Chance	0.5	→ 0	≤ 0	0.5	→ 0	≤ 0	0.5	→ 0	≤ 0

Table C.1.: Performance of selected baseline algorithms on the Kootstra, Judd/MIT, and Bruce/Toronto datasets.

AUC Method	RGB		Lab		ICOPP		
	raw	PCA	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.6661	<b>0.7031</b>	0.6974	<b>0.7072</b>	0.6881	<b>0.7032</b>	0.7019
QDCT	0.7033	<b>0.7157</b>	0.7149	<b>0.7210</b>	0.7135	0.7140	<b>0.7175</b>
EPQFT	0.7003	0.7142	<b>0.7158</b>	<b>0.7212</b>	0.7112	<b>0.7118</b>	<b>0.7156</b>
DCT'11	0.6915	<b>0.7196</b>	0.7121	0.7207	0.7114	<b>0.7184</b>	0.7166
AC'09	0.5406	0.5608	<b>0.5780</b>	<b>0.5735</b>	0.5510	0.5543	<b>0.5702</b>
GBVS'07	0.6030	<b>0.6620</b>	0.6614	0.6655	0.6374	<b>0.6637</b>	0.6617
PFT'07	0.6952	<b>0.7196</b>	0.7135	<b>0.7226</b>	0.7128	0.7179	<b>0.7189</b>
IK'98	0.6410	0.6723	<b>0.6772</b>	<b>0.6814</b>	0.6636	0.6721	<b>0.6756</b>

AUC Method	GAUSS		XYZ		CAT02LMS		
	raw	PCA	raw	PCA	raw	PCA	ZCA
CCH'12	0.6959	0.7019	<b>0.7038</b>	<b>0.6664</b>	0.6306	<b>0.6657</b>	0.6563
QDCT	0.7135	0.7153	<b>0.7171</b>	<b>0.6754</b>	0.6657	<b>0.6749</b>	0.6704
EPQFT	0.7110	0.7130	<b>0.7168</b>	<b>0.6746</b>	0.6638	<b>0.6739</b>	0.6678
DCT'11	0.7097	0.7190	<b>0.7205</b>	<b>0.6811</b>	0.6612	<b>0.6804</b>	0.6667
AC'09	0.5576	0.5620	<b>0.5622</b>	<b>0.5433</b>	0.5280	0.5438	<b>0.5451</b>
GBVS'07	0.6417	0.6618	<b>0.6619</b>	<b>0.6207</b>	0.5794	0.6203	<b>0.6206</b>
PFT'07	0.7114	0.7180	<b>0.7200</b>	<b>0.6820</b>	0.6619	<b>0.6818</b>	0.6678
IK'98	0.6702	0.6719	<b>0.6738</b>	<b>0.6393</b>	0.6099	0.6389	<b>0.6472</b>

Table C.2.: Color space decorrelation results as quantified by the AUC evaluation measure on the Bruce/Toronto dataset. This table contains color coded information and is best seen in color. Please refer to Tab. 3.3 for a color legend.

CC Method	RGB			Lab			ICOPP		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.2688	<b>0.3490</b>	0.3360	0.3408	0.3522	<b>0.3543</b>	0.3334	<b>0.3507</b>	0.3491
QDCT	0.3484	0.3943	<b>0.4034</b>	0.3988	0.4017	<b>0.4087</b>	0.3909	0.3933	<b>0.3991</b>
EPQFT	0.3023	0.3574	<b>0.3769</b>	0.3638	0.3627	<b>0.3786</b>	0.3447	0.3501	<b>0.3575</b>
DCT'11	0.3200	<b>0.4213</b>	0.4019	0.4068	<b>0.4220</b>	0.4213	0.4063	<b>0.4191</b>	0.4168
AC'09	0.0485	0.0861	<b>0.1091</b>	0.0764	0.0929	0.1151	0.0629	0.0716	<b>0.1047</b>
GBVS'07	0.2385	0.3364	<b>0.3391</b>	0.2820	<b>0.3462</b>	0.3432	0.2890	<b>0.3419</b>	0.3383
PFT'07	0.2905	<b>0.4007</b>	0.3840	0.3869	0.4046	<b>0.4050</b>	0.3857	<b>0.3998</b>	<b>0.3998</b>
IK'98	0.3204	0.3865	<b>0.3938</b>	0.3557	0.3848	<b>0.3965</b>	0.3640	<b>0.3877</b>	0.3851
CC Method	GAUSS			XYZ			CAT02LMS		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.3305	<b>0.3490</b>	0.3465	0.1836	<b>0.2620</b>	0.2485	0.1847	<b>0.2601</b>	0.2458
QDCT	0.3908	0.3918	<b>0.3949</b>	0.2637	0.2967	<b>0.3087</b>	0.2671	0.2962	<b>0.3102</b>
EPQFT	0.3516	0.3507	<b>0.3607</b>	0.2241	0.2606	<b>0.2759</b>	0.2273	0.2603	<b>0.2774</b>
DCT'11	0.3999	<b>0.4192</b>	0.4159	0.2517	<b>0.3359</b>	0.3119	0.2546	<b>0.3347</b>	0.3107
AC'09	0.0813	<b>0.0917</b>	0.0862	0.0160	0.0285	<b>0.0543</b>	0.0185	0.0295	<b>0.0498</b>
GBVS'07	0.3065	<b>0.3360</b>	0.3358	0.1760	0.2669	<b>0.2674</b>	0.1785	<b>0.2673</b>	0.2671
PFT'07	0.3776	<b>0.3974</b>	0.3963	0.2190	<b>0.3116</b>	0.2883	0.2217	<b>0.3102</b>	0.2870
IK'98	0.3813	0.3835	<b>0.3852</b>	0.2719	0.3257	<b>0.3458</b>	0.2738	0.3250	<b>0.3459</b>

Table C.3.: Color space decorrelation results as quantified by the CC evaluation measure on the Bruce/Toronto dataset. This table contains color coded information and is best seen in color. Please refer to Tab. 3.3 for a color legend.

NSS Method	RGB			Lab			ICOPP		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.9052	<b>1.1700</b>	1.1268	1.1415	1.1801	<b>1.1880</b>	1.1148	<b>1.1730</b>	1.1705
QDCT	1.1860	1.3376	<b>1.3665</b>	1.3558	1.3636	<b>1.3884</b>	1.3271	1.3347	<b>1.3533</b>
EPQFT	1.0342	1.2187	<b>1.2804</b>	1.2429	1.2349	<b>1.2917</b>	1.1752	1.1919	<b>1.2173</b>
DCT'11	1.0905	<b>1.4304</b>	1.3578	1.3861	<b>1.4357</b>	1.4315	1.3782	<b>1.4246</b>	1.4114
AC'09	0.1822	0.3203	<b>0.3950</b>	0.2807	0.3397	<b>0.4173</b>	0.2343	0.2672	<b>0.3864</b>
GBVS'07	0.7892	1.1080	<b>1.1185</b>	0.9273	<b>1.1412</b>	1.1309	0.9463	<b>1.1268</b>	1.1147
PFT'07	0.9946	<b>1.3664</b>	1.3052	1.3248	<b>1.3822</b>	1.3821	1.3143	<b>1.3651</b>	1.3604
IK'98	1.0651	1.2779	<b>1.3000</b>	1.1771	1.2729	<b>1.3079</b>	1.2055	<b>1.2816</b>	1.2714

NSS Method	GAUSS			XYZ			CAT02LMS		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	1.1093	<b>1.1718</b>	1.1631	0.6226	<b>0.8806</b>	0.8325	0.6260	<b>0.8737</b>	0.8223
QDCT	1.3285	1.3293	<b>1.3399</b>	0.9051	1.0099	<b>1.0413</b>	0.9160	1.0079	<b>1.0468</b>
EPQFT	1.2014	1.1941	<b>1.2288</b>	0.7752	0.8940	<b>0.9348</b>	0.7860	0.8933	<b>0.9399</b>
DCT'11	1.3583	<b>1.4253</b>	1.4127	0.8651	<b>1.1426</b>	1.0509	0.8747	<b>1.1376</b>	1.0457
AC'09	0.2998	<b>0.3397</b>	0.3176	0.0825	0.1258	<b>0.2038</b>	0.0898	0.1289	<b>0.1892</b>
GBVS'07	1.0093	<b>1.1071</b>	1.1062	0.5929	0.8785	<b>0.8814</b>	0.6008	0.8801	<b>0.8805</b>
PFT'07	1.2895	<b>1.3560</b>	1.3516	0.7583	<b>1.0637</b>	0.9772	0.7679	<b>1.0587</b>	0.9717
IK'98	1.2660	1.2694	<b>1.2722</b>	0.9132	1.0772	<b>1.1421</b>	0.9191	1.0748	<b>1.1423</b>

Table C.4.: Color space decorrelation results as quantified by the NSS evaluation measure on the Bruce/Toronto dataset. This table contains color coded information and is best seen in color. Please refer to Tab. 3.3 for a color legend.

AUC Method	RGB			Lab			ICOPP		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.6480	<b>0.6696</b>	<b>0.6708</b>	0.6674	<b>0.6733</b>	0.6722	0.6595	<b>0.6705</b>	0.6702
QDCT	0.6517	0.6608	<b>0.6625</b>	0.6599	0.6610	<b>0.6625</b>	0.6585	0.6593	<b>0.6613</b>
EPQFT	0.6484	0.6590	<b>0.6621</b>	0.6579	0.6581	<b>0.6609</b>	0.6547	0.6558	<b>0.6583</b>
DCT'11	0.6440	<b>0.6641</b>	0.6608	0.6581	<b>0.6645</b>	0.6638	0.6577	<b>0.6632</b>	0.6627
AC'09	0.5306	0.5513	<b>0.5810</b>	0.5493	0.5514	<b>0.5592</b>	0.5452	0.5492	<b>0.5585</b>
GBVS'07	0.5846	<b>0.6343</b>	0.6327	0.6207	<b>0.6367</b>	0.6362	0.6162	<b>0.6349</b>	0.6342
PFT'07	0.6449	<b>0.6652</b>	0.6627	0.6597	<b>0.6653</b>	0.6650	0.6590	0.6639	<b>0.6647</b>
IK'98	0.6367	0.6572	<b>0.6585</b>	0.6508	0.6581	<b>0.6582</b>	0.6493	0.6556	<b>0.6564</b>
AUC Method	GAUSS			XYZ			CAT02LMS		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.6657	0.6700	<b>0.6705</b>	0.6207	<b>0.6460</b>	0.6457	0.6202	0.6452	<b>0.6458</b>
QDCT	0.6573	0.6600	<b>0.6610</b>	0.6262	0.6385	<b>0.6448</b>	0.6262	0.6380	<b>0.6455</b>
EPQFT	0.6552	0.6570	<b>0.6595</b>	0.6240	0.6362	<b>0.6434</b>	0.6240	0.6353	<b>0.6446</b>
DCT'11	0.6549	<b>0.6639</b>	0.6632	0.6216	<b>0.6452</b>	0.6430	0.6218	<b>0.6450</b>	0.6429
AC'09	0.5459	0.5526	<b>0.5532</b>	0.5205	0.5371	<b>0.5602</b>	0.5214	0.5372	<b>0.5587</b>
GBVS'07	0.6134	<b>0.6343</b>	<b>0.6343</b>	0.5671	<b>0.6124</b>	0.6101	0.5670	<b>0.6125</b>	0.6103
PFT'07	0.6568	<b>0.6650</b>	0.6649	0.6225	<b>0.6461</b>	0.6440	0.6224	<b>0.6456</b>	0.6439
IK'98	0.6480	0.6565	<b>0.6570</b>	0.6134	0.6372	<b>0.6407</b>	0.6131	0.6365	<b>0.6409</b>

Table C.5.: Color space decorrelation results as quantified by the AUC evaluation measure on the Judd/MIT dataset. This table contains color coded information and is best seen in color. Please refer to Tab. 3.3 for a color legend.

