TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Medientechnik

Image-based Novelty Detection for Cognitive Mobile Robots

Dipl.-Ing. Univ. Werner A. Maier

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzende:		Univ.-Prof. Dr.-Ing. Sandra Hirche
Prüfer der Dissertation:	1.	Univ.-Prof. Dr.-Ing. Eckehard Steinbach
	2.	Univ.-Prof. Dr.-Ing. Klaus Diepold

Die Dissertation wurde am 08.11.2011 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 13.04.2012 angenommen.

Image-based Novelty Detection for Cognitive Mobile Robots

Dipl.-Ing. Univ. Werner A. Maier

May 14, 2012

To María and Marc.

Acknowledgments

This thesis results from my research work at the Institute for Media Technology at the Technische Universität München. Since there are a lot of people who have contributed to my work over the years, I would like to express my gratitude to them.

First of all, I would like to thank my supervisor Prof. Dr.-Ing. Eckehard Steinbach. I am grateful for the constructive advice he gave me during discussions and for his encouraging feedback on my work and plans. I appreciate very much his great expertise in the fields of computer vision and computer graphics as well as his commitment.

Furthermore, I would like to thank Prof. Dr.-Ing. Klaus Diepold for accepting to be the second examiner of this thesis and Prof. Dr.-Ing. Sandra Hirche for heading the committee of examiners.

I would like to thank Elmar Mair and Prof. Dr.-Ing. Darius Burschka from the Robotics and Embedded Systems Group at the Technische Universität München for providing the camera pose data for some of the image sequences used in this work.

Moreover, it is a pleasure to me to thank my colleague Nicolas Alt from the Institute for Media Technology for his suggestions and for the fruitful collaboration during the preparation of demonstrations. Furthermore, I thank Omiros Kourakos, Matthias Rambow, Florian Rohrmüller and Martin Lawitzky from the Insitute of Automatic Control Engineering at the Technische Universität München for granting me access to their robot platforms and for their technical assistance in case of hardware issues. I would also like to thank Daniel Althoff from the Institute of Automatic Control Engineering for the navigation of the robot during some of the experiments described in this work.

I am also indebted to my Master students for their brilliant work especially in the field of illumination-invariant image-based novelty detection. Without them it would have taken much more time to finish this part of my thesis.

I owe my deepest gratitude to my parents Maria and Alfred who always stood behind my decisions and supported me in life. This thesis would not have been possible without them.

Last but not least, I am heartfully thankful to my wife María and my son Marc for the love, warmth and joy they give me.

Munich, October 2011 *Werner Maier*

Abstract

Robots need to be able to perceive their environment when they execute object manipulation tasks or when they interact with humans. To this end, robots are equipped with sensors to acquire visual and geometric information. Geometric representations of the environment are important for tasks where the robot has to measure the distances to the objects in the surroundings, e.g. during navigation or an object grasping process. Image-based representations acquired from camera images, in turn, provide information for object recognition and a reference model of the environment for the detection of novel objects.

In the first part of this thesis a novel probabilistic appearance representation is investigated. This environment model is inspired by image-based scene representations and thus represents the colors of the robot's 3D environment at a densely spaced series of viewpoints. Instead of storing the raw color value captured by a camera, however, a Gaussian model is used for the luminance and chrominance at each pixel of a view. Using depth information and camera pose data stored at each viewpoint, the probabilistic priors can be interpolated at intermediate viewpoints, as virtual images are synthesized from image-based representations. The Gaussian distributions model the uncertainty of the appearance which can arise from erroneous pose or depth estimates or from moving objects in the scene. The expectation and uncertainty of the appearance of the scene are used for assessing the level of *surprise* of novel visual stimuli. As the performance analysis shows, the surprise measure is a detector for novelty which is superior to measures based on image differencing.

As an application of surprise detection, a method for the acquisition of feature-based object representations is presented. A region in a new camera image where a new object is exhibited in a familiar environment and where high surprise values are measured is analyzed for local image features. A selected subset of these features is added to a database such that object representations are generated which can later be used for object recognition during the robot's tasks.

In the second part of the thesis an illumination-invariant image-based environment representation is investigated. This environment representation is computed from multiple image sequences acquired under different illumination conditions. Together with statistical models for the variation of the illumination in the environment it is used for detecting the addition or removal of objects in the environment while illumination changes are suppressed. This enables the robot to distinguish between novelty from new or disappeared objects, which is relevant for its tasks, and irrelevant novelty, which can result from varying illumination. In this context, a special technique allows for the identification and suppression of specularities, which makes the approach versatile for real-world environments with arbitrary surface materials.

Kurzfassung

Roboter müssen in der Lage sein, ihre Umgebung wahrzunehmen, wenn sie in ihren Aufgaben mit Objekten oder Menschen interagieren. Zu diesem Zweck werden Roboter mit Sensoren ausgestattet, um visuelle und geometrische Information aufzunehmen. Geometrische Umgebungsbeschreibungen sind für Aufgaben wichtig, in denen der Roboter die Entfernungen der Objekte um ihn herum messen muss, z.B. während der Navigation oder während eines Greifvorgangs. Bildbasierte Umgebungsbeschreibungen, die aus Kamerabildern aufgenommen werden, liefern Informationen zur Objekterkennung und ein Referenzmodell der Umgebung, um neue Objekte zu detektieren.

Im ersten Teil dieser Dissertation wird eine neuartige Umgebungsbeschreibung untersucht, die das Aussehen der Umgebung wahrscheinlichkeitstheoretisch beschreibt. Diese Umgebungsbeschreibung lehnt sich an bildbasierte Szenenbeschreibungen an und beschreibt somit die Farben der 3D-Umgebung des Roboters an einer dichten Reihe von Blickpunkten. Anstatt die Farbwerte zu speichern, die von einer Kamera aufgenommen werden, wird jedoch an jedem Pixel einer Ansicht ein Gauß-Modell für die Luminanz und Chrominanz verwendet. Unter Verwendung von Tiefeninformation und der Lagedaten der Kamera, die an jedem Blickpunkt gespeichert sind, können die A-Priori-Verteilungen für Zwischenansichten interpoliert werden, in ähnlicher Weise, wie virtuelle Bilder aus bildbasierten Szenenbeschreibungen berechnet werden. Die Gauß-Verteilungen modellieren die Unsicherheit über das Erscheinungsbild der Umgebung, welche aus fehlerhaften Lage- oder Geometrieschätzungen, oder durch sich bewegende Objekte entstehen kann. Die Erwartung und die Unsicherheit über das Aussehen der Szene werden für die Bewertung des Überraschungsgehalts eines neuen visuellen Reizes verwendet. Wie die Performanz-Analyse zeigt, ist das vorgeschlagene Maß für *Überraschung* anderen Maßen, die auf der Berechnung von Bilddifferenzen beruhen, überlegen.

Als Anwendung für die Überraschungsdetektion wird eine Methode vorgestellt, mit der merkmalsbasierte Objektbeschreibungen erstellt werden können. Ein Bereich in einem neuen Kamerabild, der ein neues Objekt in einer sonst bekannten Umgebung zeigt und in dem hohe Überraschungswerte gemessen werden, wird nach lokalen Bildmerkmalen untersucht. Eine Auswahl an Merkmalen wird in einer Datenbank abgelegt, so dass Objektbeschreibungen erzeugt werden, mit denen der Roboter bei seinen Aufgaben Objekte wiedererkennen kann.

Im zweiten Teil dieser Dissertation wird eine beleuchtungsinvariante bildbasierte Umgebungsbeschreibung untersucht. Diese Umgebungsbeschreibung wird aus mehreren Bildersequenzen errechnet, die unter veränderlichen Beleuchtungsbedingungen aufgenommen werden. Zusammen mit statistischen Modellen für die Beleuchtungsänderungen in der

Umgebung wird sie verwendet, um das Hinzufügen oder Entfernen von Objekten festzustellen, wobei Beleuchtungsschwankungen unterdrückt werden. Dies ermöglicht es dem Roboter zu unterscheiden zwischen neuartigen Ereignissen, wie neuen oder verschwundenen Objekten, die relevant für seine Aufgaben sind, oder nicht-relevanten neuartigen Ereignissen wie Beleuchtungsänderungen. In diesem Zusammenhang erlaubt es eine spezielle Methode, Spiegelungen auf Oberflächen zu identifizieren und zu unterdrücken. Daraus entsteht ein Ansatz, der in der realen Welt mit Oberflächen, die aus beliebigen Materialien bestehen, vielseitig eingesetzt werden kann.

Contents

List of Figures

List of Tables

Abbreviations

2D Two-Dimensional

3D Three-Dimensional

3DTV Three-Dimensional Television

6D Six-Dimensional

6DoF 6-Degree of Freedom

AS Attentional Selectivity

Cg C for graphics

DoG Difference of Gaussians

FN False Negative

FP False Positive

FPR False Positive Rate

GLSL OpenGL Shading Language

GPU Graphics Processing Unit

HSV Hue-Saturation-Value

IBR Image-Based Rendering

KLT Kanade-Lucas-Tomasi

LED Light-Emitting Diode

NGC Normalized Gradient Correlation

PAS Peak Attentional Selectivity

ROC Receiver Operating Characteristic

RGB Red-Green-Blue

RT Reaction Time

SIFT Scale-Invariant Feature Transform

SLAM Simultaneous Localization And Mapping

SURF Speeded-Up Robust Features

SVD Singular Value Decomposition

TN True Negative

TP True Positive

TPR True Positive Rate

VGA Video Graphics Array

VGPS Vision-based GPS

WTA Winner-Takes-it-All

YCbCr Luminance-Chrominance

1 Introduction

Robots are becoming more and more ubiquitous in people's everyday lives. While, over the past decades, robots have been predominantly used in factories to automate production processes, to increase efficiency or to execute tasks which, on the long run, can harm the workers' health, nowadays, robots found their way into private households. Small robots, which can be employed for simple tasks like vacuum cleaning, are available as consumer products and are getting more and more popular.

The vision of many researchers is that robots take over more and more complex tasks in a household in the future, which are not restricted to floor cleaning but can also involve tasks like setting the table, cleaning the dishes etc. Robots could then help elderly people who need permanent care and cannot do these things on their own. These tasks, however, require cognitive capabilities which today's robots do not possess yet. Currently, robots primarily execute their actions according to a rigid program and hardly adapt to their environment. The execution of tasks in a household, which also includes the interaction with humans, however, requires that robots be able to autonomously create and maintain knowledge representations of their environment, which contain information about task-relevant objects and their interaction partners at multiple abstraction levels. Besides, robots have to be able to make intelligent action plans to execute their tasks in an efficient way. In this context, it is also particularly important that the action plans are adaptive such that a robot can react to unforeseen events, e.g. that a cleaning action is triggered when a plate breaks. Besides, intelligent and flexible planning can contribute to customizing industrial mass production and to further improving the quality of products. Hence, cognition for technical systems is a hot topic in research which requires collaboration between different disciplines like electrical engineering, computer sciences, psychology and neurosciences.