CC Method	RGB			Lab			ICOPP		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.2180	<b>0.2363</b>	<b>0.2389</b>	0.2317	<b>0.2404</b>	0.2343	0.2233	<b>0.2390</b>	<b>0.2335</b>
QDCT	0.2206	<b>0.2350</b>	0.2332	0.2345	<b>0.2351</b>	0.2349	0.2328	0.2338	<b>0.2340</b>
EPQFT	0.1918	0.2105	<b>0.2124</b>	0.2101	0.2097	<b>0.2114</b>	0.2050	0.2071	<b>0.2075</b>
DCT'11	0.2084	<b>0.2394</b>	0.2313	0.2336	<b>0.2389</b>	0.2372	0.2339	<b>0.2391</b>	0.2368
AC'09	0.0291	0.0494	<b>0.0873</b>	0.0490	<b>0.0504</b>	<b>0.0592</b>	0.0410	0.0469	<b>0.0566</b>
GBVS'07	0.1743	<b>0.2372</b>	0.2360	0.2203	<b>0.2393</b>	0.2389	0.2135	<b>0.2393</b>	0.2384
PFT'07	0.1860	<b>0.2195</b>	0.2132	0.2137	<b>0.2188</b>	0.2176	0.2135	<b>0.2193</b>	0.2181
IK'98	0.2616	0.2907	<b>0.2908</b>	0.2841	<b>0.2916</b>	0.2910	0.2832	<b>0.2903</b>	0.2885

CC Method	GAUSS			XYZ			CAT02LMS		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.2365	<b>0.2372</b>	0.2360	0.1785	<b>0.2091</b>	<b>0.2101</b>	0.1785	<b>0.2079</b>	<b>0.2101</b>
QDCT	0.2317	<b>0.2344</b>	<b>0.2344</b>	0.1822	<b>0.2015</b>	<b>0.2100</b>	0.1830	<b>0.2009</b>	<b>0.2130</b>
EPQFT	0.2070	<b>0.2086</b>	<b>0.2107</b>	0.1586	<b>0.1789</b>	<b>0.1915</b>	0.1593	<b>0.1781</b>	<b>0.1956</b>
DCT'11	0.2310	<b>0.2391</b>	0.2384	0.1753	<b>0.2152</b>	0.2107	0.1761	<b>0.2150</b>	0.2110
AC'09	0.0421	0.0500	<b>0.0514</b>	0.0203	<b>0.0304</b>	<b>0.0644</b>	0.0208	<b>0.0299</b>	<b>0.0664</b>
GBVS'07	0.2102	0.2375	<b>0.2378</b>	0.1393	<b>0.2108</b>	0.2063	0.1395	<b>0.2104</b>	0.2064
PFT'07	0.2109	<b>0.2196</b>	0.2193	0.1557	<b>0.1984</b>	0.1950	0.1564	<b>0.1978</b>	0.1952
IK'98	0.2825	0.2901	<b>0.2909</b>	0.2259	<b>0.2626</b>	<b>0.2694</b>	0.2266	<b>0.2619</b>	<b>0.2700</b>

Table C.6.: Color space decorrelation results as quantified by the CC evaluation measure on the Judd/MIT dataset. This table contains color coded information and is best seen in color. Please refer to Tab. 3.3 for a color legend.

NSS Method	RGB			Lab			ICOPP		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH <sup>12</sup>	0.8736	<b>0.9543</b>	<b>0.9661</b>	0.9357	<b>0.9707</b>	0.9464	0.9033	<b>0.9648</b>	0.9438
QDCT	0.8895	<b>0.9469</b>	0.9392	0.9444	<b>0.9472</b>	0.9464	0.9380	0.9426	<b>0.9432</b>
EPQFT	0.7780	0.8535	<b>0.8599</b>	0.8506	0.8499	<b>0.8565</b>	0.8308	0.8401	<b>0.8409</b>
DCT <sup>11</sup>	0.8397	<b>0.9632</b>	0.9304	0.9395	<b>0.9611</b>	0.9539	0.9413	<b>0.9625</b>	0.9530
AC <sup>09</sup>	0.1259	0.2130	<b>0.3703</b>	0.2099	0.2167	<b>0.2528</b>	0.1792	0.2031	<b>0.2418</b>
GBVS <sup>07</sup>	0.6939	<b>0.9460</b>	0.9410	0.8769	<b>0.9538</b>	0.9526	0.8517	<b>0.9545</b>	0.9511
PFT <sup>07</sup>	0.7539	<b>0.8882</b>	0.8624	0.8642	<b>0.8852</b>	0.8801	0.8649	<b>0.8875</b>	0.8830
IK <sup>98</sup>	1.0391	1.1531	<b>1.1540</b>	1.1260	<b>1.1573</b>	1.1544	1.1236	<b>1.1523</b>	1.1450
NSS Method	GAUSS			XYZ			CAT02LMS		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH <sup>12</sup>	0.9533	<b>0.9575</b>	0.9532	0.7153	0.8457	<b>0.8516</b>	0.7148	0.8408	<b>0.8516</b>
QDCT	0.9330	<b>0.9445</b>	0.9443	0.7348	0.8147	<b>0.8509</b>	0.7380	0.8122	<b>0.8635</b>
EPQFT	0.8383	0.8452	<b>0.8539</b>	0.6432	0.7273	<b>0.7800</b>	0.6461	0.7238	<b>0.7975</b>
DCT <sup>11</sup>	0.9293	<b>0.9623</b>	0.9588	0.7060	<b>0.8705</b>	0.8531	0.7089	<b>0.8694</b>	0.8543
AC <sup>09</sup>	0.1820	0.2150	<b>0.2214</b>	0.0881	0.1320	<b>0.2785</b>	0.0899	0.1298	<b>0.2862</b>
GBVS <sup>07</sup>	0.8374	0.9470	<b>0.9484</b>	0.5549	<b>0.8440</b>	0.8256	0.5555	<b>0.8425</b>	0.8258
PFT <sup>07</sup>	0.8538	<b>0.8884</b>	0.8872	0.6312	<b>0.8075</b>	0.7937	0.6336	<b>0.8049</b>	0.7951
IK <sup>98</sup>	1.1211	1.1509	<b>1.1543</b>	0.8969	1.0446	<b>1.0731</b>	0.8994	1.0419	<b>1.0756</b>

Table C.7.: Color space decorrelation results as quantified by the NSS evaluation measure on the Judd/MIT dataset. This table contains color coded information and is best seen in color. Please refer to Tab. 3.3 for a color legend.



AUC Method	RGB		Lab		ICOPP				
	raw	PCA	PCA	ZCA	raw	PCA	ZCA		
CCH'12	0.5838	<b>0.6030</b>	<b>0.6045</b>	0.6018	<b>0.6043</b>	<b>0.6037</b>	0.6027	0.6040	<b>0.6042</b>
QDCT	0.5974	0.6068	<b>0.6148</b>	0.6041	0.6049	<b>0.6088</b>	0.6045	0.6069	<b>0.6092</b>
EPQFT	0.5955	0.6050	<b>0.6140</b>	0.6021	0.6032	<b>0.6069</b>	0.6016	0.6050	<b>0.6070</b>
DCT'11	0.5891	<b>0.6148</b>	0.6143	0.6063	0.6126	<b>0.6147</b>	0.6074	0.6134	<b>0.6173</b>
AC'09	0.5415	0.5509	<b>0.5633</b>	0.5464	0.5487	<b>0.5544</b>	0.5463	0.5488	<b>0.5534</b>
GBVS'07	0.5584	<b>0.5897</b>	0.5879	0.5788	<b>0.5914</b>	0.5906	0.5764	<b>0.5912</b>	0.5901
PFT'07	0.5936	<b>0.6180</b>	0.6147	0.6087	0.6157	<b>0.6172</b>	0.6100	0.6159	<b>0.6190</b>
IK'98	0.5740	0.5951	<b>0.5965</b>	0.5882	0.5936	<b>0.5950</b>	0.5881	0.5943	<b>0.5966</b>

AUC Method	GAUSS		XYZ		CAT02LMS				
	raw	PCA	PCA	ZCA	raw	PCA	ZCA		
CCH'12	0.6025	<b>0.6037</b>	<b>0.6041</b>	0.5684	0.5834	<b>0.5862</b>	0.5682	0.5832	<b>0.5861</b>
QDCT	0.6050	0.6075	<b>0.6091</b>	0.5821	0.5960	<b>0.6047</b>	0.5829	0.5958	<b>0.6057</b>
EPQFT	0.6037	0.6064	<b>0.6068</b>	0.5820	0.5927	<b>0.6040</b>	0.5827	0.5928	<b>0.6045</b>
DCT'11	0.6058	<b>0.6161</b>	0.6155	0.5786	<b>0.6063</b>	0.6050	0.5793	<b>0.6069</b>	0.6065
AC'09	0.5484	0.5515	<b>0.5523</b>	0.5343	0.5436	<b>0.5522</b>	0.5340	0.5424	<b>0.5541</b>
GBVS'07	0.5736	0.5893	<b>0.5898</b>	0.5540	<b>0.5843</b>	0.5804	0.5539	<b>0.5846</b>	0.5805
PFT'07	0.6092	<b>0.6185</b>	0.6172	0.5816	<b>0.6092</b>	0.6072	0.5826	<b>0.6095</b>	0.6068
IK'98	0.5886	0.5957	<b>0.5959</b>	0.5643	0.5864	<b>0.5893</b>	0.5645	0.5867	<b>0.5890</b>

Table C.8.: Color space decorrelation results as quantified by the AUC evaluation measure on the Kootstra dataset. This table contains color coded information and is best seen in color. Please refer to Tab. 3.3 for a color legend.

CC Method	RGB			Lab			ICOPP		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.2193	<b>0.2204</b>	0.2194	0.2195	<b>0.2307</b>	0.2147	0.2130	<b>0.2197</b>	0.2120
QDCT	0.2449	0.2636	<b>0.2714</b>	0.2565	0.2599	<b>0.2664</b>	0.2574	0.2633	<b>0.2682</b>
EPQFT	0.2008	0.2209	<b>0.2333</b>	0.2138	0.2167	<b>0.2248</b>	0.2113	0.2200	<b>0.2248</b>
DCT'11	0.2300	<b>0.2766</b>	0.2752	0.2553	0.2722	<b>0.2760</b>	0.2598	0.2766	<b>0.2796</b>
AC'09	0.0610	0.0704	<b>0.0786</b>	0.0654	0.0681	<b>0.0724</b>	0.0609	<b>0.0703</b>	0.0694
GBVS'07	0.2201	<b>0.2844</b>	0.2798	0.2622	<b>0.2852</b>	<b>0.2852</b>	0.2590	<b>0.2868</b>	0.2847
PFT'07	0.1965	<b>0.2427</b>	0.2401	0.2203	0.2365	<b>0.2402</b>	0.2225	0.2412	<b>0.2445</b>
IK'98	0.2790	0.3136	<b>0.3203</b>	0.3012	0.3141	<b>0.3172</b>	0.2998	0.3130	<b>0.3179</b>
CC Method	GAUSS			XYZ			CAT02LMS		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	<b>0.2195</b>	0.2175	0.2146	0.1800	0.1880	<b>0.1895</b>	0.1789	0.1879	0.1911
QDCT	0.2614	<b>0.2665</b>	<b>0.2679</b>	0.2088	<b>0.2310</b>	<b>0.2445</b>	0.2106	0.2308	<b>0.2466</b>
EPQFT	0.2207	<b>0.2254</b>	0.2246	0.1712	0.1917	<b>0.2124</b>	0.1726	0.1918	<b>0.2144</b>
DCT'11	0.2601	<b>0.2787</b>	0.2774	0.2035	<b>0.2518</b>	0.2513	0.2045	<b>0.2521</b>	0.2513
AC'09	0.0699	0.0689	<b>0.0704</b>	0.0574	0.0577	<b>0.0713</b>	0.0559	0.0571	<b>0.0722</b>
GBVS'07	0.2489	0.2830	<b>0.2853</b>	0.1922	<b>0.2609</b>	0.2533	0.1922	<b>0.2619</b>	0.2517
PFT'07	0.2270	<b>0.2436</b>	0.2418	0.1697	<b>0.2238</b>	0.2218	0.1711	<b>0.2242</b>	0.2217
IK'98	0.3050	0.3158	<b>0.3162</b>	0.2515	0.2907	<b>0.2960</b>	0.2535	0.2906	<b>0.2952</b>

Table C.9.: Color space decorrelation results as quantified by the CC evaluation measure on the Kootstra dataset. This table contains color coded information and is best seen in color. Please refer to Tab. 3.3 for a color legend.

NSS Method	RGB			Lab			ICOPP		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.4783	<b>0.4996</b>	0.4991	0.4972	<b>0.5219</b>	0.4863	0.4851	<b>0.4993</b>	0.4836
QDCT	0.5268	0.5760	<b>0.6042</b>	0.5588	0.5666	<b>0.5846</b>	0.5592	0.5771	<b>0.5911</b>
EPQFT	0.4384	0.4895	<b>0.5282</b>	0.4737	0.4801	<b>0.5010</b>	0.4644	0.4887	<b>0.5024</b>
DCT'11	0.4909	<b>0.6153</b>	0.6132	0.5562	0.6027	<b>0.6116</b>	0.5671	0.6156	<b>0.6221</b>
AC'09	0.1390	0.1650	<b>0.2014</b>	0.1561	0.1629	<b>0.1726</b>	0.1439	<b>0.1682</b>	0.1662
GBVS'07	0.4580	<b>0.6099</b>	0.5980	0.5550	0.6094	<b>0.6097</b>	0.5509	<b>0.6147</b>	0.6098
PFT'07	0.4273	<b>0.5526</b>	0.5460	0.4918	0.5364	<b>0.5442</b>	0.4978	0.5490	<b>0.5575</b>
IK'98	0.5851	0.6690	<b>0.6892</b>	0.6407	0.6708	<b>0.6800</b>	0.6374	0.6691	<b>0.6828</b>

NSS Method	GAUSS			XYZ			CAT02LMS		
	raw	PCA	ZCA	raw	PCA	ZCA	raw	PCA	ZCA
CCH'12	0.4916	<b>0.4942</b>	0.4875	0.3877	0.4195	<b>0.4236</b>	0.3854	0.4194	<b>0.4273</b>
QDCT	0.5684	0.5831	<b>0.5886</b>	0.4433	0.4981	<b>0.5382</b>	0.4477	0.4974	<b>0.5415</b>
EPQFT	0.4871	0.4995	<b>0.5001</b>	0.3694	0.4183	<b>0.4738</b>	0.3729	0.4185	<b>0.4760</b>
DCT'11	0.5664	<b>0.6193</b>	0.6176	0.4306	<b>0.5549</b>	0.5543	0.4332	<b>0.5553</b>	0.5545
AC'09	0.1635	0.1629	<b>0.1663</b>	0.1225	0.1271	<b>0.1716</b>	0.1190	0.1254	<b>0.1733</b>
GBVS'07	0.5230	0.6066	<b>0.6120</b>	0.3992	<b>0.5566</b>	0.5398	0.3989	<b>0.5589</b>	0.5357
PFT'07	0.5043	<b>0.5546</b>	0.5512	0.3658	<b>0.5040</b>	0.4990	0.3696	<b>0.5051</b>	0.4984
IK'98	0.6458	0.6735	<b>0.6777</b>	0.5245	0.6183	<b>0.6323</b>	0.5291	0.6187	<b>0.6303</b>

Table C.10.: Color space decorrelation results as quantified by the NSS evaluation measure on the Kootstra dataset. This table contains color coded information and is best seen in color. Please refer to Tab. 3.3 for a color legend.



# D

## Center Bias Integration Methods

In the main evaluation, see Sec. 4.2.3.B, we present the results that we achieved with a convex combination in Eq. 4.1 and Eq. 4.8. However, we have considered and evaluated alternative integration methods.

To investigate the question how good other combination types are, we tested the minimum, maximum, and product as alternative combinations. To account for the influence of different value distributions within the normalized value range, we also weighted the input of the min and max operation (e.g.,  $S_P^{\min} = \min(w_C S_C, w_B S_B)$ ). The results of the algorithms using different combination types are shown in Tab. D.1. The presented results are the results that we achieve with the center bias weight that results in the highest  $F_1$  score.

In Tab. D.1, we can see that the linear combination is clearly the best choice for LDRC+CB. It is interesting to note that LDRC+CB with the product as combination achieves similar results to RC'10. However, LDRC+CB remains the algorithm that provides the best performance in terms of  $F_1$  score and  $F_\beta$  score whereas RC'10+CB provides the best performance in terms of PHR. Interestingly, LDRC+CB and RC'10+CB achieve a nearly identical  $\int$ ROC.

Method	Combination	$F_1$	$F_\beta$	$\int$ ROC	PHR
LDRC+CB	Linear/Convex	<u>0.8034</u>	<u>0.8183</u>	<u>0.9624</u>	0.9240
LDRC+CB	Max	0.7504	0.7561	0.9422	0.8630
LDRC+CB	Min	0.7897	0.8049	0.9535	0.8880
LDRC+CB	Product	0.7883	0.8024	0.9578	0.9130
RC'10+CB	Linear/Convex	0.7973	0.8120	0.9620	0.9340
RC'10+CB	Max	0.7855	0.7993	0.9568	0.9140
RC'10+CB	Min	0.7962	0.8150	0.9603	0.9180
RC'10+CB	Product	0.7974	0.8136	<u>0.9623</u>	<u>0.9460</u>
CB <sub>S</sub>	–	0.5793	0.5764	0.8623	0.6980
CB <sub>P</sub>	–	0.5604	0.5452	0.8673	0.7120

Table D.1.: Salient object detection results that we obtain using different center bias integration types. Please compare to the results in Tab. 4.1, Sec. 4.2.



# E

## Who's Waldo?

“Where’s Waldo?”<sup>1</sup> is a book series for children created by the British illustrator Martin Handford. The books consist of illustrations that depict groups of people doing various things, see Fig. E.1(b), and the reader’s task is to find Waldo, who is hidden in the crowd. Waldo always wears very distinctive clothing: a red-and-white-striped shirt, bobble hat, and glasses, see Fig. E.1(a). To make it more interesting, the illustrations often contain distractors with similar features, which – as you should understand after having read this thesis – makes it harder for the reader to find Waldo.

<sup>1</sup>Also known as “Where’s Wally” outside of North America. Waldo’s German name is Walter.

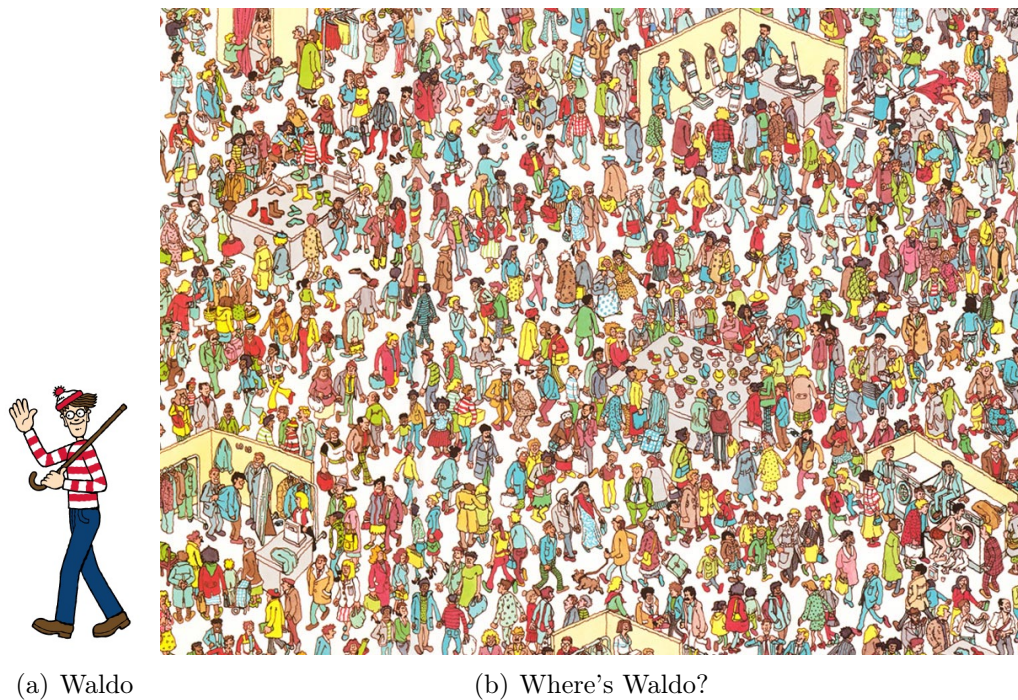


Figure E.1.: Who and where is Waldo? This illustration is best viewed in color. TM & © 2008 Entertainment Rights Distribution Limited. All rights reserved.





# Publications

- [JSF12] H. Jaspers, B. Schauerte, and G. A. Fink, “Sift-based camera localization using reference objects for application in multi-camera environments and robotics,” in *Proc. 1st International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, Vilamoura, Algarve, Portugal, February 2012.
- [KSKS12] B. Kühn, B. Schauerte, K. Kroschel, and R. Stiefelhagen, “Multimodal saliency-based attention: A lazy robot’s approach,” in *Proc. 25th International Conference on Intelligent Robots and Systems (IROS)*. Vilamoura, Algarve, Portugal: IEEE/RSJ, October 2012.
- [KSS13] D. Koester, B. Schauerte, and R. Stiefelhagen, “Accessible section detection for visual guidance,” in *Proc. IEEE/NSF Workshop on Multimodal and Alternative Perception for Visually Impaired People (MAP4VIP)*, San Jose, CA, USA, July 2013.
- [KSSK12] B. Kühn, B. Schauerte, R. Stiefelhagen, and K. Kroschel, “A modular audio-visual scene analysis and attention system for humanoid robots,” in *Proc. 43rd International Symposium on Robotics (ISR)*, Taipei, Taiwan, August 2012.
- [MCS<sup>+</sup>14] M. Martinez, A. Constantinescu, B. Schauerte, D. Koester, and R. Stiefelhagen, “Cognitive evaluation of haptic and audio feedback in short range navigation tasks,” in *Proc. 14th Int. Conf. Computers Helping People with Special Needs (ICCHP)*. Paris, France: Springer, July 2014.
- [MSS13] M. Martinez, B. Schauerte, and R. Stiefelhagen, “BAM! Depth-based body analysis in critical care,” in *Proc. 15th International Conference on Computer Analysis of Images and Patterns (CAIP)*. York, UK: Springer, August 2013.
- [RSAHS14] L. Rybok, B. Schauerte, Z. Al-Halah, and R. Stiefelhagen, “Important stuff, everywhere! Activity recognition with salient proto-objects as context,” in *Proc. 14th IEEE Winter Conference on Applications of Computer Vision (WACV)*, Steamboat Springs, CO, USA, March 2014.
- [SF10a] B. Schauerte and G. A. Fink, “Focusing computational visual attention in multi-modal human-robot interaction,” in *Proc. 12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI)*. Beijing, China: ACM, November 2010.
- [SF10b] B. Schauerte and G. A. Fink, “Web-based learning of naturalized color models for human-machine interaction,” in *Proc. 12th International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. Sydney, Australia: IEEE, December 2010.
- [SKKS11] B. Schauerte, B. Kühn, K. Kroschel, and R. Stiefelhagen, “Multimodal saliency-based attention for object-based scene analysis,” in *Proc. 24th International Conference on Intelligent Robots and Systems (IROS)*. San Francisco, CA, USA: IEEE/RSJ, September 2011.