Understanding the environment and learning representations of it relies on the perception of stimuli. To this end, robots can be equipped with sensors to acquire visual, auditory and tactile data, just like humans do with their eyes, ears and hands. Although multimodal data acquisition is important for an integral perception of the environment, this thesis focuses on environment representations obtained from visual data. Images, which are formed on the retinas of the human eyes or on the imaging sensors of digital cameras, provide plenty of information which is useful to locate and recognize subjects and objects, to interpret the emotional state and intentions of humans or to extract geometric information about the environment. Especially the recognition of landmarks and the inference of geometric data is essential for a series of algorithms in the field of robotics which deal with Simultaneous Localization and Mapping (SLAM). The self-localization in a known environment enables a robot to plan trajectories to get from one point in the environment to another in a fast and safe way. The computation of a geometric environment map is particularly important for obstacle avoidance during navigation. Besides, the acquisition of 3D information is indispensable for grasping an object. The 3D shape of an object enables a robot to determine the object class and optimal grasping parameters like the grasping point and the approaching direction.

To acquire visual information, robots are usually equipped with passive imaging devices like cameras and/or active imaging devices for range finding which emit light (laser) and leverage the time-of-flight of a light ray reflected by the scene to measure distances. The latter outperform passive stereo vision in 3D reconstruction if the environment exhibits surfaces with little texture or under dark lighting conditions. Nevertheless, active range finding usually fails for transparent surfaces since the emitted light passes through them and is hardly reflected. A similar issue arises with shiny metallic surfaces which usually do not reflect the emitted light in the direction of the imaging sensor. Passive stereo vision, in general, facilitates the reconstruction of the edges of transparent objects, since there acceptable pixel correspondences can be found between the camera images. However, surface regions where the light rays pass through the object without being refracted by the glass also provide poor results in depth estimation. The estimated depth corresponds then, as in case of active imaging, to the depth of the background behind the transparent object.

Geometric representations are not the only way of describing objects. A lot of information for recognizing objects is contained in the gradient structure and the colors in a camera image. Furthermore, many models for human visual attention are primarily based on appearance cues. Hence, appearance-based environment models can support the robot in object recognition and visual search tasks which are driven by attention. Having a realistic internal representation of the environment's appearance, a robot can rapidly find novel interesting regions in the currently captured camera image and concentrate its attention on them to extract visual features on a high level of detail. In the field of computer graphics, the appearance of a virtual 3D environment is traditionally visualized by mapping textures on the faces of a geometric mesh of the environment. Textures are two-dimensional color arrays which can be obtained by real captured camera images. However, for modeling the appearance of an environment with transparent and shiny surfaces as well as complex illumination requires sophisticated algorithms like ray tracing which are costly.

To this end, image-based rendering methods have been developed as an alternative way of visualizing the appearance of virtual 3D environments. Image-based representations encompass a large number of densely acquired camera views (*reference views*) and can contain additional information as view-dependent geometric information and camera pose data. View interpolation methods aim at the synthesis of photorealistic virtual images from real images stored in the representation. The transfer of pixel colors from the real images to the virtual image, using a geometric approximation of the scene, is independent of the complexity of the scene, of the illumination and of the surface materials. Thus, no knowledge about the refraction coefficient of transparent surfaces or other material properties is required for fast and high-qualitative rendering.

Apart from its applications in the area of computer graphics, image-based rendering is a promising technique for the realistic and fast prediction of the appearance of a robot's environment to support processes like surprise detection and attentional selection of image regions. Probabilistic extensions of image-based representation enable a robot to assess the novelty of perceived visual stimuli and to detect unexpected and surprising events in the environment. Surprise detection in cognitive systems can trigger a replanning of actions in case of unforeseen situations. Furthermore, surprise can contribute to learning and to the update and the extension of a robot's knowledge representation.

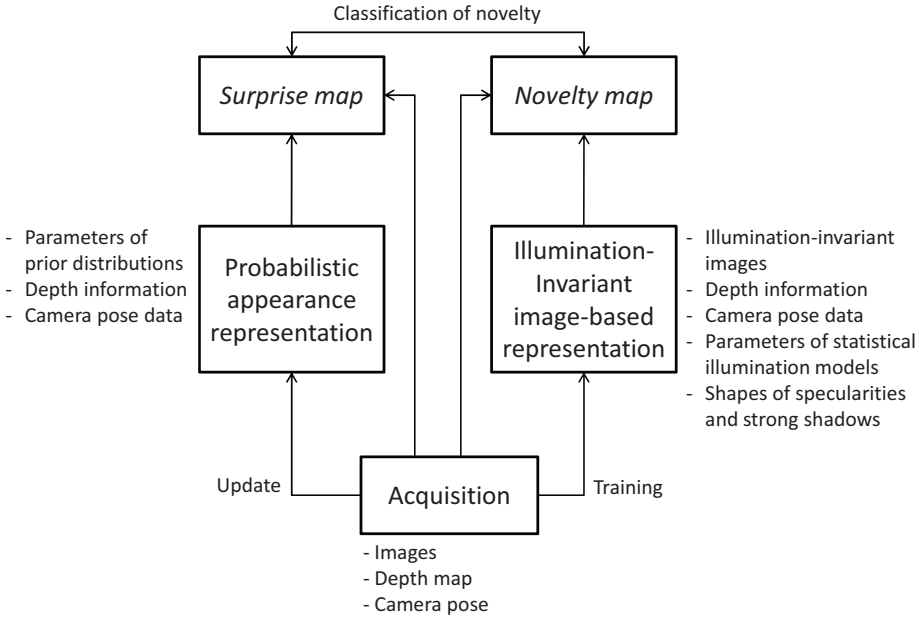


Figure 1.1: *An overview of the two types of environment representations proposed in this thesis and their application to novelty detection.*

1.1 Overview of the dissertation

This thesis describes two types of environment representations which coexist as components of an image-based internal model of the environment of a cognitive mobile robot. Techniques are presented for the detection of novel events from image data, while these two representations are used as prior knowledge of the robot's environment, as depicted in Figure 1.1.

The first type of environment representation presented in this thesis is a probabilistic appearance representation. It stores prior distributions describing the expectation and uncertainty of the appearance of the environment in pixel arrays at densely spaced viewpoints. In addition, a depth map and the camera pose is stored at each viewpoint. This data serves for the interpolation of the priors at intermediate viewpoints. The uncertainty of the appearance can arise from erroneous pose or depth estimates or from moving objects in the scene. The priors are used to assess the surprise level of each pixel in a new captured image. The representation is continuously updated from observations which the robot makes along its way through the environment. The surprise maps indicate all kind of novelty, including novelty from changing illumination.

The second type of environment representation is acquired over a longer period of time from training image sequences captured under varying lighting conditions. It consists of illumination-invariant images, which are recovered at densely spaced viewpoints from the training images. As the probabilistic appearance representation it stores depth maps and the camera poses at each viewpoint, so that illumination-invariant images can be interpolated at intermediate positions. Statistical models which describe the effects of lighting on the variation of intensity and color saturation are trained and used to compute a novelty measure for all pixels in a new captured camera image. Thus, illumination changes usually lead to low

values in the resulting novelty map.

The surprise and novelty maps can be combined to distinguish relevant novelty from irrelevant novelty. Examples for relevant novelty are in this thesis the addition or removal of new objects, which can trigger tidying-up-tasks or object-search-tasks in the robot's action planner. In general, it is important to detect lighting effects in the surprise maps. A sudden moving shadow in the image, e.g., can imply that an object or a human might enter the field of view of the robot's camera and can cross its way. Hence, the detection of the shadow can prepare a replanning of the robot's trajectory during navigation at an early stage. However, for object manipulation tasks, novelty from varying lighting conditions, is considered in this thesis as irrelevant.

Furthermore, this thesis presents a technique which, driven by surprise, facilitates the acquisition of object representations from local image features which can be used afterwards for object recognition.

The remainder of this thesis is structured as follows. In the next chapter some mathematical background is given on the inference of probability models. Besides, related work from different research areas like image-based rendering, illumination modeling, novelty and surprise detection is revisited. Chapter 3 describes a system which has been developed within this thesis for the interpolation of virtual images from an image-based representation which contains explicit view-dependent geometry information. The proposed probabilistic appearance representation is presented in Chapter 4 together with a method for surprise detection. In Chapter 5, the application of surprise detection to the autonomous acquisition of object representations is described. Chapter 6 treats illumination-invariant image-based environment representations and a technique for the detection of novel objects from images which is robust against varying illumination. Finally, chapter 7 concludes this thesis.

1.2 Contributions of the dissertation

In the following the main contributions of this thesis are stated.

A probabilistic appearance representation for mobile robots

The probabilistic appearance representation presented in this thesis is inspired by image-based scene representations but extends them in a way that the pixels of the reference views do not store raw intensity and color values but parameters of statistical models for intensity and color. This, on the one hand, allows a mobile robot to make a photorealistic prediction of the appearance of a 3D environment from images captured in the past. On the other hand, the robot can evaluate the uncertainty of the appearance of the environment in the internal representation and thus the consistency of the past observations. Modeling the uncertainty of the appearance of the environment is crucial for a better assessment of the novelty of new visual stimuli. Furthermore, image-based representations are appropriate for modeling static scenes. In this thesis, in turn, update rules are presented, which adapt the probabilistic appearance representation to dynamic environments.

Bayesian surprise detection

The method for surprise detection which is presented in this thesis is closely related to the probabilistic appearance representation of the 3D world. This allows for the computation of per-pixel surprise maps. Existing approaches for surprise detection are based on feature-based environment representations. Hence, objects which do not exhibit the type of features which the detector responds to cannot be detected. Furthermore, many models for visual attention and surprise presented in literature do not take into account the geometric properties of the environment and the resulting correspondences between visual stimuli acquired at different viewpoints.

Illumination-invariant image-based environment representations

This thesis presents an approach for the acquisition of illumination-invariant image-based scene representations, which represent the appearance of a 3D environment free of illumination effects. Existing approaches only consider the recovery of an illumination-invariant image at a static camera viewpoint.

Novel statistical and shape-based methods for modeling illumination effects

The detection of novel environment changes from images, which is based on statistical illumination models and thus is robust against illumination changes, is proposed for the first time in this thesis. Furthermore, the shape-based modeling of specularities and strong shadows allows for the identification of these illumination effects in new images and their suppression during novelty detection. This method has not been presented before.

Acquisition of object representations

Another main contribution is a method for the acquisition of object representations for object recognition. Regions of high values in a surprise map are used to isolate a new object in the environment and to selectively extract and learn features from the corresponding image region. Since this whole process is driven by surprise, the learning phase is triggered automatically when a new object is present in the environment. In contrast to existing approaches, the robot's manipulator does not have to interact with the object and no depth information is required for object segmentation.

Parts of this dissertation have been published in [MMBS09], [MBS+09], [MS10], [MBM+10], [MS11] and [MES11].

2 Background and Related Work

This chapter presents mathematical background on the inference of the parameters of probability distributions. These concepts are applied in the approaches for surprise detection and illumination-invariant novelty detection described in Chapters 4 and 6 of this thesis. Besides, a number of approaches from the fields image-based rendering, illumination modeling, change, novelty and surprise detection as well as autonomous acquisition of object representations are revisited.

2.1 Statistical learning

Robots process sensor data in order to get information about their surroundings. In real-world environments, sensor measurements usually do not reflect the true state of the environment but are afflicted by noise. Furthermore, the environment of a robot changes over time, e.g. due to new objects which appear in the scene or disappear, or due to illumination, which depends on the time of day. Noise and the dynamics of the world make it difficult to make reliable predictions about how the robot will perceive the environment in the future. The concepts of probability theory provide a way to represent and quantify the uncertainty which is inherent in the acquired sensor data. The variation of the appearance of the environment (e.g. the luminance) can be described by a continuous random variable x. A probability distribution $p(x)$ represents the probability density over a given value range of x. The probability that x lies in an interval $[a, b]$ is

$$p\left(x \geq a \wedge x \leq b\right) = \int_a^b p(x)\mathrm{d}x \tag{2.1}$$

In probability theory, non-parametric approaches have been presented to model probability distributions [CH67, DHS01]. Histogram methods, e.g., divide the value range of the random variable x into equally spaced distinct bins and count the number of samples which fall into each bin. The advantage of non-parametric density estimation is that no particular assumption is made with respect to the form of the probability distribution. Parametric approaches, in contrast, use models for probability distributions which depend on a small set of parameters and have a special functional form. Although these methods often provide an approximation of the true probability distribution of the random variable, the probability densities over the value range of the random variable are determined by specifying the parameter values. Storing the parameter values usually requires less memory than storing the density estimates in non-parametric approaches, which allows for a much more compact representation of the probability distribution.

In order to assess the probability of future observations, the robot has to learn its probability models of the world from its percept history. In case of parametric probability models this means that an optimal set of parameters has to be determined which explains the distribution of the observed samples best. In this section, Maximum Likelihood estimation and

Bayesian inference, two popular methods for learning the parameters of a given probability distribution, will be revisited.

2.1.1 Maximum Likelihood estimation

Maximum Likelihood estimation starts from a set of samples from a random variable x. The set of samples is denoted here by $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$. The goal is to find the parameter values $\mathbf{w} = \{w_1, w_2, \ldots, w_L\}$ of a probability distribution $p(x \mid \mathbf{w})$ which maximize the likelihood function $p(\mathcal{D} \mid \mathbf{w})$. In other words, out of all models which result from different parameterizations the model is chosen for which the set of samples achieves the highest probability. In the following, the approach is described on the basis of the multivariate Gaussian distribution and the gamma distribution as examples. The latter plays an important role for the statistical modeling of the illumination variation in camera images in the approach presented in Chapter 6 of this thesis.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ denote a set of N samples which are drawn from a multivariate Gaussian distribution

$$p_{\text{Gauss}}(\mathbf{x} \mid \mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} \left|\mathbf{\Sigma}\right|^{\frac{1}{2}}} \cdot \exp\left\{-\tfrac{1}{2}\left(\mathbf{x}-\mu\right)^{\mathrm{T}} \mathbf{\Sigma}^{-1} \left(\mathbf{x}-\mu\right)\right\}. \tag{2.2}$$

The random variable $\mathbf{x}$ is a D-dimensional vector $\mathbf{x} = [\xi_1, \xi_2, \ldots, \xi_D]^{\mathrm{T}}$, which is reflected in the normalization constant of the distribution. The parameters of the Gaussian model, which are estimated in the Maximum Likelihood approach, are the mean vector μ and the covariance matrix $\mathbf{\Sigma}$. $\left|\mathbf{\Sigma}\right|$ denotes the determinant of the covariance matrix.

It is assumed that the samples in $\mathbf{X}$ are independent and identically distributed (i.i.d.), that means that they are independently drawn from the same Gaussian distribution. Hence, the likelihood function for the Gaussian, which is to be maximized and which depends on the mean vector and the covariance matrix, is given by

$$l_{\text{Gauss}}(\mathbf{X} \mid \mu, \mathbf{\Sigma}) = \prod_{i=1}^{N} p_{\text{Gauss}}(\mathbf{x}_i \mid \mu, \mathbf{\Sigma}). \tag{2.3}$$

For probability distributions which belong to the exponential family [BS00, DHS01] it is more convenient to consider the natural logarithm of the likelihood function. Furthermore, the numerical stability of the calculations is improved, especially if some probabilities in (2.3) take very small values. The log likelihood function for the Gaussian is then

$$\ln l_{\text{Gauss}}(\mathbf{X} \mid \mu, \mathbf{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln\left|\mathbf{\Sigma}\right| - \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{x}_i-\mu\right)^{\mathrm{T}} \mathbf{\Sigma}^{-1}\left(\mathbf{x}_i-\mu\right). \tag{2.4}$$

Since the logarithm is a strictly monotonic increasing function, maximizing the likelihood function is equivalent to maximizing the log likelihood function. The mean vector μ_{ML} for

which the probability of the set of samples is maximum is computed by

$$\begin{aligned} \frac{\partial}{\partial \mu} \ln l_{\text{Gauss}}(\mathbf{X} \mid \mu, \mathbf{\Sigma}) &= \mathbf{0} \\ \sum_{i=1}^{N} \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mu) &= \mathbf{0} \\ \mu_{\text{ML}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i. & \end{aligned} \tag{2.5}$$

Estimating the covariance matrix $\mathbf{\Sigma}_{\text{ML}}$ from the sample data is a little bit more tedious. In [MN99] it is shown that a solution for $\mathbf{\Sigma}_{\text{ML}}$ can be obtained from the first differential of the log-likelihood function in (2.4) using Matrix Differential Calculus. An estimate for the covariance matrix in (2.2), which is symmetric and positive definite, as required for covariance matrices, is given by

$$\mathbf{\Sigma}_{\text{ML}} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \mu_{\text{ML}}) (\mathbf{x}_i - \mu_{\text{ML}})^{\text{T}}. \tag{2.6}$$

The Maximum Likelihood approach underestimates the covariances of the Gaussian distribution, especially for small set of samples. Considering various sets of samples drawn from the Gaussian distribution in (2.2), the expectation of the estimator in (2.6) is [Bis06]

$$\mathbb{E}\left\{\mathbf{\Sigma}_{\text{ML}}\right\} = \frac{N-1}{N} \cdot \mathbf{\Sigma}. \tag{2.7}$$

Thus, $\mathbf{\Sigma}_{\text{ML}}$ is biased with respect to the true covariance matrix $\mathbf{\Sigma}$. An unbiased estimator for the covariance matrix is provided by

$$\mathbf{\Sigma}_{\text{ML}}^{\text{u}} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \mu_{\text{ML}}) (\mathbf{x}_i - \mu_{\text{ML}})^{\text{T}}. \tag{2.8}$$

While the Maximum Likelihood estimates for the parameters of the Gaussian distribution can be directly computed from a set of samples using (2.5) and (2.8), an iterative method has to be applied in order to obtain the parameters of a gamma distribution. Let $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$ represent a set of N samples which are drawn from a gamma distribution

$$p_{\text{Gamma}}(y \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot y^{\alpha-1} \exp\left\{-\beta y\right\} \tag{2.9}$$

where $\Gamma(\alpha)$ denotes the gamma function

$$\Gamma(\alpha) = \int_0^{\infty} u^{\alpha-1} \exp\{-u\} \mathrm{d}u, \quad \alpha > 0. \tag{2.10}$$

Using this set of samples, Maximum Likelihood estimation provides a solution for the parameters α and β which determine the shape of the distribution. The log-likelihood function for the gamma distribution which is maximized is

$$\ln l_{\text{Gamma}}(\mathbf{y} \mid \alpha, \beta) = N\alpha \ln \beta - N \ln \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{N} \ln y_i - \beta \sum_{i=1}^{N} y_i. \tag{2.11}$$

The parameter β_{ML} for which the set of samples has the highest probability is obtained by

$$\begin{aligned} \frac{\partial}{\partial\beta} \ln l_{\mathrm{Gamma}}(\mathbf{y} \mid \alpha, \beta) &= 0 \\ \frac{N\alpha}{\beta} - \sum_{i=1}^{N} y_i &= 0 \\ \beta_{\mathrm{ML}} &= \frac{N\alpha}{\sum_{i=1}^{N} y_i}. \end{aligned} \tag{2.12}$$

Replacing the parameter β in (2.11) by (2.12) yields the log-likelihood function

$$\ln l_{\mathrm{Gamma}}(\mathbf{y} \mid \alpha) = N\alpha \left[\ln \alpha - \ln \left(\frac{1}{N} \sum_{i=1}^{N} y_i \right) \right] - N \ln \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{N} \ln y_i - N\alpha. \tag{2.13}$$

Setting the partial derivative of (2.13) with respect to the parameter α to zero provides the equation

$$\ln \alpha - \psi(\alpha) = \ln \left(\frac{1}{N} \sum_{i=1}^{N} y_i \right) - \frac{1}{N} \sum_{i=1}^{N} \ln y_i \tag{2.14}$$

with the digamma function

$$\psi(\alpha) = \frac{\frac{\mathrm{d}}{\mathrm{d}\alpha}\Gamma(\alpha)}{\Gamma(\alpha)}. \tag{2.15}$$

Since no closed-form solution can be found for (2.14), numerical methods have to be used in order to solve for α. In [CW69] the Newton-Raphson algorithm is applied to get the Maximum Likelihood estimate α_{ML}. α_{ML} is iteratively computed by

$$\alpha_{\mathrm{ML}} \leftarrow \alpha_{\mathrm{ML}} - \frac{\ln \alpha_{\mathrm{ML}} - \psi(\alpha_{\mathrm{ML}}) - \ln \left(\frac{1}{N} \sum_{i=1}^{N} y_i \right) + \frac{1}{N} \sum_{i=1}^{N} \ln y_i}{\frac{1}{\alpha_{\mathrm{ML}}} - \psi'(\alpha_{\mathrm{ML}})} \tag{2.16}$$

where $\psi'(\,\cdot\,)$ denotes the trigamma function, i.e. the derivative of the digamma function in (2.15). An initial estimate for α_{ML} can be taken as

$$\alpha_{\mathrm{ML},0} = \frac{1}{2 \left[\ln \left(\frac{1}{N} \sum_{i=1}^{N} y_i \right) - \frac{1}{N} \sum_{i=1}^{N} \ln y_i \right]}. \tag{2.17}$$

2.1.2 Bayesian inference

Maximum Likelihood estimation provides a very simple method to find the parameters of a probability distribution from a set of samples. If the set of samples is large, the parameter estimates are accurate and close to the true parameters of the probability model. This can be seen in (2.7), where the expression $\frac{N-1}{N}$ tends to 1 for large values of N. However, the downside is that Maximum Likelihood estimation only provides one probability model and ignores that there might also be other models with different parameterization which fit the set of samples as well. Especially in case of small sets of samples the probability distribution found by Maximum Likelihood estimation can be very different from the true probability

distribution of the random variable. Consider as an example a box which contains 50 red, 50 green and 50 blue balls. If three red balls are taken from the box, Maximum Likelihood estimation will come to the conclusion that the probability to take a red ball out of the box is 100% whereas the probability to get a green or blue ball is 0%. This result is obviously not true.