- [SKMS14] B. Schauerte, D. Koester, M. Martinez, and R. Stiefelhagen, "Way to Go! Detecting open areas ahead of a walking person," in *ECCV Workshop on Assistive Computer Vision and Robotics (ACVR)*. Springer, 2014.
- [SMCS12] B. Schauerte, M. Martinez, A. Constantinescu, and R. Stiefelhagen, "An assistive vision system for the blind that helps find lost things," in *Proc. 13th International Conference on Computers Helping People with Special Needs (ICCHP)*. Linz, Austria: Springer, July 2012.
- [SPF09] B. Schauerte, T. Plötz, and G. A. Fink, "A multi-modal attention system for smart environments," in *Proc. 7th International Conference on Computer Vision Systems (ICVS)*, ser. Lecture Notes in Computer Science, vol. 5815. Liège, Belgium: Springer, October 2009.
- [SRF10] B. Schauerte, J. Richarz, and G. A. Fink, "Saliency-based identification and recognition of pointed-at objects," in *Proc. 23rd International Conference on Intelligent Robots and Systems (IROS)*. Taipei, Taiwan: IEEE/RSJ, October 2010.
- [SRP<sup>+</sup>09] B. Schauerte, J. Richarz, T. Plötz, C. Thureau, and G. A. Fink, "Multi-modal and multi-camera attention in smart environments," in *Proc. 11th International Conference on Multimodal Interfaces (ICMI)*. Cambridge, MA, USA: ACM, November 2009.
- [SS12a] B. Schauerte and R. Stiefelhagen, "Learning robust color name models from web images," in *Proc. 21st International Conference on Pattern Recognition (ICPR)*. Tsukuba, Japan: IEEE, November 2012.
- [SS12b] B. Schauerte and R. Stiefelhagen, "Predicting human gaze using quaternion DCT image signature saliency and face detection," in *Proc. IEEE Workshop on the Applications of Computer Vision (WACV)*. Breckenridge, CO, USA: IEEE, January 2012.
- [SS12c] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Proc. 12th European Conference on Computer Vision (ECCV)*. Firenze, Italy: Springer, October 2012.
- [SS13a] B. Schauerte and R. Stiefelhagen, "How the distribution of salient objects in images influences salient object detection," in *Proc. 20th International Conference on Image Processing (ICIP)*. Melbourne, Australia: IEEE, September 2013.
- [SS13b] B. Schauerte and R. Stiefelhagen, "Wow! Bayesian surprise for salient acoustic event detection," in *Proc. 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, Canada: IEEE, May 2013.
- [SS14] B. Schauerte and R. Stiefelhagen, "Look at this! Learning to guide visual saliency in human-robot interaction," in *Proc. International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ, 2014.
- [SSS14] T. Schneider, B. Schauerte, and R. Stiefelhagen, "Manifold alignment for person independent appearance-based gaze estimation," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*. Stockholm, Sweden: IEEE, August 2014.

# Bibliography

- [3M] 3M, “3M visual attention service,” [http://solutions.3m.com/wps/portal/3M/en\\_US/VAS-NA?MDR=true](http://solutions.3m.com/wps/portal/3M/en_US/VAS-NA?MDR=true).
- [ABGA04] S. R. Arnott, M. A. Binns, C. L. Grady, and C. Alain, “Assessing the auditory dual-pathway model in humans,” *Neuroimage*, vol. 22, pp. 401–408, 2004.
- [ADF10] B. Alexe, T. Deselaers, and V. Ferrari, “What is an object?” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2010, pp. 73–80.
- [AHES09] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, “Frequency-tuned Salient Region Detection,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2009.
- [AHW<sup>+</sup>10] A. Andreopoulos, S. Hasler, H. Wersing, H. Janssen, J. Tsotsos, and E. Körner, “Active 3D object localization using a humanoid robot,” *IEEE Trans. Robotics*, pp. 47–64, 2010.
- [All96] R. E. Alley, *Algorithm Theoretical Basis Document for Decorrelation Stretch*. NASA, JPL, 1996.
- [ARA<sup>+</sup>06] T. Asfour, K. Regenstein, P. Azad, J. Schröder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann, “ARMAR-III: An integrated humanoid platform for sensory-motor control,” in *Humanoids*, 2006.
- [AS07] S. Avidan and A. Shamir, “Seam carving for content-aware image resizing,” *ACM Trans. Graph.*, vol. 26, no. 3, Jul. 2007.
- [AS10] R. Achanta and S. Süsstrunk, “Saliency detection using maximum symmetric surround,” in *Proc. Int. Conf. Image Process.*, 2010.
- [AS13] A. Alsam and P. Sharma, “A robust metric for the evaluation of visual saliency algorithms,” *Journal of the Optical Society of America*, 2013.
- [AWA<sup>+</sup>08] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, “The Karlsruhe Humanoid Head,” in *Humanoids*, 2008.
- [AWB88] Y. Aloimonos, I. Weiss, and A. Bandopadhyay, “Active vision,” *International Journal of Computer Vision*, vol. 1, no. 4, pp. 333–356, 1988.
- [Ban04] A. Bangertner, “Using pointing and describing to achieve joint focus of attention in dialogue,” *Psychological Science*, vol. 15, no. 6, pp. 415–419, 2004.
- [Bar61] H. Barlow, “Possible principles underlying the transformation of sensory messages,” *Sensory Communication*, pp. 217–234, 1961.
- [Bat81] E. Batschelet, *Circular statistics in biology*, 1981, vol. 24, no. 4.
- [BC98] R. Beun and A. Cremers, “Object reference in a shared domain of conversation,” *Pragmatics and Cognition*, vol. 1, no. 6, pp. 111–142, 1998.
- [BDT08] S. Bock, P. Dicke, and P. Thier, “How precise is gaze following in humans?” *Vision Research*, vol. 48, pp. 946–957, 2008.

- [Ber76] J. M. Bernardo, "Algorithm as 103 psi(digamma function) computation," *Applied Statistics*, vol. 25, pp. 315–317, 1976.
- [BG83] G. Buchsbaum and A. Gottschalk, "Trichromacy, opponent colours coding and optimum colour information transmission in the retina," *Proceedings of the Royal Society*, vol. B, no. 220, pp. 89–113, 1983.
- [BH05] C. Bundesen and T. Habekost, *Handbook of Cognition*. Sage Publications, 2005, ch. Attention.
- [BHCS07] P. Becouze, C. Hann, J. Chase, and G. Shaw, "Measuring facial grimacing for quantifying patient agitation in critical care," in *Computer Methods and Programs in Biomedicine*, 2007.
- [BI00] G. Butterworth and S. Itakura, "How the eyes, head and hand serve definite reference," *Br. J. Dev. Psychol.*, vol. 18, pp. 25–50, 2000.
- [BI13] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 185–207, 2013.
- [BJ80] A. K. Bera and C. M. Jarque, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Economics Letters*, vol. 5, no. 3, pp. 255–259, 1980.
- [BK69] B. Berlin and P. Kay, *Basic color terms: their universality and evolution*. University of California Press, 1969.
- [BK11] M. Begum and F. Karray, "Integrating visual exploration and visual search in robotic visual attention: The role of human-robot interaction," in *Proc. Int. Conf. Robot. Autom.*, 2011.
- [BKMG10] M. Begum, F. Karray, G. K. I. Mann, and R. G. Gosine, "A probabilistic model of overt visual attention for cognitive robots," *IEEE Trans. Syst., Man, Cybern. B*, vol. 40, pp. 1305–1318, 2010.
- [BO06] A. Bangerter and D. M. Oppenheimer, "Accuracy in detecting referents of pointing gestures unaccompanied by language," *Gesture*, vol. 6, no. 1, pp. 85–102, 2006.
- [Bor12] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2012.
- [BP94] S. Baluja and D. Pomerleau, "Non-Intrusive Gaze Tracking Using Artificial Neural Networks." in *Advances in Neural Information Processing Systems*, 1994.
- [Bre] C. Breazeal, "MIT AI – Sociable machines – Ongoing research," <http://www.ai.mit.edu/projects/sociable/ongoing-research.html>.
- [Bre90] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sounds*. MIT Press, 1990.
- [Bri95] E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," *Comp. Ling.*, vol. 21, no. 4, pp. 543–565, 1995.

- [Bro07] A. G. Brooks, “Coordinating human-robot communication,” Ph.D. dissertation, MIT, 2007.
- [BS97] A. J. Bell and T. J. Sejnowski, “The independent components of scenes are edge filters,” *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.
- [BS99] C. Breazeal and B. Scassellati, “A context-dependent attention system for a social robot,” in *Proc. Int. Joint Conf. Artif. Intell.*, 1999.
- [BSF11] M. Brown, S. Susstrunk, and P. Fua, “Spatio-chromatic decorrelation by shift-invariant filtering,” in *CVPR Workshop*, 2011.
- [BSI12] A. Borji, D. N. Sihite, and L. Itti, “Salient object detection: A benchmark,” in *Proc. European Conf. Comp. Vis.*, 2012.
- [BSI13a] A. Borji, D. Sihite, and L. Itti, “What/where to look next? modeling top-down visual attention in complex interactive environments,” *IEEE Trans. Syst., Man, Cybern. A*, no. 99, 2013.
- [BSI13b] A. Borji, D. N. Sihite, and L. Itti, “Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study,” *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, 2013.
- [BT09] N. Bruce and J. Tsotsos, “Saliency, attention, and visual search: An information theoretic approach,” *Journal of Vision*, vol. 9, no. 3, pp. 1–24, 2009.
- [Bus35] G. T. Busswell, *How people look at pictures: A study of the psychology of perception in art*. University of Chicago Press, 1935.
- [BZ09] P. Bian and L. Zhang, “Biological plausibility of spectral domain approach for spatiotemporal visual saliency,” in *Proc. Ann. Conf. Neural Inf. Process. Syst.*, 2009.
- [BZCM08] N. Butko, L. Zhang, G. Cottrell, and J. R. Movellan, “Visual saliency model for robot cameras,” in *Proc. Int. Conf. Robot. Autom.*, 2008.
- [CB10] B. D. Coensel and D. Botteldooren, “A model of saliency-based auditory attention to environmental sound,” in *Proc. Int. Congress on Acoustics*, 2010.
- [CBB<sup>+</sup>97] G. A. Calvert, E. Bullmore, M. Brammer, R. Campbell, S. C. Williams, P. K. McGuire, P. W. Woodruff, S. D. Iversen, and A. S. David, “Activation of auditory cortex during silent lipreading,” *Science*, vol. 276, pp. 593–596, 1997.
- [CC03] C. Cashon and L. Cohen, *The construction, deconstruction, and reconstruction of infant face perception*. NOVA Science Publishers, 2003, ch. The development of face processing in infancy and early childhood: Current perspectives, pp. 55–68.
- [CFK08] M. Cerf, P. Frady, and C. Koch, “Subjects’ inability to avoid looking at faces suggests bottom-up attention allocation mechanism for faces,” in *Proc. Soc. Neurosci.*, 2008.