Statistical learning under the Bayesian paradigm is less exclusive and considers multiple hypotheses for the parameterization of the probability model. The uncertainty of the model parameters $\mathbf{w}$ is represented by a prior distribution $p(\mathbf{w})$. By observing a set of N samples $\mathbf{z} = \{z_1, z_2, \ldots, z_N\}$ drawn from a probability distribution $p(z)$, a posterior distribution over the model parameters is obtained by Bayes' formula

$$p(\mathbf{w} \mid \mathbf{z}) = \frac{p(\mathbf{z} \mid \mathbf{w}) \cdot p(\mathbf{w})}{p(\mathbf{z})}. \tag{2.18}$$

$p(\mathbf{z} \mid \mathbf{w})$ is a likelihood function. The posterior distribution is the belief distribution over the model parameters given the observed samples and has a smaller variance than the prior distribution. This means that the uncertainty with respect to the model parameters decreases the more samples of the random variable z are taken.

In Bayesian parameter learning it is often very convenient to use conjugate priors. Conjugacy means that the posterior distribution in (2.18) has the same functional form as the prior. This simplifies the calculus, since the Bayesian update in (2.18) results in a simple set of equations for the computation of the hyperparameters of the posterior subject to the hyperparameters of the prior and the sample data. The hyperparameters denote the parameters of the prior and posterior distribution.

In the following, the Bayesian inference of the parameters of a univariate Gaussian distribution

$$p(z \mid \mu, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} \cdot \exp\left\{-\tfrac{\lambda}{2}(z-\mu)^2\right\} \tag{2.19}$$

will be discussed as an example. It is assumed that both the mean μ and the precision λ of the Gaussian distribution are unknown. The precision is the reciprocal value of the variance. As a conjugate joint prior over both parameters the normal-gamma distribution

$$p(\mu, \lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)\sqrt{2\pi\sigma}} \cdot \lambda^{\alpha-\frac{1}{2}} \cdot \exp\left\{-\beta\lambda\right\} \cdot \exp\left\{-\tfrac{\lambda(\mu-\tau)^2}{2\sigma}\right\} \tag{2.20}$$

is chosen. Its hyperparameters are α, β, τ and σ. Note that the normal-gamma distribution is not separable with respect to the random variables μ and λ. While the first exponential term in (2.20) only contains the precision, the second one depends on both the mean value and the precision.

A set of samples $\mathbf{z} = \{z_1, z_2, \ldots, z_N\}$ is observed and used to compute the posterior

$$\begin{aligned}
p(\mu, \lambda \mid \mathbf{z}) &\propto \prod_{i=1}^{N} p(z_i \mid \mu, \lambda) \cdot p(\mu, \lambda) \qquad (2.21)\\
p(\mu, \lambda \mid \mathbf{z}) &\propto \lambda^{\frac{N}{2}} \cdot \exp\left\{-\tfrac{\lambda}{2}\sum_{i=1}^{N}(z_i-\mu)^2\right\} \cdot \lambda^{\alpha-\frac{1}{2}} \cdot \exp\left\{-\beta\lambda\right\} \cdot \exp\left\{-\tfrac{\lambda}{2} \cdot \tfrac{(\mu-\tau)^2}{\sigma}\right\}\\
p(\mu, \lambda \mid \mathbf{z}) &\propto \lambda^{\alpha'-\frac{1}{2}} \cdot \exp\left\{-\beta'\lambda\right\} \cdot \exp\left\{-\tfrac{\lambda}{2} \cdot \tfrac{(\mu-\tau')^2}{\sigma'}\right\}
\end{aligned}$$

where

$$\alpha' = \alpha + \frac{N}{2} \tag{2.22}$$

$$\beta' = \beta + \frac{1}{2} \cdot \sum_{i=1}^{N} \left(z_i - \frac{1}{N} \sum_{i=1}^{N} z_i \right)^2 + \frac{1}{2} \cdot \frac{\left(\frac{1}{N} \sum_{i=1}^{N} z_i - \tau \right)^2}{\frac{\sigma N + 1}{N}} \tag{2.23}$$

$$\tau' = \frac{\sigma \cdot \sum_{i=1}^{N} z_i + \tau}{\sigma N + 1} \tag{2.24}$$

$$\sigma' = \frac{\sigma}{\sigma N + 1}. \tag{2.25}$$

2.2 Image-based rendering

In computer graphics, the traditional approach for modeling a virtual environment is based on a geometric model of the scene. Textures are mapped on the primitives of the geometry meshes in order to give the surfaces of the objects a realistic appearance. A virtual camera, whose pose is determined by the user in interactive applications like games, renders images of the scene, mimicking the imaging process of a real camera in the real world. Local or global illumination models describe the interaction of the light rays emitted by one or several light sources with the surfaces of the objects and determine the shading of the objects. Traditional computer graphics approaches have advanced a lot and sophisticated techniques like ray tracing [Whi80, SJ00] and photon mapping [Jen01] have been presented.

Image-based modeling and rendering [SCK07] has been developed as an alternative approach to traditional geometry-based techniques for image synthesis with the goal of achieving photorealistic rendering results of complex real-world scenes using a captured set of camera images of the environment. The synthesis of virtual views of a scene does not require detailed knowledge of the material properties of the objects' surfaces nor the lighting, since the visual appearance of complex illumination effects can be directly transfered from the camera images. From a signal processing view, image-based scene representations can be seen as a sampled discrete representation of the continuous plenoptic function. The seven-dimensional plenoptic function $P_7(V_x, V_y, V_z, \theta, \phi, \lambda, t)$, as it is defined in [AB91], measures the "intensity of light rays passing through the center of the pupil [at a 3D position (V_x, V_y, V_z),] at every possible angle (θ, ϕ), for every wavelength λ, at every time t". Since signals in high-dimensional spaces are difficult to handle, several assumptions are made in order to eliminate some of the parameters which the plenoptic function depends on. In [MB95], e.g., static environments illuminated by monochromatic light are considered, which results in a five-dimensional plenoptic function $P_5(V_x, V_y, V_z, \theta, \phi)$. Image-based rendering techniques reconstruct a "continuous representation of the plenoptic function from discrete samples" [SCK07] by interpolating the intensities of light rays captured by the virtual camera.

Over the past two decades, image-based rendering has received much attention since it allows for a photorealistic visualization of complex environments with little computational effort. Numerous image-based scene representations have been proposed, which all form a sampled representation of a plenoptic function with reduced dimensionality. They are classified in the image-geometry continuum (IBR continuum) [Len98, SKC03]. As illustrated in

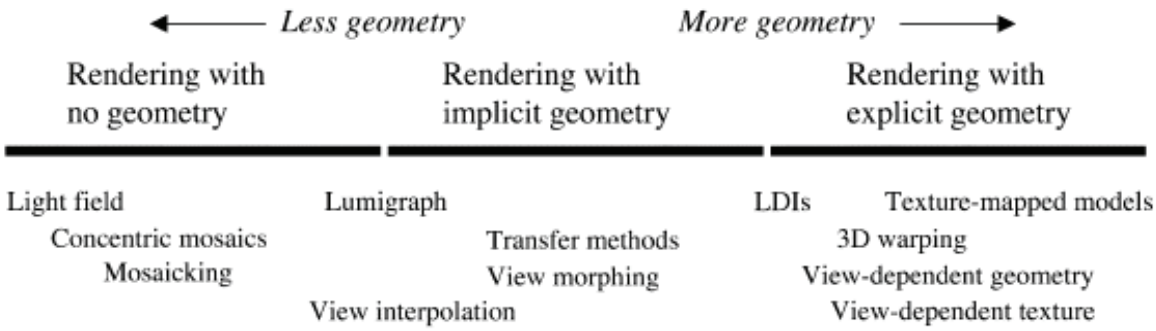


Figure 2.1: *IBR continuum with three categories for the classification of image-based rendering techniques and image-based representations (source: [SKC03]).*

Figure 2.1, the main categories in the IBR continuum are "Rendering with no geometry", "Rendering with implicit geometry" and "Rendering with explicit geometry". In the following, selected representatives from these categories will be discussed.

2.2.1 Rendering with no geometry

Unlike the term "Rendering with no geometry" might suggest, the approaches presented in the following do use a geometric model of the environment, which is, however, a very coarse approximation of the true scene structure (geometric proxy). The model is not recovered from intensity correspondences between images or range sensor data but is defined by a simple parametric surface. Light fields [LH96], e.g., which are captured by an array of cameras arranged in a regularly spaced planar grid, assume that the whole scene resides on a plane at a given distance from the cameras (focal plane). In many cases the scene consists of a single object. The cameras capture a collection of light rays. Each ray is defined by the indices (u, v), which address the cameras in the 2D grid, and the indices (s, t), which describe the intersection of the ray with the focal plane. Effectively, the five-dimensional plenoptic function is simplified to a four-dimensional plenoptic function $P_4(u, v, s, t)$, under the assumption that the viewpoint of the virtual camera is always outside a bounding box around the modeled object.

The synthesis of a virtual image can be understood as the interpolation of the intensities of light rays which are captured by the virtual camera from the intensities of the real light rays. Each point on the focal plane is seen in multiple camera images, which results in a bunch of rays running between the point and the projection centers of the cameras. For interpolation the nearest neighboring rays in the light field are chosen.

The Lumigraph [GGSC96] is an image-based representation which is very similar to light fields. As a difference to light field rendering Lumigraph rendering incorporates a more accurate geometry model of the scene for intensity interpolation and compression. Although the Lumigraph is indeed an image-based representation with explicit geometry, it is arranged between the categories "Rendering with no geometry" and "Rendering with implicit geometry" in the IBR continuum in Figure 2.1, due to "its strong similarity with the light field" [SCK07].

Instead of capturing images in a regular planar two-dimensional grid, the images can also

be captured by cameras moving along planar concentric circles. This representation is called Concentric Mosaics and is proposed in [SH99]. Two setups are possible in which the cameras are mounted either in tangential or normal direction to the circle. The virtual images are synthesized by interpolating pixel columns from pixel columns of nearby reference images. For the interpolation of the slices it is often assumed that the scene is infinitely far away from the acquisition setup. Another common assumption is that the geometry of the scene is the surface of a cylinder around the Concentric Mosaics. Fast model acquisition and high computational efficiency make Concentric Mosaics attractive for image-based rendering. In [BMS08] a scheme is presented for the progressive synthesis of virtual images from a Concentric Mosaics representation which is streamed over a network and optimized with respect to storage rate, distortion, transmission rate and decoding complexity [BS08a, BS08b].

Due to the heavy constraints on the regularity of the structure of image-based scene representations used for rendering with no geometry, these representations have a limited applicability in the field of mobile robots. Although it is shown in [DW11] that light fields can be used for modeling the appearance of the sea ground by Autonomous Underwater Vehicles, it is hardly impossible for mobile robots navigating on wheels in indoor environments to capture image-based representations on perfect circles.

Furthermore, the very coarse approximation of the scene geometry requires that the camera views be captured in a very dense manner which leads to a vast amount of data and high memory consumption for the storage of the environment representation. To this end, methods have been developed for the efficient compression of image-based scene representations [SKC03].

2.2.2 Rendering with implicit geometry

The approaches revisited in the following use pixel correspondences between a small number of reference images to synthesize novel virtual images. Since a geometry model of the scene in terms of 3D point clouds or 3D meshes is not available, they are embraced by the term “implicit geometry” [SKC03].

The method in [CW93] uses two input images and computes virtual images at arbitrary viewpoints between them using dense optical flow. The best results are achieved if the two input images are close to each other so that ambiguities in the correspondences between them are avoided.

Instead of using dense optical flow, virtual images can also be interpolated from sparse point correspondences between two input images and the resulting fundamental matrix [Fau93], as shown in [LF94]. The synthesis of a view from sparse point correspondences leads to holes in the virtual image. To avoid these black pixel regions a reverse mapping is done which finds the image correspondences in the two reference views for each pixel in the virtual image. In this mapping epipolar constraints between the two reference views and the virtual view are exploited.

There might be cases where the two epipolar lines in the virtual image coincide. In these degenerate cases it is useful to add a third reference image. View interpolation is then done [AS97] using the trifocal tensor [HZ03] which describes the structural relationship between the three reference views.

2.2.3 Rendering with explicit geometry

In [CTCS00] the minimum sampling rate is analyzed for light fields. The minimum sampling rate indicates how dense the cameras in the two-dimensional grid have to be arranged such that virtual images can be interpolated without aliasing artifacts. One important finding is that the minimum sampling rate is subject to the minimum and maximum depth of the scene and does not depend on the depth variation in the scene Furthermore, Chai et al. [CTCS00] consider plenoptic sampling in joint image and geometry space and draw important conclusions with respect to the minimum sampling curve, which relates the number of images in the representation to the number of depth layers of a geometry model which is stored in addition to the images. They find that the number of images, and thus the density of the camera views, strongly decreases with the number of depth layers. Hence, the finer and the more accurate a geometry model used for view interpolation is, the less images have to be acquired of the scene. The reduction of the amount of image data makes rendering techniques using an explicit geometric model very attractive.

In [DTM96] and [DYB98] an approach for the realistic visualization of virtual 3D models of buildings is presented. The approach is closely related to traditional geometry-based methods for image synthesis. However, view-dependent texture mapping is used, where multiple textures from different viewpoints are mapped on the same surface of the geometric model. This method allows for capturing effects like specular highlights, whose appearance is subject to the viewpoint of the observer.

In addition to view-dependent textures, view-dependent geometry models have become very popular [PCD+97, KS04]. View-dependent geometry models in terms of depth maps take into account that the depths which are estimated from stereo images are only valid in a small viewpoint region if non-Lambertian effects are exhibited.

An intersting aspect is also to relax the strong constraints which are imposed on the structure of the camera views in image-based representations like light fields and Concentric Mosaics. In [HKP+99] a system is presented which interpolates novel virtual images from an unstructured image-based representation acquired by a handheld camera. The poses of the captured camera views are estimated using a structure-from-motion approach which computes the fundamental matrices between image pairs. For view synthesis the scene geometry is approximated by one or more multiple planes and the camera images are mapped on them as textures. A system for the reconstruction of a scene from a handheld moving camera with similar modules is presented in [PvGV+04]. In [ESK03] an unstructured image-based representation is acquired using a handheld multi-camera system. Three criteria are proposed for the selection of a subset of camera views for view synthesis. They consider the proximity of the reference cameras to the virtual camera, the viewing directions of the reference cameras and the virtual camera as well as visibility aspects. The local geometry of the scene is reconstructed subject to the viewpoint of the virtual camera by fusing the depth maps of the selected reference views. In [ESNK06] an improved version of the system is presented which recovers coarse depth information in an off-line step and refines the geometry model during view synthesis by a tile-based photoconsistency check.

Buehler et al. define in [BBM+01] a list of properties which an image-based rendering system should have. Besides, they propose a rendering approach which generalizes rendering techniques with no knowledge about the scene geometry and techniques for image-based rendering using explicit geometric models.

Image-based representations with view-dependent geometry have also become very popular for view synthesis in Free Viewpoint Video and Three-dimensional Television (3DTV). Zitnick et al. propose in [ZKU+04] a system which captures multiple synchronized video streams from different viewpoints. A two-layered image-based representation which treats object boundaries in a separate layer is acquired to reduce artifacts in virtual views. Besides, a stereo matching algorithm which is based on color segmentation provides high-quality depth maps, which allows for a small number of calibrated video cameras capturing the scene. A similar multi-view video plus depth representation which consists of multiple layers is presented in [SMD+08] for 3DTV applications. Taguchi et al. present in [TTN08] a system which adopts the layered light-field representation in [TN06]. The pixel colors in virtual views are determined from the reference cameras, which are arranged in a two-dimensional grid, using a depth-from-focus method.

Since image-based representations with explicit geometric models allow for unstructured input and reduce memory consumption, they are very convenient for realistic environment modeling for mobile robots. The algorithms in Chapter 3 and the representation in Chapter 4 of this thesis are inspired by the approaches in [ESK03], [PvGV+04] and [ZKU+04]. However, compared to image-based representations, the environment model in Chapter 4 stores the parameters of probability distributions which do not only represent the expectation about the appearance of the scene but also the uncertainty. It is shown that this concept allows for the detection of surprising events in the robot's surroundings.

2.3 Illumination modeling and intrinsic images

A popular illumination model in computer graphics for the computation of the intensity reflected by a point on a surface which is a nonperfect reflector is proposed by Phong in [Pho75]. In [FvDFH96] Phong's model is given by

$$I_\lambda = I_{\mathrm{a}\lambda} k_\mathrm{a} O_{\mathrm{d}\lambda} + f_\mathrm{att} I_{\mathrm{p}\lambda} \left[k_\mathrm{d} O_{\mathrm{d}\lambda} \left(\vec{N} \cdot \vec{L} \right) + k_\mathrm{s} \left(\vec{R} \cdot \vec{V} \right)^n \right]. \tag{2.26}$$

The reflected intensity I_λ is the sum of three reflection components which describe the reflection of ambient light, the diffuse reflection and the specular reflection of light emitted by a light source, respectively. I_a is the intensity of the ambient light and k_a indicates the proportion of the ambient light which is reflected by the surface point. $O_{\mathrm{d}\lambda}$ represents the diffuse color of the surface point. The intensity of the light source is $I_{\mathrm{p}\lambda}$ and f_att is an attenuation factor which accounts for the attenuation of the light intensity at the surface point with increasing distance of the light source from the surface or in case of occlusions. k_d and k_s describe the proportion of light which is reflected under diffuse and specular reflection, respectively. As illustrated in Figure 2.2, the vectors $\vec{L}$ and $\vec{N}$ indicate the opposite direction of the incident light and the direction of the normal of the surface, respectively. The vectors $\vec{R}$ and $\vec{V}$ represent the direction of specularly reflected light and the direction of an observer's viewpoint, respectively. $\vec{R}$ depends on the surface normal and the opposite direction of the incoming light as

$$\vec{R} = 2\vec{N} \left(\vec{N} \cdot \vec{L} \right) - \vec{L}. \tag{2.27}$$

The specular exponent n determines how strong the intensity of specularly reflected light falls off subject to the angle between the direction of the reflected light and the direction of

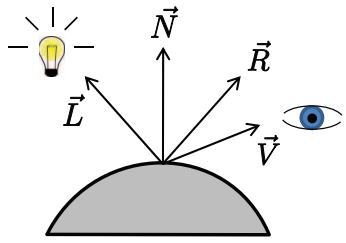


Figure 2.2: *Light which is emitted by a light source and illuminates the scene from the opposite direction of $\vec{L}$ is mirrored about the surface normal $\vec{N}$ under specular reflection and leaves the surface point in direction $\vec{R}$. The Phong illumination model takes into account that the intensity of the specular component falls off as the angle between $\vec{R}$ and the direction of the observer's viewpoint $\vec{V}$ gets larger.*

the observer's viewpoint.

In environments with objects that have Lambertian surfaces and thus only reflect diffuse light the specular term in (2.26) can be omitted. Assuming that there is no ambient light and that the intensity is constant along a ray of light, a simple illumination model is given by

$$I_\lambda = k_\mathrm{d} O_{\mathrm{d}\lambda} \cdot I_{\mathrm{p}\lambda} \left(\vec{N} \cdot \vec{L} \right) = R_\lambda \cdot L_\lambda. \tag{2.28}$$

Thus, the intensity of light which an observer perceives at a surface point depends, on the one hand, on intrinsic material properties like the spectral reflectance of the surface $R_\lambda = k_\mathrm{d} O_{\mathrm{d}\lambda}$ and, on the other hand, on the characteristics of the light source, which are summarized in $L_\lambda = I_{\mathrm{p}\lambda} \left(\vec{N} \cdot \vec{L} \right)$ and include its intensity, the spectrum of the emitted light as well as the position of the light source with respect to the object surface. Barrow and Tenenbaum suggest in [BT78] that the separation of intrinsic scene characteristics from an intensity image plays a central role in early visual processing. Intrinsic scene characteristics are range (scene depth), surface orientation, reflectance and incident illumination. The representation of the intrinsic properties of the scene in different images is denoted by *intrinsic images*. The recovery of the intrinsic features is considered as a crucial preprocessing step for higher-level scene analysis. However, the computation of intrinsic images from an intensity image is in general very challenging since the estimation of both R_λ and L_λ from a single equation is an ill-posed mathematical problem.