- [CFK09] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *Journal of Vision*, vol. 9, 2009.
- [Che08] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, pp. 975–979, 2008.
- [CHEK07] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Proc. Ann. Conf. Neural Inf. Process. Syst.*, 2007.
- [CLE] CLEAR2007, "Classification of events, activities and relationships evaluation and workshop," <http://www.clear-evaluation.org>.
- [CLR90] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*. MIT Press and McGraw-Hill, 1990.
- [CM02] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 603–619, 2002.
- [Cox64] R. T. Cox, "Probability, frequency, and reasonable expectation," *American Journal of Physics*, vol. 14, pp. 1–13, 1964.
- [CSB83] H. H. Clark, R. Schreuder, and S. Buttrick, "Common ground and the understanding of demonstrative reference," *Journal of Verbal Learning and Verbal Behavior*, no. 22, pp. 245–258, 1983.
- [CXF<sup>+</sup>03] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou, "A visual attention model for adapting images on small displays," *Multimedia Systems*, vol. 9, no. 4, pp. 353–64, 2003.
- [CZM<sup>+</sup>11] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2011.
- [DBZ07] A. Dankers, N. Barnes, and A. Zelinsky, "A reactive vision system: Active-dynamic saliency," in *Proc. Int. Conf. Vis. Syst.*, 2007.
- [DEHR07] P. H. Delano, D. Elgueda, C. M. Hamame, and L. Robles, "Selective attention to visual stimuli reduces cochlear sensitivity in chinchillas," *Journal of Neuroscience*, vol. 27, pp. 4146–4153, 2007.
- [Dom13] J. Domke, "Learning graphical model parameters with approximate marginal inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2454–2467, 2013.
- [DSB01] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Robust localization in reverberant rooms*. Springer, 2001, ch. 8, pp. 157–180.
- [DSCM07] L. De Santis, S. Clarke, and M. M. Murray, "Automatic and intrinsic auditory what and where processing in humans revealed by electrical neuroimaging," *Cereb Cortex*, vol. 17, pp. 9–17, 2007.
- [DSHB11] D. Droschel, J. Stückler, D. Holz, and S. Behnke, "Towards joint attention for a domestic service robot - person awareness and gesture

- recognition using time-of-flight cameras,” in *Proc. Int. Conf. Robot. Autom.*, 2011.
- [DSMS02] V. Dragoi, J. Sharma, E. K. Miller, and M. Sur, “Dynamics of neuronal sensitivity in visual cortex and local feature discrimination,” *Nature Neuroscience*, pp. 883–891, 2002.
- [DSS06] M. Doniec, G. Sun, and B. Scassellati, “Active learning of joint attention,” in *Humanoids*, 2006.
- [DT05] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2005.
- [Dun84] J. Duncan, “Selective attention and the organization of visual information,” *Journal of Experimental Psychology: General*, vol. 113, no. 4, pp. 501–517, 1984.
- [DWM<sup>+</sup>11] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, “Visual saliency detection by spatially weighted dissimilarity,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2011.
- [EDR94] R. Egly, J. Driver, and R. D. Rafal, “Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects,” *Journal of Experimental Psychology: General*, vol. 123, no. 2, 1994.
- [Ehr05] M. Ehrgott, *Multicriteria Optimization*. Springer, 2005.
- [EI08] L. Elazary and L. Itti, “Interesting objects are visually salient,” *Journal of Vision*, vol. 8, no. 3, pp. 1–15, 2008.
- [EIV07] A. Elmagarmid, P. Ipeirotis, and V. Verykios, “Duplicate record detection: A survey,” *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 1, pp. 1–16, 2007.
- [Ell93] T. Ell, “Quaternion-fourier transforms for analysis of two-dimensional linear time-invariant partial differential systems,” in *Int. Conf. Decision and Control*, 1993.
- [Enca] Encyclopaedia Britannica Online, “basilar membrane: analysis of sound frequencies,” <http://www.britannica.com/EBchecked/media/537/The-analysis-of-sound-frequencies-by-the-basilar-membrane>, retrieved 3 April 2014.
- [Encb] —, “cochlea: cross section,” <http://www.britannica.com/EBchecked/media/534/A-cross-section-through-one-of-the-turns-of-the>, retrieved 3 April 2014.
- [Encc] —, “ear: structure of the human ear,” <http://www.britannica.com/EBchecked/media/530/Structure-of-the-human-ear>, retrieved 3 April 2014.
- [ES07] T. Ell and S. Sangwine, “Hypercomplex fourier transforms of color images,” *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 22–35, 2007.
- [ESJ86] C. W. Eriksen and J. D. St. James, “Visual attention within and around the field of focal attention: A zoom lens model,” *Perception and Psychophysics*, vol. 40, no. 4, pp. 225–240, 1986.

- [ESP08] W. Einhäuser, M. Spain, and P. Perona, “Objects predict fixations better than early saliency,” *Journal of Vision*, vol. 8, no. 14, 2008.
- [Ess00] I. Essa, “Ubiquitous sensing for smart and aware environments,” *IEEE Personal Communications*, vol. 7, no. 5, pp. 47–49, 2000.
- [EVGW<sup>+</sup>] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results,” <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>.
- [Eye] EyeQuant, “The 3 most surprising insights from a 200 website eye-tracking study,” <http://blog.eyequant.com/2014/01/15/the-3-most-surprising-insights-from-a-200-website-eye-tracking-study/>, retrieved 3 April 2014.
- [FBH<sup>+</sup>08] M. E. Foster, E. G. Bard, R. L. Hill, M. Guhe, J. Oberlander, and A. Knoll, “The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue,” in *Proc. Int. Conf. Human-Robot Interaction*, 2008, pp. 295–302.
- [FBR05] S. Frintrop, G. Backer, and E. Rome, “Selecting what is important: Training visual attention,” in *Proc. KI*, 2005.
- [FE04] B. Fröba and A. Ernst, “Face detection with the modified census transform,” in *Proc. Int. Conf. Automatic Face and Gesture Rec.*, 2004.
- [FEDS07] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, “Auditory attention – focusing the searchlight on sound,” *Current Opinion in Neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.
- [FFFP03] L. Fei-Fei, R. Fergus, and P. Perona, “A bayesian approach to unsupervised one-shot learning of object categories,” in *Proc. Int. Conf. Comp. Vis.*, 2003.
- [FH04] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.
- [FH08] W. Feng and B. Hu, “Quaternion discrete cosine transform and its application in color template matching,” in *Int. Cong. Image and Signal Processing*, 2008, pp. 252–256.
- [FJ08] S. Frintrop and P. Jensfelt, “Attentional landmarks and active gaze control for visual slam,” *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1054–1065, 2008.
- [FND03] T. Fong, I. Nourbakhsh, and K. Dautenhahn, “A survey of socially interactive robots,” *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 143–166, 2003.
- [FPB06] K. A. Fleming, R. A. Peters II, and R. E. Bodenheimer, “Image mapping and visual attention on a sensory ego-sphere,” in *Proc. Int. Conf. Intell. Robots Syst.*, 2006.
- [Fra79] Francis, W. N., and Kucera, H., compiled by, “A standard corpus of present-day edited american english, for use with digital computers (brown),” 1964, 1971, 1979.



- 
- [FRC10] S. Frintrop, E. Rome, and H. I. Christensen, “Computational visual attention systems and their cognitive foundation: A survey,” *ACM Trans. Applied Perception*, vol. 7, no. 1, pp. 6:1–6:39, 2010.
- [Fri06] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, ser. Lecture Notes in Computer Science. Springer, 2006.
- [Fri11] —, “Towards attentive robots,” *Paladyn*, vol. 2, no. 2, pp. 64–70, 2011.
- [GAK<sup>+</sup>07] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. Bradski, P. Baumstarck, S. Chung, and A. Y. Ng, “Peripheral-foveal vision for real-time object recognition and tracking in video,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2007.
- [Gaz] GazeHawk, “Gazehawk – eye tracking for everyone,” <http://www.gazehawk.com/>.
- [GCAS<sup>+</sup>04] J. Geoffrey Chase, F. Agogue, C. Starfinger, Z. Lam, G. Shaw, A. Rudge, and H. Sirisena, “Quantifying agitation in sedated icu patients using digital imaging,” in *Computer Methods and Programs in Biomedicine*, 2004.
- [Gil00] D. Gillies, “The subjective theory,” in *Philosophical Theories of Probability*. Routledge, 2000, ch. 4.
- [GKW87] A. R. Gillespie, A. B. Kahle, and R. E. Walker, “Color enhancement of highly correlated images. II. channel ratio and chromaticity transformation techniques,” *Remote Sensing of Environment*, vol. 22, no. 3, pp. 343–365, 1987.
- [GMV08] D. Gao, V. Mahadevan, and N. Vasconcelos, “On the plausibility of the discriminant center-surround hypothesis for visual saliency,” *Journal of Vision*, vol. 8, no. 7, pp. 1–18, 2008.
- [GMZ08] C. Guo, Q. Ma, and L. Zhang, “Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2008.
- [Goo] Google, “Eye-tracking studies: more than meets the eye,” <http://googleblog.blogspot.de/2009/02/eye-tracking-studies-more-than-meets.html>.
- [GRK07] D. Gergle, C. P. Rosé, and R. E. Kraut, “Modeling the impact of shared visual information on collaborative reference,” in *Proc. Int. Conf. Human Factors Comput. Syst. (CHI)*, 2007, pp. 1543–1552.
- [GS06] A. A. Ghazanfar and C. E. Schroeder, “Is neocortex essentially multisensory?” *Trends in Cognitive Science*, vol. 10, pp. 278–285, 2006.
- [GSvdW03] J.-M. Geusebroek, A. Smeulders, and J. van de Weijer, “Fast anisotropic gauss filtering,” *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 938–943, 2003.
- [GvdBSG01] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts, “Color invariance,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1338–1350, 2001.

- [GZ10] C. Guo and L. Zhang, “A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression,” *IEEE Trans. Image Process.*, vol. 19, pp. 185–198, 2010.
- [GZMT10] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2010.
- [GZMT12] —, “Context-aware saliency detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.
- [HA04] V. J. Hodge and J. Austin, “A survey of outlier detection methodologies,” *Artificial Intelligence Review*, vol. 22, pp. 85–126, 2004.
- [Ham66] W. R. Hamilton, *Elements of Quaternions*. University of Dublin Press., 1866.
- [HB95] D. J. Heeger and J. R. Bergen, “Pyramid-based texture analysis/synthesis,” in *Proc. Ann. Conf. Special Interest Group on Graphics and Interactive Techniques*, 1995, pp. 229–238.
- [HB13] H. Hadizadeh and I. Bajic, “Saliency-aware video compression,” *IEEE Trans. Image Process.*, no. 99, 2013.
- [HBD75] T. Huang, J. Burnett, and A. Deczky, “The importance of phase in image processing filters,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23, no. 6, pp. 529–542, 1975.
- [HEA<sup>+</sup>05] B. Haslinger, P. Erhard, E. Altenmuller, U. Schroeder, H. Boecker, and A. O. Ceballos-Baumann, “Transmodal sensorimotor networks during action observation in professional pianists,” *Journal of Cognitive Neuroscience*, vol. 17, pp. 282–293, 2005.
- [Hen03] J. M. Henderson, “Human gaze control during real-world scene perception,” *Trends in Cognitive Science*, pp. 498–504, 2003.
- [Her64] E. Hering, *Outlines of a Theory of the Light Sense*. Harvard University Press, 1964.
- [HHK12] X. Hou, J. Harel, and C. Koch, “Image signature: Highlighting sparse salient regions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, 2012.
- [HJ10] D. W. Hansen and Q. Ji, “In the Eye of the Beholder: A Survey of Models for Eyes and Gaze,” *Trans. PAMI*, vol. 32, pp. 478–500, 2010.
- [HKM<sup>+</sup>09] M. Heracles, U. Körner, T. Michalke, G. Sagerer, J. Fritsch, and C. Goerick, “A dynamic attention system that reorients to unexpected motion in real-world traffic environments,” in *Proc. Int. Conf. Intell. Robots Syst.*, 2009.
- [HKP07] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Proc. Ann. Conf. Neural Inf. Process. Syst.*, 2007.
- [HL08] D. Hall and J. Linas, *Handbook of Multisensor Data Fusion: Theory and Practice*. CRC Press, 2008.