To make the problem tractable, Freeman and Viola consider in [FV97] two special cases in which either the reflectance is uniform across a surface varying in shape or the reflectance changes across the image but the surface is flat and thus exhibits uniform shading. A prior distribution is placed over the magnitudes of the image gradients and the goal is to determine in a probabilistic approach whether an image shows only changes in shape or only changes in reflectance. In experiments the performance of the algorithm is compared to the classification results by human subjects.

The approach in [TFA05] is less restrictive with respect to the reflectance and shape properties of the scene and uses machine learning techniques to classify each gradient in an intensity image as a reflectance or shading gradient. For the classification of the gradients

both color and intensity patterns are analyzed. The color vectors of two adjacent pixels which show surface points of the same intrinsic color under different illumination are usually collinear. Else, if the two color vectors are linearly independent, and thus the chromaticity changes, a reflectance change is exhibited. However, if the scene exhibits only grey tones a reflectance change can also lead two two adjacent pixels with collinear color vectors. Hence, using only color information does not allow for an unambiguous decision in favor of a reflectance or a shading gradient. Therefore, to make the classification more reliable, the approach in [TFA05] also analyzes the output of nonlinear filters applied to intensity patches. The filter output compared to a threshold is a weak classifier which assigns the label 1 to reflectance gradients and the label -1 to shading gradients. In a training phase the coefficients of the filter and the threshold are determined using intensity patches from synthetic images. After learning the classifier, the gradients in an intensity image which are classified as shading gradients are set to zero while the reflectance gradients are left unchanged. A reflectance image is computed by pseudo-inverse filtering as proposed in [Wei01]. The strong point of the approach in [TFA05] is that, once the classifier is trained, the reflectance and the shading image can be computed directy from one single intensity image. However, the approach also requires a very low false positive rate during classification, which is in general difficult to achieve with the used machine learning techniques. The misclassification of a single gradient can lead to strong artifacts in the integrated reflectance image.

In [FDL04], an approach is presented for the recovery of an illumination-invariant image which is computed in the logarithmic 2D chromaticity color space $\{\log\left(\frac{G}{R}\right), \log\left(\frac{B}{R}\right)\}$, where R, G and B denote the red, green and blue primaries from the RGB color space. Finlayson et al. note that the chromaticity values acquired from the same scene under different illumination conditions accumulate along parallel straight lines for a given camera. Hence, a grayscale illumination-invariant image can be recovered by projecting the chromaticity values on a straight line which is perpendicular to the set of lines. The orientation of the line is found by testing several hypothesis for the orientation and evaluating the entropy of the recovered grayscale image. The orientation for which the entropy is minimum is selected. On the one hand, the approach offers a simple and fast method for the computation of illumination-invariant images. On the other hand, the approach poses several constraints on the scene, the light source and the camera. As in [TFA05], the scene must consist of objects with exclusively Lambertian surfaces. Furthermore, the light source must emit light with a spectral power density according to Planck's law [Pla00]. Besides, the three sensitivity functions of the camera sensor should be narrowband. These conditions can be relaxed to a certain degree, however, the approach performs poorly if they are ignored. In [FHLD06], it is shown that a full shadow-free color image can be computed by extending the grayscale illumination-invariant image to an equivalent 2D chromaticity representation and, finally, to a RGB color image.

As already mentioned before, the computation of an illumination-invariant image and the corresponding shading image from one single intensity image is a difficult problem since it is ill-posed in a mathematical sense. Therefore, Weiss looks at an easier version of this problem in [Wei01] by considering not only one single camera image but a series of camera images taken of a static scene at the same viewpoint and under different lighting conditions. Modeling T intensity images $I_t, t = 1, \ldots, T$ as a decomposition in their reflectance part R

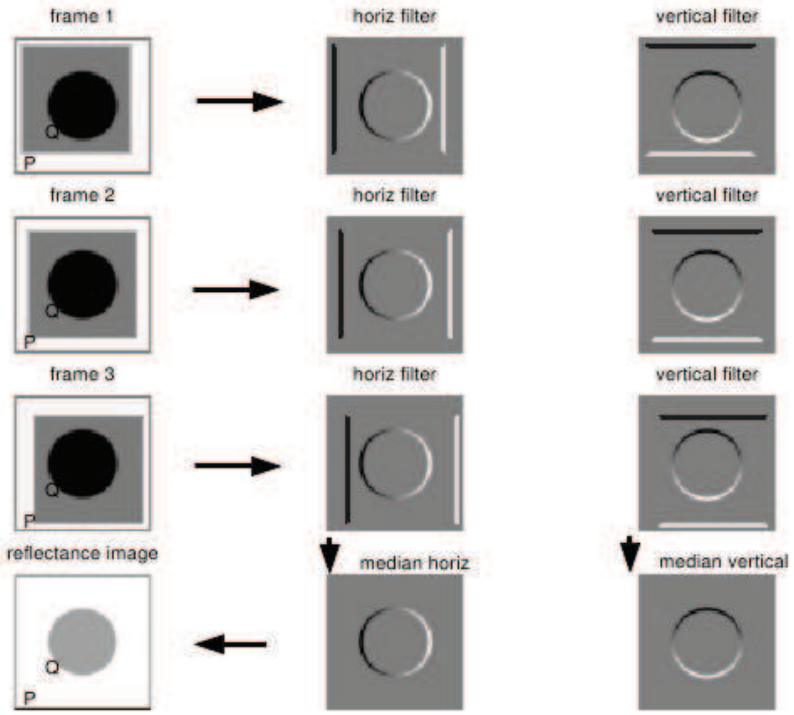


Figure 2.3: *An illustration of the computation of a reflectance image from a sequence of multiple images of a scene taken under different illumination conditions. (source: [Wei01])*

and their shading parts $L_t, t = 1, \ldots, T$ results in a system of equations given as

$$I_1 = R \cdot L_1 \tag{2.29}$$
$$I_2 = R \cdot L_2 \tag{2.30}$$
$$\vdots$$
$$I_T = R \cdot L_T. \tag{2.31}$$

This is still an underconstrained system of equations. The number of unknowns (the reflectance and the T shading images) exceeds the number of equations by 1. Hence, with one additional constraint the system of equations has a unique solution. In [Wei01], it is therefore assumed that the horizontal and vertical shading gradients are sparsely distributed in the images and mostly close to zero. This is a valid assumption since in previous works it has already been shown that the outputs of derivative filters applied to natrual images are sparse and can be modeled by a Laplacian distribution [OF96, Sim97]. Weiss shows that the reflectance gradient at a given image location can be computed by the median of the intensity gradients at that location in all camera images. A reflectance image is recovered by pseudo-inverse filtering from the horizontal and vertical reflectance gradients. Figure 2.3 illustrates the algorithm. As shown in the bottom row, the pixel-wise median of the horizontal and vertical image gradients removes the edges of the shadow over the gray circle whose position varies in the frames 1 to 3. After the pseudo-inverse filtering step, a reflectance image is obtained which is free of illumination effects.

In [Wei01], grayscale reflectance images are computed. However, applying the method to

the three color channels of the camera images, a color reflectance image is obtained. The requirements for a high-quality reflectance image are that both the camera and the scene are static during the acquisition of the images. Furthermore, the illumination in the captured images should vary as much as possible such that all illumination effects are removed in the reflectance image. Illumination effects which are exhibited in the majority of the camera images will also be visible in the reflectance image. Since the algorithm does not perform a hard classificaton of the gradients in reflectance and shading gradients as in [TFA05], there are no artifacts due to misclassification. Due to the natural appearance of the illumination-invariant images, the algorithm in [Wei01] has been chosen as the basis for the computation of the illumination-invariant image-based environment representation descibed in Chapter 6. One challenge which arises from the acquisition of image sequences with mobile robot platforms, as considered in this thesis, is the accurate registration of the images taken under different lighting conditions.

A common assumption which is made in all approaches for the computation of intrinsic images revisited in this section is that the shading of the scene can be described by the Lambertian illumination model in (2.28). However, to handle general real-world scenes, a number of approaches have been presented for the separation of specular reflections and interreflections from diffuse reflections in images. The diffuse and the specular component of a scene can be modeled as separate layers which are superimposed in the image formation process. If a camera moves with constant velocity on a linear trajectory which is parallel to the camera's image plane, it is shown in [SKS+02] that the motion of a surface point in the captured images is subject to the epipolar constraint. The motion of a surface point in the images perceived under specular reflection, in turn, does not. Specular highlights and reflections are perceived at a virtual depth behind the surface, depending on the surface curvature, the surface orientation and the camera distance. In [TKS03] a method is presented for the estimation of the depths of the diffuse layer and the reflection layer of a scene, while in contrast to [SKS+02], the type of reflection is not limited to specular highlights.

The approach in [STT+11] uses specular cues for the detection of screws in a cluttered bin and the estimation of their pose in an industrial assembly process. Shroff et al. note, as Swaminathan et al. in [SKS+02], that the position of a specularity does not vary much with the position of the light source if the curvature of the surface is high. Hence, surfaces with high curvature (like on a screw) can be detected using a multi-flash camera [RTF+04], which consists of an image sensor and several LEDs which are uniformly placed around the sensor. With the LEDs flashing light in a sequential order, several views are captured under varying light direction. By matching specular features of a screw in multiple views taken by the multi-flash camera at different viewpoints, the pose of the screw is estimated and passed to a grasping system.

Levin et al. propose in [LZW04] a method for the separation of reflections on specular transparent surfaces from the scene behind the surface. The reflection and the distant scene are treated as two superimposed statistically independent images, which are separated using Independent Component Analysis.

A method for the removal of specularities in facial images is presented in [LB05]. First, the image is processed by Luminance Multi-Scale Retinex in [FBBC97], which is based on the Retinex theory for modeling color constancy [LM71]. Using the luminance and saturation component of the processed image, the centers of the specularities are found and used as a seed point for a wavefront propagation algorithm, which stops at the border of the specular-

ity or the object. This provides the specularity regions in the image, which are then colored from the boundaries towards the interior with the average color of neighboring pixels just outside the region.

2.4 Illumination-invariant change detection

The segmentation of an object from a background scene can be tremendously impaired if the illumination in the background image differs from the illumination in the image containing the additional new object. To this end, a lot of research has been done in the area of change detection under varying lighting conditions [RAAKR05]. One of the earliest straightforward techniques to make image-based object segmentation robust to illumination changes was to normalize the intensity values in two images so that they have the same mean and the same variance. This accounts for global illumination changes.

Under the assumption that the illumination of the scene causes patterns with lower spatial frequencies than the reflectance of the scene exhibits, the reflectance and the shading component can be separated by homomorphic filtering [ADMT01]. This procedure is similar to the decomposition of a camera image into intrinsic images, as described in Section 2.3. Using the reflectance components for change detection, illumination effects do not influence the segmentation of the object of interest. However, as already pointed out in Section 2.3, the separation of reflectance and shading from single images is in general very difficult and the approach in [ADMT01] fails if pronounced shadow edges are present in the images.

In [XRB04] an algorithm for the suppression of sudden illumination changes between subsequently captured camera frames is presented. The approach uses the Phong illumination model as described in Section 2.3 and assumes monotone and nonlinear camera response function as well as locally constant but spatially varying illumination. It is shown that the sign of the difference between two pixel values is the same across global illumination changes.

Another method for illumination-invariant image analysis is the transformation to an appropriate color space as a preprocessing step. In [CGZ08] a perception-based color space for illumination-invariant image processing is presented and used for object segmentation and image inpainting. The transformation of a color vector in XYZ color space to a color vector in the color space proposed in [CGZ08] is chosen such that difference vectors between colors are constant under changing illumination conditions. Furthermore, the length of a difference vector matches the perceptual distance between the two colors.

Color space transformations have become popular in mobile robot applications for a vision-based segmentation of the environment by colors [SS09] because they are computationally cheap and therefore very fast. In [SS07] the objects in a mobile robot's environment, which is rich in colors, are segmented in a spherically distributed color space which is also used in [MMHM01]. The transformation of a RGB color vector to this color space, also known as

Spherical Coordinate Transform (SCT), is given by

$$I = \sqrt{R^2 + G^2 + B^2} \tag{2.32}$$

$$\phi = \tan^{-1}\left(\frac{G}{R}\right) \tag{2.33}$$

$$\theta = \cos^{-1}\left(\frac{B}{I}\right), \tag{2.34}$$

where I is intensity and the angles ϕ and θ represent color independently from intensity. The Euclidean distance of two tupels (ϕ_1, θ_1) and (ϕ_2, θ_2) is used for color matching.

The approach in [OH07] proposes the normalized correlation of corresponding intensity derivatives in two images as a measure for changes which is robust against varying illumination. Be $\mathbf{\Delta}_1 = \left(\frac{\partial I_1(x,y)}{\partial x}, \frac{\partial I_1(x,y)}{\partial y}\right)^{\mathrm{T}}$ and $\mathbf{\Delta}_2 = \left(\frac{\partial I_2(x,y)}{\partial x}, \frac{\partial I_2(x,y)}{\partial y}\right)^{\mathrm{T}}$ the intensity gradients in two equally sized images at pixel location (x, y). Then the normalised correlation between the two gradients describes the cosine of the angle θ between them and is given by

$$\rho' = \cos\theta = \frac{\mathbf{\Delta}_1^{\mathrm{T}} \cdot \mathbf{\Delta}_2}{\left\|\mathbf{\Delta}_1\right\| \cdot \left\|\mathbf{\Delta}_2\right\|}, \tag{2.35}$$

where $\left\|\mathbf{\Delta}\right\| = \sqrt{\mathbf{\Delta}^{\mathrm{T}} \cdot \mathbf{\Delta}}$. The Normalized Gradient Correlation (NGC) coefficient ρ' is robust against illumination variations of the form $I_2(x, y) = \alpha \cdot I_1(x, y) + \beta$, where α is a scaling factor and β an additive constant. If the position of the light source changes, the edges of a 3D object change their appearance due to shading. Depending on how much light falls on the surface at the object edge, it might appear either brighter or darker than the background. Consequently, the sign of the intensity gradients at the object edge changes with varying shading. To account for this, the algorithm in [OH07] uses a non-linear absolute value operation in the computation of the inner product of (2.35). A robust change measure is obtained by evaluating the correlation in block neighborhoods which are bounded horizontally by $x_{\min}$ and $x_{\max}$ and vertically by $y_{\min}$ and $y_{\max}$. The correlation coefficient results in

$$\rho = \frac{\sum_{x=x_{\min}}^{x_{\max}} \sum_{y=y_{\min}}^{y_{\max}} \left| \frac{\partial I_1(x,y)}{\partial x} \frac{\partial I_2(x,y)}{\partial x} + \frac{\partial I_1(x,y)}{\partial y} \frac{\partial I_2(x,y)}{\partial y} \right|}{\sqrt{\sum_{x=x_{\min}}^{x_{\max}} \sum_{y=y_{\min}}^{y_{\max}} \left(\left(\frac{\partial I_1(x,y)}{\partial x}\right)^2 + \left(\frac{\partial I_1(x,y)}{\partial y}\right)^2 \right) \sum_{x=x_{\min}}^{x_{\max}} \sum_{y=y_{\min}}^{y_{\max}} \left(\left(\frac{\partial I_2(x,y)}{\partial x}\right)^2 + \left(\frac{\partial I_2(x,y)}{\partial y}\right)^2 \right)}}. \tag{2.36}$$

The smaller ρ the higher is the probability to find changes which are not due to illumination but due to new/disappeared objects or people in the scene. In [OH07] the correlation of gradient structures is analyzed on multiple scales, i.e. in neighborhoods of varying size. An efficient implementation is proposed to achieve low computation times. The approach is simple and fast. However, it cannot cope with high frequent illumination effects like shadow borders since these effects also change the gradient structures and lead to a small correlation coefficient. Futhermore, specularities pose a problem to the algorithm.

2.5 Attention models, novelty and surprise detection

Modeling human attention has aroused great interest across various research disciplines. Computational approaches which measure how interesting a part of a signal is are useful

for the implementation of attentional mechanisms on robots to extract relevant information from the acquired sensor data for their actions. In electrical engineering and computer science, attention models can be used for the compression of images and videos since they predict which parts of an image will probably receive much attention and which will not. Psychologists and neuroscientists investigate attentional mechanisms in the human brain to get insights in the role of attention in visual search. In the following, various approaches from the different disciplines are revisited.

2.5.1 Saliency-based visual attention

A common notion is that an image region attracts the attention of a human observer if it is salient, i.e. if it exhibits visual features like color, intensity and orientation which are different from the features in neighboring regions. In [Li99], a model is proposed which makes contextual influences between the neurons in the visual cortex area "V1" responsible for the pop out effects of visual stimuli which differ from the stimuli in the surrounding region. Horizontal connections between nearby pyramid cells suppress the cells' responses if the stimuli are similar or excite the cells if the stimuli are different.

Itti et al. propose in [IKN98] a model of saliency-based bottom-up visual attention which is related to the "feature integration theory", which states that the conjunction of separable primary visual features needs attention [TG80]. In [IKN98] color, intensity and orientation are considered as primary visual features. In a later work [IK01], motion, stereo disparity, shape from shading etc. are named as additional features. The model presented in [IKN98] is depicted in Figure 2.4.

After filtering the input image, center-surround differences are computed in the color, intensity and orientation maps on multiple scales. Center-surround differences are computed across different scales while a pixel of a fine scale is the center and the corresponding pixel on a coarser scale describes the surround region. Color information is encoded by red-green and blue-yellow opponencies. Orientation cues are provided by the output of Gabor filters with the prefered orientations $0°$, $45°$, $90°$ and $135°$ [GBG+94]. Since the numbers of the center-surround differences are in different value ranges for different feature types, they are normalized to a fixed value range. Conspicuity maps are computed by multiplying the center-surround differences by their squared difference from the mean value across the image. This suppresses center-surround differences which are close to the mean value and amplifies center-surround differences which are much larger. Finally, the conspicuity maps of the feature channels are combined to a saliency map. Several highly salient regions in the map compete for the focus of attention while the region associated with the fastest firing neurons wins (winner-takes-it-all, WTA). Then, an inhibition-of-return mechanism locally inhibits the area of the focus of attention, which leads to a shift of attention to the next salient location in the map and prevents that the focus of attention immediately returns to the current location.

Attention models are used in active vision to position a robot's camera based on the feedback from the attention model about interesting, salient objects in the environment (see [BK11] for a survey). Using graphical processing units (GPUs), saliency maps at VGA resolution can be computed within a few milliseconds, which allows for real-time attentional control in robotic applications [XPKB09].

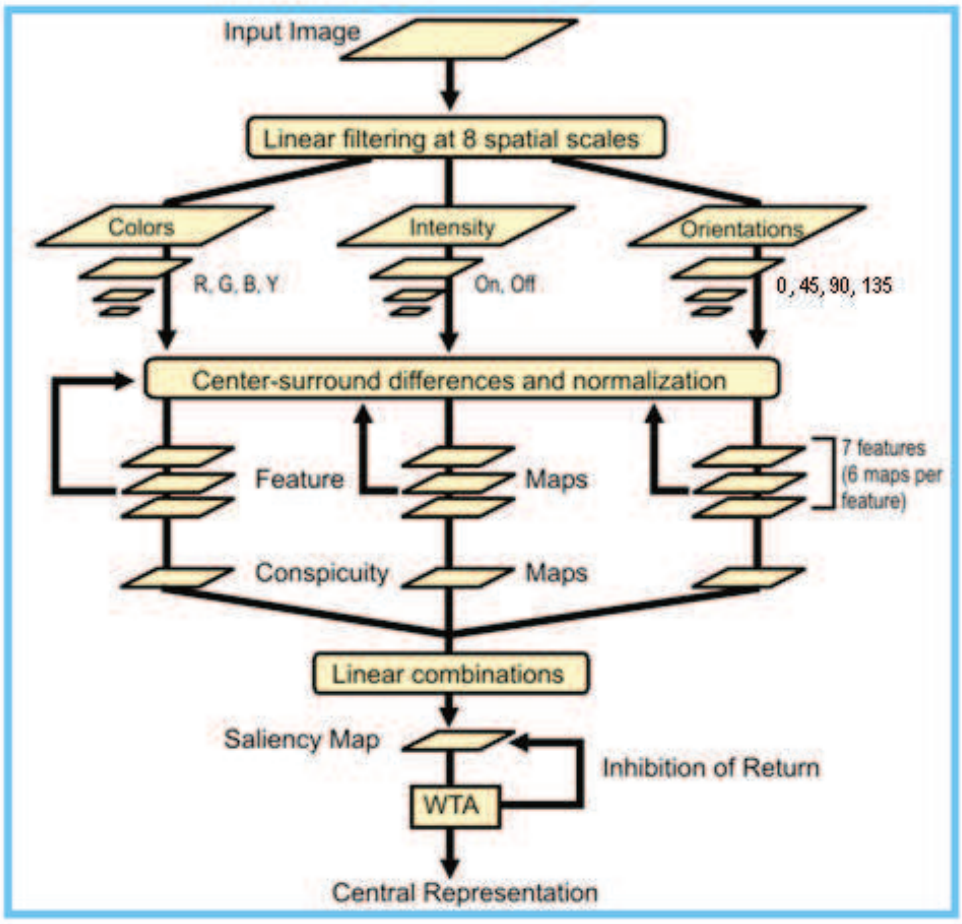


Figure 2.4: *The model of saliency-based visual attention proposed in [IKN98] (source: [mba]).*

In video coding and streaming applications, saliency-based attention models are used to determine regions-of-interest in a video frame which attract the attention of a human viewer. A region-weighted rate-distortion model allocates more bits to the regions-of-interest and less bits to regions which are less salient [LGW+04]. This can improve the subjective visual quality of video streams compared to traditional lossy video coding schemes under the same bandwidth.