- [HO07] J. Hershey and P. Olsen, “Approximating the kullback leibler divergence between gaussian mixture models,” in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2007.
- [Hob05] R. Hobson, *Joint attention: Communication and other minds*. Oxford University Press, 2005, ch. What puts the jointness in joint attention?, pp. 185–204.
- [HPSJ56] R. Hernandez-Peon, H. Scherrer, and M. Jouvet, “Modification of electric activity in cochlear nucleus during attention in unanesthetized cats,” *Science*, vol. 123, pp. 331–332, 1956.
- [HRB<sup>+</sup>04] G. Heidemann, R. Rae, H. Bekel, I. Bax, and H. Ritter, “Integrating context-free and context-dependent attentional mechanisms for gestural object reference,” *Machine Vision and Applications*, vol. 16, no. 1, pp. 64–73, 2004.
- [HRM11] F. Huettig, J. Rommers, and A. S. Meyer, “Using the visual world paradigm to study language processing: A review and critical evaluation,” *Acta psychologica*, vol. 137, no. 2, pp. 151–171, 2011.
- [HSK<sup>+</sup>10] Y. Hato, S. Satake, T. Kanda, M. Imai, and N. Hagita, “Pointing to space: modeling of deictic interaction referring to regions,” in *Proc. Int. Conf. Human-Robot Interaction*, 2010, pp. 301–308.
- [HSL07] E. R. Hafter, A. Sarampalis, and P. Loui, *Auditory Perception of Sound Sources*. Springer, 2007, ch. Auditory attention and filters (review).
- [HTL10] M. J. Huiskes, B. Thomee, and M. S. Lew, “New trends and ideas in visual concept detection,” in *ACM International Conference on Multimedia Information Retrieval*, 2010.
- [HY09] J. Holsopple and S. Yang, “Designing a data fusion system using a top-down approach,” in *Proc. Int. Conf. Military Comm.*, 2009.
- [HZ07] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2007.
- [IB05] L. Itti and P. F. Baldi, “A principled approach to detecting surprising events in video,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2005.
- [IB06] —, “Bayesian surprise attracts human attention,” in *Proc. Ann. Conf. Neural Inf. Process. Syst.*, 2006.
- [IK00] L. Itti and C. Koch, “A saliency-based search mechanism for overt and covert shifts of visual attention,” *Vision Research*, vol. 40, no. 10-12, pp. 1489–1506, 2000.
- [IK01a] —, “Computational modelling of visual attention.” *Nature Reviews: Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.
- [IK01b] —, “Feature combination strategies for saliency-based visual attention systems,” *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

- [IKN98] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [Jay03] E. T. Jaynes, *Probability Theory. The Logic of Science*. Cambridge University Press, 2003.
- [JDT11] T. Judd, F. Durand, and A. Torralba, “Fixations on low-resolution images,” *Journal of Vision*, vol. 11, no. 4, 2011.
- [JDT12] T. Judd, F. Durand, and A. Torralba., “A benchmark of computational models of saliency to predict human fixations,” MIT, Tech. Rep., 2012.
- [JEDT09] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” in *Proc. Int. Conf. Comp. Vis.*, 2009.
- [JM97] D. Johnson and L. McGeoch, “The traveling salesman problem: A case study in local optimization,” *Local search in combinatorial optimization*, pp. 215–310, 1997.
- [JOvW<sup>+</sup>05] T. Jost, N. Ouerhani, R. von Wartburg, R. Mäuri, and H. Häugli, “Assessing the contribution of color in visual attention,” *Computer Vision and Image Understanding*, vol. 100, pp. 107–123, 2005.
- [JWY<sup>+</sup>11] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng, “Automatic salient object segmentation based on context and shape prior,” in *Proc. British Conf. Comp. Vis.*, 2011.
- [Kal09] O. Kalinli, “Biologically inspired auditory attention models with applications in speech and audio processing,” Ph.D. dissertation, University of Southern California, Los Angeles, CA, USA, 2009.
- [KB01] T. Kadir and M. Brady, “Saliency, scale and image description,” *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.
- [KBS<sup>+</sup>10] B. Kühn, A. Belkin, A. Swerdlow, T. Machmer, J. Beyerer, and K. Kroschel, “Knowledge-driven opto-acoustic scene analysis based on an object-oriented world modelling approach for humanoid robots,” in *Proc. 41st Int. Symp. Robotics and 6th German Conf. Robotics*, 2010.
- [KC06] P. Knoeferle and M. W. Crocker, “The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking,” *Cognitive Science*, no. 30, pp. 481–529, 2006.
- [KCP<sup>+</sup>13] E. A. Krause, R. Cantrell, E. Potapova, M. Zillich, and M. Scheutz, “Incrementally biasing visual search using natural language input,” in *Proc. Int. Conf. Autonomous Agents and Multi-Agent Systems*, 2013.
- [KF11] D. A. Klein and S. Frintrop, “Center-surround divergence of feature statistics for salient object detection,” in *Proc. Int. Conf. Comp. Vis.*, 2011.
- [KH06] F. Kaplan and V. Hafner, “The challenges of joint attention,” *Interaction Studies*, vol. 7, no. 2, pp. 135–169, 2006.
- [KJS<sup>+</sup>02] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen, “Visual fixation patterns during viewing of naturalistic social situations as predictors of

- social competence in individuals with autism,” *Arch. Gen. Psychiatry*, vol. 59, no. 9, pp. 809–816, 2002.
- [KLP<sup>+</sup>06] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth, “Deixis: How to determine demonstrated objects using a pointing cone,” in *Proc. Int. Gesture Workshop*, vol. 3881, 2006.
- [KN07] O. Kalinli and S. Narayanan, “A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech,” in *Proc. Ann. Conf. Int. Speech Communication Association*, 2007.
- [KN09] —, “Prominence detection using auditory attention cues and task-dependent high level information,” *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 17, no. 5, pp. 1009–1024, 2009.
- [KNd08] G. Kootstra, A. Nederveen, and B. de Boer, “Paying attention to symmetry,” in *Proc. British Conf. Comp. Vis.*, 2008.
- [KPLL05] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, “Mechanisms for allocating auditory attention: an auditory saliency map,” *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.
- [KS13] H. S. Koppula and A. Saxena, “Anticipating human activities using object affordances for reactive robotic response,” in *RSS*, 2013.
- [KT00] D. Kahneman and A. Treisman, *Varieties of attention*. Academic Press, 2000, ch. Changing views of attention and automaticity, pp. 26–61.
- [KTG92] D. Kahneman, A. Treisman, and B. J. Gibbs, “The reviewing of object files: Object-specific integration of information,” *Cognitive Psychology*, vol. 24, no. 2, pp. 175–219, 1992.
- [KU85] C. Koch and S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry,” *Human Neurobiology*, vol. 4, pp. 219–27, 1985.
- [LB05] M. Louwerse and A. Bangerter, “Focusing attention with deictic gestures and linguistic expressions,” in *Proc. Ann. Conf. Cog. Sci. Soc.*, 2005.
- [LBW<sup>+</sup>02] P. Laurienti, J. H. Burdette, M. T. Wallace, Y. F. Yen, A. S. Field, and B. E. Stein, “Deactivation of sensory-specific cortex by cross-modal stimuli,” *Journal of Cognitive Neuroscience*, vol. 14, pp. 420–429, 2002.
- [LHR05] J. Lichtenauer, E. Hendriks, and M. Reinders, “Isophote properties as features for object detection,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2005.
- [Lil67] H. W. Lilliefors, “On the kolmogorov-smirnov test for normality with mean and variance unknown,” *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.
- [LL12] S. Lu and J.-H. Lim, “Saliency modeling from image histograms,” in *Proc. European Conf. Comp. Vis.*, 2012.
- [LLAH11] J. Li, M. D. Levine, X. An, and H. He, “Saliency detection based on frequency and spatial domain analysis,” in *Proc. British Conf. Comp. Vis.*, 2011.

- [LLLN12] W. Luo, H. Li, G. Liu, and K. N. Ngan, "Global salient information maximization for saliency detection," *Signal Processing: Image Communication*, vol. 27, pp. 238–248, 2012.
- [LMP01] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Machine Learning*, 2001.
- [LMSR08] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2008.
- [Low04] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.
- [LSL00] Y. Liang, E. Simoncelli, and Z. Lei, "Color channels decorrelation by ica transformation in the wavelet domain for color texture analysis and synthesis," in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, vol. 1, 2000, pp. 606–611.
- [LSZ<sup>+</sup>07] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2007.
- [LT09] K. Liebal and M. Tomasello, "Infants appreciate the social intention behind a pointing gesture: Commentary on "children's understanding of communicative intentions in the middle of the second year of life" by T. Aureli, P. Perucchini and M. Genco," *Cognitive Development*, vol. 24, no. 1, pp. 13–15, 2009.
- [LXG12] J. Li, D. Xu, and W. Gao, "Removing label ambiguity in learning-based visual saliency estimation," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1513–1525, 2012.
- [LZG<sup>+</sup>12] K.-H. Lin, X. Zhuang, C. Goudeseune, S. King, M. Hasegawa-Johnson, and T. S. Huang, "Improving faster-than-real-time human acoustic event detection by saliency-maximized audio visualization," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2012.
- [Mar82] D. Marr, *VISION – A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company, 1982.
- [MCB06] O. L. Meur, P. L. Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, no. 19, pp. 2483–2498, 2006.
- [MCUP04] 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.
- [MFL<sup>+</sup>07] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, and D. G. Lowe, "Curious George: An attentive semantic robot," in *IROS Workshop: From sensors to human spatial concepts*, 2007.

- [MFL<sup>+</sup>08] —, “Curious george: An attentive semantic robot,” *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 503–511, 2008.
- [MFOF08] M. A. McDowell, C. D. Fryar, C. L. Ogden, and K. M. Flegal, “Anthropometric reference data for children and adults: United states, 2003–2006,” National Health Statistics Reports, Tech. Rep., 2008.
- [MHS10] P. Matikainen, M. Hebert, and R. Sukthankar, “Representing pairwise spatial and temporal relations for action recognition,” in *Proc. European Conf. Comp. Vis.*, 2010.
- [MMKL99] J. R. Muller, A. B. Metha, J. Krauskopf, and P. Lennie, “Rapid adaptation in visual cortex to the structure of images,” *Science*, vol. 285, pp. 1405–1408, 1999.
- [MMS<sup>+</sup>09] T. Machmer, J. Moragues, A. Swerdlow, L. Vergara, J. Gosalbez-Castillo, and K. Kroschel, “Robust impulsive sound source localization by means of an energy detector for temporal alignment and pre-classification,” in *Proc. Europ. Sig. Proc. Conf.*, 2009.
- [MN07a] P. Mundy and L. Newell, “Attention, joint attention, and social cognition,” *Curr. Dir. Psychol. Sci.*, vol. 16, no. 5, pp. 269–274, 2007.
- [MN07b] —, “Attention, joint attention, and social cognition,” *Curr. Dir. Psychol. Sci.*, vol. 16, no. 5, pp. 269–274, 2007.
- [Moj05] A. Mojsilovic, “A computational model for color naming and describing color composition of images,” *IEEE Trans. Image Process.*, vol. 14, no. 5, pp. 690–699, 2005.
- [MON08] M. Minock, P. Olofsson, and A. Näslund, “Towards building robust natural language interfaces to databases,” in *Proc. Int. Conf. Appl. Nat. Lang. Inf. Sys.*, 2008, pp. 187–198.
- [MPI01] F. Miau, C. Papageorgiou, and L. Itti, “Neuromorphic algorithms for computer vision and attention,” in *Proc. SPIE 46 Annual International Symposium on Optical Science and Technology*, B. Bosacchi, D. B. Fogel, and J. C. Bezdek, Eds., vol. 4479, 2001, pp. 12–23.
- [MPK09] R. Messing, C. Pal, and H. Kautz, “Activity recognition using the velocity histories of tracked keypoints,” in *Proc. Int. Conf. Comp. Vis.*, 2009.
- [MS13] M. Martinez and R. Stiefelhagen, “Automated multi-camera system for long term behavioral monitoring in intensive care units,” in *Proc. Int. Conf. Machine Vision Applications*, 2013.
- [MSKK10] T. Machmer, A. Swerdlow, B. Kühn, and K. Kroschel, “Hierarchical, knowledge-oriented opto-acoustic scene analysis for humanoid robots and man-machine interaction,” in *Proc. Int. Conf. Robot. Autom.*, 2010.
- [NHMA03] Y. Nagai, K. Hosoda, A. Morita, and M. Asada, “A constructive model for the development of joint attention,” *Connection Science*, vol. 15, no. 4, pp. 211–229, 2003.
- [NI06] V. Navalpakkam and L. Itti, “An integrated model of top-down and bottom-up attention for optimizing detection speed,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2006.

- [NI07] —, “Search goal tunes visual features optimally.” *Neuron*, vol. 53, no. 4, pp. 605–617, 2007.
- [NIS12] NIST/SEMATECH, *Engineering Statistics Handbook*, 2012.
- [NJW<sup>+</sup>98] S. B. Nickerson, P. Jasiobedzki, D. Wilkes, M. Jenkin, E. Milios, J. K. Tsotsos, A. Jepson, and O. N. Bains, “The ark project: Autonomous mobile robots for known industrial environments,” *Robotics and Autonomous Systems*, vol. 25, pp. 83–104, 1998.
- [NM00] C. Nass and Y. Moon, “Machines and mindlessness: Social responses to computers,” *Journal of Social Issues*, vol. 56, no. 1, pp. 81–103, 2000.
- [NS07] K. Nickel and R. Stiefelhagen, “Visual recognition of pointing gestures for human-robot interaction,” *Image and Vision Computing*, vol. 25, no. 12, pp. 1875–1884, 2007.
- [NSK14] J. Nakajima, A. Sugimoto, and K. Kawamoto, “Incorporating audio signals into constructing a visual saliency map,” in *Image and Video Technology*, ser. Lecture Notes in Computer Science, R. Klette, M. Rivera, and S. Satoh, Eds. Springer Berlin Heidelberg, 2014, vol. 8333.
- [OBH<sup>+</sup>01] N. Ouerhani, J. Bracamonte, H. Hugli, M. Ansorge, and F. Pellandini, “Adaptive color image compression based on visual attention,” in *Proc. Int. Conf. Image Analysis and Processing*, 2001, pp. 416–421.
- [OF96] B. A. Olshausen and D. J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, pp. 607–609, 1996.
- [OK04] A. Olmos and F. A. A. Kingdom, “A biologically inspired algorithm for the recovery of shading and reflectance images,” *Perception*, vol. 33, pp. 1463–1473, 2004.
- [OL81] A. Oppenheim and J. Lim, “The importance of phase in signals,” *Proc. IEEE*, vol. 69, no. 5, pp. 529–541, 1981.
- [OLK07] S. Onat, K. Libertus, and P. König, “Integrating audiovisual information for the control of overt attention,” *Journal of Vision*, vol. 7, no. 10, 2007.
- [OMS08] F. Orabona, G. Metta, and G. Sandini, “A proto-object based visual attention model,” in *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, L. Paletta and E. Rome, Eds., 2008, pp. 198–215.
- [OSI13] L. Onofri, P. Soda, and G. Iannello, “Multiple subsequence combination in human action recognition,” *IET Computer Vision*, 2013.
- [Pas08] D. Pascale, “A review of RGB color spaces...from xyY to R’G’B’,” 2008.
- [Pea00] K. Pearson, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” *Philosophical Magazine*, vol. 50, no. 302, pp. 157–175, 1900.
- [Pet03] B. F. Peterson, *Learning to see creatively*. Amphoto Press, 2003.



- [PFS12] A. Prest, V. Ferrari, and C. Schmid, “Explicit modeling of human-object interactions in realistic videos,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, pp. 835–848, 2012.
- [PGMC05] D. Perez-Gonzalez, M. S. Malmierca, and E. Covey, “Novelty detector neurons in the mammalian auditory midbrain,” *European Journal of Neuroscience*, vol. 22, pp. 2879–2885, 2005.
- [PI08a] R. J. Peters and L. Itti, “Applying computational tools to predict gaze direction in interactive visual environments,” *ACM Transactions on Applied Perception*, vol. 5, no. 2, 2008.
- [PI08b] R. Peters and L. Itti, “The role of fourier phase information in predicting saliency,” *Journal of Vision*, vol. 8, no. 6, p. 879, 2008.
- [PIIK05] R. Peters, A. Iyer, L. Itti, and C. Koch, “Components of bottom-up gaze allocation in natural images,” *Vision Research*, vol. 45, no. 18, pp. 2397–2416, 2005.
- [Piw07] P. L. A. Piwek, “Modality choice for generation of referring acts: Pointing versus describing,” in *Proc. Int. Workshop on Multimodal Output Generation*, 2007.
- [PLN02] D. Parkhurst, K. Law, and E. Niebur, “Modeling the role of salience in the allocation of overt visual attention,” *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.
- [PN03] D. Parkhurst and E. Niebur, “Scene content selected by active vision,” *Spatial Vision*, vol. 16, no. 2, pp. 125–154, 2003.
- [Pos80] M. I. Posner, “Orienting of attention,” *Quarterly Journal of Experimental Psychology*, vol. 32, no. 1, pp. 3–25, 1980.
- [PPI09] G. Passino, I. Patras, and E. Izquierdo, “Latent semantics local distribution for crf-based image semantic segmentation,” in *Proc. British Conf. Comp. Vis.*, 2009.
- [RAGS01] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, “Color transfer between images,” *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 2001.
- [Rat65] F. Ratliff, *Mach Bands: Quantitative Studies on Neural Networks in the Retina*. Holden-Day, San Francisco, 1965.
- [RB99] R. P. Rao and D. H. Ballard, “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects,” *Nature Neuroscience*, pp. 79–87, 1999.
- [RBC06] U. Rajashekar, A. C. Bovik, and L. K. Cormack, “Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis,” *Journal of Vision*, vol. 6, no. 4, pp. 379–386, 2006.
- [RCC98] D. Ruderman, T. Cronin, and C. Chiao, “Statistics of cone responses to natural images: Implications for visual coding,” *Journal of the Optical Society of America*, vol. 15, no. 8, pp. 2036–2045, 1998.

- [RDM<sup>+</sup>13] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, “Saliency and human fixations: State-of-the-art and study of comparison metrics,” in *Proc. Int. Conf. Comp. Vis.*, 2013.
- [Ren00a] R. A. Rensink, “The dynamic representation of scenes,” *Visual Cognition*, vol. 7, pp. 17–42, 2000.
- [Ren00b] —, “Seeing, sensing, and scrutinizing,” *Vision Research*, vol. 40, pp. 1469–1487, 2000.
- [RFHS11] L. Rybok, S. Friedberger, U. D. Hanebeck, and R. Stiefelhagen, “The KIT robo-kitchen data set for the evaluation of view-based activity recognition systems,” in *Humanoids*, 2011.
- [RLB<sup>+</sup>08] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, “Multimodal saliency-based bottom-up attention: A framework for the humanoid robot iCub,” in *Proc. Int. Conf. Robot. Autom.*, 2008.
- [RLS98] P. R. Roelfsema, V. A. F. Lamme, and H. Spekreijse, “Object-based attention in the primary visual cortex of the macaque monkey,” *Nature*, vol. 395, pp. 376–381, 1998.
- [RMDB<sup>+</sup>13] S. Ramenahalli, D. R. Mendat, S. Dura-Bernal, E. Culurciello, E. Niebur, and A. Andreou, “Audio-Visual Saliency Map: Overview, Basic Models and Hardware Implementation,” in *Ann. Conf. Information Sciences and Systems*, 2013.
- [Rob] RobotCub Consortium, “iCub – an open source cognitive humanoid robotic platform,” <http://www.icub.org>.
- [Roy05] D. Roy, “Grounding words in perception and action: computational insights,” *Trends in Cognitive Sciences*, vol. 9, no. 8, pp. 389–396, 2005.
- [RP99] M. Riesenhuber and T. Poggio, “Hierarchical models of object recognition in cortex,” *Nature Neuroscience*, no. 2, pp. 1019–1025, 1999.
- [RP11] E. Reinhard and T. Pouli, “Colour spaces for colour transfer,” in *Computational Color Imaging*, ser. Lecture Notes in Computer Science, 2011, vol. 6626, pp. 1–15.
- [RPF08] J. Richarz, T. Plötz, and G. A. Fink, “Real-time detection and interpretation of 3D deictic gestures for interaction with an intelligent environment,” in *Proc. Int. Conf. Pat. Rec.*, 2008, pp. 1–4.
- [RR13] X. Ren and D. Ramanan, “Histograms of sparse codes for object detection,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2013.
- [RSA08] M. Rubinstein, A. Shamir, and S. Avidan, “Improved seam carving for video retargeting,” in *Proc. Ann. Conf. Special Interest Group on Graphics and Interactive Techniques*, 2008.
- [RVES10] L. Rybok, M. Voit, H. K. Ekenel, and R. Stiefelhagen, “Multi-view based estimation of human upper-body orientation,” in *Proc. Int. Conf. Pat. Rec.*, 2010.
- [RZ99] P. Reinagel and A. M. Zador, “Natural scene statistics at the centre of gaze,” in *Network: Computation in Neural Systems*, 1999, pp. 341–350.

- [SAD<sup>+</sup>06] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, “Gaze-based interaction for semi-automatic photo cropping,” in *Proc. Int. Conf. Human Factors Comput. Syst. (CHI)*, 2006.
- [San96] S. J. Sangwine, “Fourier transforms of colour images using quaternion or hypercomplex, numbers,” *Electronics Letters*, vol. 32, no. 21, pp. 1979–1980, 1996.
- [SC09] M. Staudte and M. W. Crocker, “Visual attention in spoken human-robot interaction,” in *Proc. Int. Conf. Human-Robot Interaction*, 2009, pp. 77–84.
- [SE00] S. Sangwine and T. Ell, “Colour image filters based on hypercomplex convolution,” *IEEE Proc. Vision, Image and Signal Processing*, vol. 147, no. 2, pp. 89–93, 2000.
- [SG31] T. Smith and J. Guild, “The C.I.E. colorimetric standards and their use,” *Transactions of the Optical Society*, vol. 33, no. 3, p. 73, 1931.
- [SHH<sup>+</sup>08] J. Schmidt, N. Hofemann, A. Haasch, J. Fritsch, and G. Sagerer, “Interacting with a mobile robot: Evaluating gestural object references,” in *Proc. Int. Conf. Intell. Robots Syst.*, 2008, pp. 3804–3809.
- [SI09] C. Siagian and L. Itti, “Biologically inspired mobile robot vision localization,” *IEEE Trans. Robot.*, vol. 25, no. 4, pp. 861–873, 2009.
- [SIK11] J. Schnupp, I. N. I., and A. King, *Auditory Neuroscience*. MIT Press, 2011.
- [SKI<sup>+</sup>07] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, “Natural deictic communication with humanoid robots,” in *Proc. Int. Conf. Intell. Robots Syst.*, 2007.
- [SLBJ03] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs, “Automatic thumbnail cropping and its effectiveness,” in *ACM Symposium on User interface Software and Technology*, 2003.
- [SLNG07] V. Setlur, T. Lechner, M. Nienhaus, and B. Gooch, “Retargeting images and video for preserving information saliency,” *IEEE Comput. Graph. Appl.*, vol. 27, no. 5, pp. 80–88, 2007.
- [SMI] SMIVision, “Sensomotoric instruments gmbh,” <http://www.smivision.com/>.
- [SMSK08] Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike, “An Incremental Learning Method for Unconstrained Gaze Estimation,” in *ECCV*, 2008.
- [SPG12] G. Song, D. Pellerin, and L. Granjon, “How different kinds of sound in videos can influence gaze,” in *Int. Workshop on Image Analysis for Multimedia Interactive Services*, 2012.
- [Spi83] D. J. Spiegelhalter, “Diagnostic tests of distributional shape,” *Biometrika*, vol. 70, no. 2, pp. 401–409, 1983.
- [SPSS12] J. Sung, C. Ponce, B. Selman, and A. Saxena, “Unstructured human activity detection from RGBD images,” in *Proc. Int. Conf. Robot. Autom.*, 2012.

- [SS06] C. Simion and S. Shimojo, "Early interactions between orienting, visual sampling and decision making in facial preference," *Vision Research*, vol. 46, no. 20, pp. 3331–3335, 2006.
- [SS07] F. Shic and B. Scassellati, "A behavioral analysis of computational models of visual attention," *International Journal of Computer Vision*, vol. 73, pp. 159–177, 2007.
- [SSYK07] F. Saidi, O. Stasse, K. Yokoi, and F. Kanehiro, "Online object search with a humanoid robot," in *Proc. Int. Conf. Intell. Robots Syst.*, 2007.
- [STET01] M. J. Spivey, M. J. Tyler, K. M. Eberhard, and M. K. Tanenhaus, "Linguistically mediated visual search," *Psychological Science*, vol. 12, pp. 282–286, 2001.
- [Sus05] E. S. Sussman, "Integration and segregation in auditory scene analysis," *Journal of the Acoustic Society America*, vol. 117, pp. 1285–1298, 2005.
- [SW65] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, pp. 591–611, 1965.
- [SW87] G. L. Shulman and J. Wilson, "Spatial frequency and selective attention to spatial location," *Perception*, vol. 16, no. 1, pp. 103–111, 1987.
- [SW01] E. S. Sussman and I. Winkler, "Dynamic sensory updating in the auditory system," *Cognitive Brain Research*, vol. 12, pp. 431–439, 2001.
- [SY06] J. T. Serences and S. Yantis, "Selective visual attention and perceptual coherence," *Trends in Cognitive Sciences*, vol. 10, no. 1, pp. 38–45, 2006.
- [SYH07] E. Sato, T. Yamaguchi, and F. Harashima, "Natural interface using pointing behavior for human-robot gestural interaction," *IEEE Trans. Ind. Electron.*, vol. 54, no. 2, pp. 1105–1112, 2007.
- [SYW97] R. Stiefelhagen, J. Yang, and A. Waibel, "Tracking Eyes and Monitoring Eye Gaze," in *Workshop on Perceptual User Interfaces*, 1997.
- [Tat07] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14, 2007.
- [TB00] E. F. Tjong Kim Sang and S. Buchholz, "Introduction to the CoNLL-2000 shared task: Chunking," in *Proc. Int. Workshop on Comp. Nat. Lang. Learn.*, 2000, pp. 127–132.
- [TBG05] B. Tatler, R. Baddeley, and I. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision Research*, vol. 45, no. 5, pp. 643–659, 2005.
- [TCC<sup>+</sup>09] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *Journal of Vision*, vol. 9, no. 7, 2009.
- [TCW<sup>+</sup>95] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nufflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, 1995.

- [TDW91] S. P. Tipper, J. Driver, and B. Weaver, “Object-centred inhibition of return of visual attention,” *Quarterly Journal of Experimental Psychology*, vol. 43, pp. 289–298, 1991.
- [TG80] A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.
- [TG88] A. M. Treisman and S. Gormican, “Feature analysis in early vision: Evidence from search asymmetries,” *Psychological Review*, vol. 95, no. 1, pp. 15–48, 1988.
- [TKA02] K.-H. Tan, D. J. Kriegman, and N. Ahuja, “Appearance-based eye gaze estimation,” in *IEEE Workshop on Applications of Computer Vision*, 2002.
- [TMZ<sup>+</sup>07] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, “Clear evaluation of acoustic event detection and classification systems,” ser. Lecture Notes in Computer Science, R. Stiefelhagen and J. Garofolo, Eds. Springer Berlin Heidelberg, 2007, vol. 4122, pp. 311–322.
- [TOCH06] A. Torralba, A. Oliva, M. S. Castelhana, and J. M. Henderson, “Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search,” *Psychological review*, vol. 113, no. 4, 2006.
- [Tom03] M. Tomasello, *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, 2003.
- [Tso89] J. K. Tsotsos, “The complexity of perceptual search tasks,” in *Proc. Int. Joint Conf. Artif. Intell.*, 1989.
- [Tso95] —, “Behaviorist intelligence and the scaling problem,” *Artif. Intell.*, vol. 75, pp. 135–160, 1995.
- [Tso11] —, *A Computational Perspective on Visual Attention*. The MIT Press, 2011.
- [TTDC06] J. Triesch, C. Teuscher, G. O. Deák, and E. Carlson, “Gaze following: why (not) learn it?” *Developmental Science*, vol. 9, no. 2, pp. 125–147, 2006.
- [Uni] University of British Columbia, “Curious George Project,” [https://www.cs.ubc.ca/labs/lci/curious\\_george/](https://www.cs.ubc.ca/labs/lci/curious_george/), retrieved 3 April 2014.
- [Usa] UsableWorld – User Experience Strategy, Usability Testing, Eye Tracking, “You look where they look,” <http://usableworld.com.au/2009/03/16/you-look-where-they-look/>, retrieved 3 April 2014.
- [VCSS01] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, “Overt visual attention for a humanoid robot,” in *Proc. Int. Conf. Intell. Robots Syst.*, 2001.
- [vdWSV07] J. van de Weijer, C. Schmid, and J. J. Verbeek, “Learning color names from real-world images,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2007.

- [VK89] R. M. Vogel and C. N. Kroll, “Low-flow frequency analysis using probability-plot correlation coefficients,” *Journal of Water Resources Planning and Management*, vol. 115, no. 3, pp. 338–357, 1989.
- [vR79] C. J. van Rijsbergen, *Information Retrieval*, 2nd ed. Butterworth, 1979.
- [VS08] M. Voit and R. Stiefelhagen, “Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios,” in *Proc. Int. Conf. Multimodal Interfaces*, 2008, pp. 173–180.
- [VSG12] R. Valenti, N. Sebe, and T. Gevers, “Combining head pose and eye location information for gaze estimation.” *IEEE Transactions on Image Processing*, vol. 21, pp. 802–815, 2012.
- [VT08] J. Verbeek and B. Triggs, “Scene segmentation with crfs learned from partially labeled images,” in *Proc. Ann. Conf. Neural Inf. Process. Syst.*, vol. 20, 2008, pp. 1553–1560.
- [VW87] R. R. Vallacher and D. M. Wegner, “What do people think they’re doing? action identification and human behavior,” *Psychological Review*, vol. 94, no. 1, pp. 3–15, 1987.
- [WAD09] K. Welke, T. Asfour, and R. Dillmann, “Active multi-view object search on a humanoid head,” in *Proc. Int. Conf. Robot. Autom.*, 2009.
- [WAD11] —, “Inhibition of return in the bayesian strategy to active visual search.” in *Proc. Int. Conf. Machine Vision Applications*, 2011.
- [WBB<sup>+</sup>96] P. W. Woodruff, R. R. Benson, P. A. Bandettini, K. K. Kwong, R. J. Howard, T. Talavage, J. Belliveau, and B. R. Rosen, “Modulation of auditory and visual cortex by selective attention is modality-dependent,” *Neuroreport*, vol. 7, pp. 1909–1913, 1996.
- [WBM12] C.-A. Wang, S. Boehnke, and D. Munoz, “Pupil dilation evoked by a salient auditory stimulus facilitates saccade reaction times to a visual stimulus,” *Journal of Vision*, vol. 12, no. 9, p. 1254, 2012.
- [WCD<sup>+</sup>13] P.-H. Wu, C.-C. Chen, J.-J. Ding, C.-Y. Hsu, and Y.-W. Huang, “Salient region detection improved by principle component analysis and boundary information,” *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3614–3624, 2013.
- [WCF89] J. M. Wolfe, K. Cave, , and S. Franzel, “Guided search: An alternative to the feature integration model for visual search,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, pp. 419–433, 1989.
- [WCS<sup>+</sup>05] I. Winkler, I. Czigler, E. Sussman, J. Horvath, and L. Balazs, “Preattentive binding of auditory and visual stimulus features,” *Journal of Cognitive Neuroscience*, vol. 17, pp. 320–339, 2005.
- [WCW11] J. Wang, Z. Chen, and Y. Wu, “Action recognition with multiscale spatio-temporal contexts,” in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2011.
- [Weg05] I. Wegener, *Theoretische Informatik – eine algorithmenorientierte Einführung*. Teubner, 2005.

- [Wel11] K. Welke, “Memory-based active visual search for humanoid robots,” Ph.D. dissertation, Karlsruhe Institute of Technology, 2011.
- [WHK<sup>+</sup>04] J. M. Wolfe, T. S. Horowitz, N. Kenner, M. Hyle, and N. Vasan, “How fast can you change your mind? the speed of top-down guidance in visual search,” *Vision Research*, vol. 44, pp. 1411–1426, 2004.
- [Wik] Wikimedia Common (Googolplexbyte), “Diagram of the opponent process,” [http://commons.wikimedia.org/wiki/File:Diagram\\_of\\_the\\_opponent\\_process.png](http://commons.wikimedia.org/wiki/File:Diagram_of_the_opponent_process.png), retrieved 3 April 2014.
- [WJ08] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*. Hanover, MA, USA: Now Publishers Inc., 2008.
- [WK06] D. Walther and C. Koch, “Modeling attention to salient proto-objects,” *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.
- [Wol94] J. M. Wolfe, “Guided search 2.0: A revised model of visual search,” *Psychonomic Bulletin and Review*, vol. 1, pp. 202–238, 1994.
- [WS13] S. Winkler and R. Subramanian, “Overview of eye tracking datasets,” in *Int. Workshop on Quality of Multimedia Experience*, 2013.
- [WTSH<sup>+</sup>03] I. Winkler, W. A. Teder-Salejarvi, J. Horvath, R. Naatanen, and E. Sussman, “Human auditory cortex tracks task-irrelevant sound sources,” *Neuroreport*, vol. 14, pp. 2053–2056, 2003.
- [WWW04] D. H. Weissman, L. M. Warner, and M. G. Woldorff, “The neural mechanisms for minimizing cross-modal distraction,” *Journal of Neuroscience*, vol. 24, pp. 10 941–10 949, 2004.
- [XCKB09] T. Xu, N. Chenkov, K. Kühnlenz, and M. Buss, “Autonomous switching of top-down and bottom-up attention selection for vision guided mobile robots,” in *Proc. Int. Conf. Intell. Robots Syst.*, 2009.
- [XPKB09] T. Xu, T. Pototschnig, K. Kühnlenz, and M. Buss, “A high-speed multi-GPU implementation of bottom-up attention using CUDA,” in *Proc. Int. Conf. Robot. Autom.*, 2009.
- [XZKB10] T. Xu, T. Zhang, K. Kühnlenz, and M. Buss, “Attentional object detection of an active multi-vocal vision system,” *Int. J. of Humanoid*, vol. 7, no. 2, 2010.
- [Yar67] A. L. Yarbus, *Eye Movements and Vision*. Plenum Press, 1967.
- [YGMG13] Y. Yu, J. Gu, G. Mann, and R. Gosine, “Development and evaluation of object-based visual attention for automatic perception of robots,” *IEEE Trans. Automation Science and Engineering*, vol. 10, no. 2, pp. 365–379, 2013.
- [YL13] Y. Yi and Y. Lin, “Human action recognition with salient trajectories,” *Signal Processing*, 2013.
- [YScM<sup>+</sup>13] Z. Yücel, A. A. Salah, Çetin Meriçli, T. Meriçli, R. Valenti, and T. Gevers, “Joint attention by gaze interpolation and saliency,” *IEEE Trans. Syst., Man, Cybern.*, vol. 43, no. 3, pp. 829–842, 2013.

- [YSS10] C. Yu, M. Scheutz, and P. Schermerhorn, “Investigating multimodal real-time patterns of joint attention in an hri word learning task,” in *Proc. Int. Conf. Human-Robot Interaction*, 2010.
- [Zad65] L. Zadeh, “Fuzzy sets,” *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.
- [ZJY12] J. Zhou, Z. Jin, and J. Yang, “Multiscale saliency detection using principle component analysis,” in *Int. Joint Conf. on Neural Networks*, 2012, pp. 1–6.
- [ZK11] Q. Zhao and C. Koch, “Learning a saliency map using fixated locations in natural scenes,” *Journal of Vision*, vol. 11, no. 3, pp. 1–15, 2011.
- [ZTM<sup>+</sup>08] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, “Sun: A bayesian framework for saliency using natural statistics,” *Journal of Vision*, vol. 8, no. 7, 2008.