2.5.2 Novelty detection

Novelty detection aims at the identification of data which is unknown to a machine learning system based on a prior model which the system has learned during a training phase. Popular approaches for novelty detection are based on the inference of statistical models from training data samples [MS03]. The probability of new incoming data samples can be computed using the learned models. If the probability lies below a given threshold, the data sample is classified as a novel data sample. Another wide-spread novelty measure is the

Mahalanobis distance [Mah36], which describes the distance of a new data sample from the mean of the probability distribution divided by the standard deviation of the distribution. The parametric or non-parametric probability models in statistical novelty detection methods are usually infered by learning techniques as described in Section 2.1.

In contrast to the attention model in [IKN98], which explains the phenomenon of visual salience by neural mechanisms, the method in [SH03] computes saliency maps using a statistical approach for novelty detection. Sajda and Han describe the processing of orientation cues in the primary visual cortex (V1) using filters which receive input from the retina and which are organized in an array of pinwheels whose wedges describe the prefered orientation of the filters. A pinwheel's response can be represented in an N-dimensional space where N is the number of wedges of the pinwheel. The pinwheels' responses are considered as random variables and their distribution is modeled by a mixture of multivariate Gaussian distributions. The novelty of a pinwheel's response is measured by its negative log-likelihood given the mixture of Gaussians model. This measure quantifies the salience of the scene structure.

2.5.3 Computational approaches for surprise detection

Saliency-based attention models allow for the detection of locations within an image which pop out from their neighborhood and therefore capture the attention of a human observer. However, spatial saliency is not the only cue which guides a human's gaze. Temporal events like the sudden appearance of an object in the scene or unexpected motion can be surprising and also attract a human's interest in a bottom-up manner.

Itti and Baldi present in [IB09] a formal way of defining and measuring surprise in a temporal image sequence from low-level visual features like intensity, color, orientation, motion and flicker. As in earlier attention models [IK01], center-surround differences of these features are computed on multiple scales of an image. It is assumed that each feature evokes a series of spikes emitted by a neuron in the primary visual cortex at a certain firing rate. Dynamic visual stimuli lead to varying firing rates of the neurons. Hence, the number of spikes k in a time window is modeled by a Poisson distribution [IB05]

$$p(k) = \frac{\lambda^k}{k!} \cdot \exp\left\{-\lambda\right\}, \tag{2.37}$$

where λ is the expected number of spikes and determines the shape of the Poisson distribution. The parameter λ is learned from observations of the neural firing rates over time by Bayesian inference, as described in Section 2.1. As a conjugate prior the gamma distribution is chosen. The hyperparameters α' and β' of the posterior distribution after observing a firing rate k_0 are

$$\alpha' = \alpha + k_0 \tag{2.38}$$

$$\beta' = \beta + 1, \tag{2.39}$$

where α and β are the hyperparameters of the gamma prior, as given in (2.9). To prevent them from growing towards infinity, both α and β in (2.38) and (2.39) are multiplied with a forgetting factor, which is smaller than 1. The Kullback-Leibler divergence [Kul59] between posterior and prior

$$\mathcal{KL} = \int_\lambda p\left(\lambda \mid \alpha', \beta'\right) \ln \frac{p\left(\lambda \mid \alpha', \beta'\right)}{p\left(\lambda \mid \alpha, \beta\right)} \, \mathrm{d}\lambda \tag{2.40}$$

is used to measure how strongly a visual stimulus changes the prior distribution over the expected neural firing rates and to quantify surprise. As Itti and Baldi show in [IB09], Bayesian surprise is superior to other information-theoretic novelty measures and saliency detectors with respect to the prediction of locations in images which attract human attention. However, the surprise model relies on the simulation of neurons in the primary visual cortex which is difficult to realize on the graphics hardware in a mobile robot for the computation of surprise maps in real-time. Furthermore, Itti and Baldi do not consider correspondences between visual features at different image locations in consecutive frames resulting from camera motion.

In [RD09] an approach is presented for the automatic detection of landmarks for the acquisition of a topological map of the environment. A Bayesian surprise metric is used to determine the novelty of a landmark in the environment. The representation of the environment can be either appearance-based or geometry-based. The appearance-based component is based on the bag-of-words paradigm [SRE+05], which models each camera image as a set of visual words which are formed by quantized SIFT descriptors. Here SIFT stands for the Scale-Invariant Feature Transform, as proposed by Lowe in [Low04]. The number of occurences of each visual word in an image is stored in a histogram associated with the image. Several images captured from a site in the environment show SIFT histograms which are similar but not identical. This type of noise is modeled by a multinomial distribution over the histograms. The parameters of the multinomial distribution are infered in a Bayesian approach. If there are SIFT descriptors in an image which change the prior over the parameters of the multinomial distribution, a novel landmark is detected. While the environment representation in [RD09] is very compact and scales well with large environments, the surprise detector shows an increased number of false positives in cluttered environments.

The method in [HWP10] provides a concept for the detection of interesting parts in video frames. A Latent Dirichlet Allocation model [BNJ03] is used to describe a set of video events like human actions and object motion. Surprising events like a U-turn of a car in a traffic scenario or a pedestrian crossing a street intersection in a diagonal way are identified using a Bayesian framework similar to [IB09] and [RD09]. Hence, the approach is not only limited to the detection of novel objects and landmarks but also considers the detection of unexpected actions.

2.5.4 The role of surprise in learning and visual search

In [SSH95], an approach is presented for reinforcement-driven information acquisition during the exploration of an unknown environment by a robot. This method evaluates the information gain which is achieved between two subsequent states along the robot's way through the environment and uses this metric for assessing the reward of a given exploration policy in a reinforcement learning framework. Similar to the approach by Itti and Baldi [IB09], which is revisited in Section 2.5.3, the Kullback-Leibler divergence is used for the computation of the information gain.

Reinforcement learning is in general a promising means for the autonomous mental development of intrinsically motivated systems [Sch05, SLBS10]. Intrinsic motivation, which results from the pursuit of maximum internal reward, can be driven by learning progress [OKH07] or by novelty and surprise [HW02, Sch10] and leads to the development of complex action sequences. However, the downside of reinforcement learning is that it is not able

to cope with high-dimensional state and action spaces. Hence, the environment has to be abstracted from the robot's sensor data and thus its internal representation is often not as realistic as the appearance representation presented in Chapter 4 of this thesis. Experiments are often performed in an artificial gridworld.

Apart from the body of work on surprise detection published by computer scientists, surprise is also investigated in neurosciences, especially in the context of associative learning. In [dOFD+09] it is shown that there are areas in the primary visual cortex and putamen which respond progressively more to unpredicted and progressively less to predicted visual stimuli. A similar behavior has been found in the prefrontal cortex, which shows high activation if the prediction of associations between perceived stimuli fails [FAS+01].

Besides, the effect of novelty and surprise in visual search is investigated in psychology. In experiments a sequence of search displays, which contain a fixed or varying number of objects, are presented to human subjects. The search display used by Theeuwes [The92] consists of objects with a simple shape, e.g. a square or a circle, in a given color, e.g. red or green. The task of the subject is to find a target which is defined by its shape among a set of nontarget objects which have a different shape. Likewise, the target can be defined by its color among nontarget objects with different color. Once the human has found the target he/she gives a response and the time between the onset of the display and the response is measured. This time interval is denoted by reaction time (RT). The repetition of this visual search task for all displays in the sequence results in several trials per subject. In some trials one of the nontarget objects differs from the other nontarget objects (distractor) and is salient due to its color (color-singleton) in case of shape-defined targets ot due to its shape (shape-singleton) in case of color-defined targets. Müller et al. find in [MGZK09] that the RT of human subjects varies with the frequency of distractor trials. In case of a rare distractor the RTs are higher than in case of a a frequent distractor. Similar effects have been found by Neo and Chua in [NC06] and Horstmann in [Hor02]. This provides evidence that novel distractors evoke surprise and capture the attention of the subject during visual search.

2.6 Autonomous acquisition of object representations

Robots which are able to autonomously acquire representations of unknown objects in cluttered environments can flexibly develop skills for the execution of tasks which require the manipulation and recognition of new objects. Learning visual representations of new objects from camera images requires attentional selection for the segmentation of the object from the background.

In [SGW+07] an active vision system focusses on objects which are sequentially presented by a human. The Adaptive Scene Dependent Filter hierarchy is used for the segmentation of the objects. Maps which contain low-level visual features like color and disparity are computed from the input image. At each pixel the features are summarized across the maps in a vector which also contains the horizontal and vertical pixel position. In a vector quantization step a codebook with prototypes for feature vectors is trained. The Euclidean distance of the feature vector at a pixel from the nearest prototype indicates the novelty of the feature combination. A mask is computed by binarizing the novelty map. This mask is combined with a relevance mask which favors pixel positions near the image center and large disparity values. In addition a skin color mask excludes image regions which exhibit the hand and

arm of the human for object segmentation.

In [SLL+07] a humanoid robot walking around a table acquires a representation of an object on it. At a series of viewpoints around the object images are taken using a stereo camera. Both SIFT [Low04] and color descriptors are computed in the images and clustered to visual words. To segment the object of interest from the background a disparity map is computed and only features whose disparity values lie above a given threshold are selected from the image. The 3D position of the features is also used for the estimation of the robot's motion between two captured images when the object is near. This implies that the object of interest has to be the closest object.

Another approach for the autonomous acquisition of object models by a humanoid robot is presented in [WIS+10]. In contrast to [SLL+07] the humanoid robot does not move around a static object but rotates an object in its hand while capturing a series of images. This enables the robot to acquire visual features from the object across a larger range of viewpoints. The rotation of the robot's hand is preprogrammed so that the images are taken on defined locations on a viewpoint sphere. The object is segmented by comparing the camera image of the object to a reference image, which is computed from a previously taken set of images of the background using Eigenvalue decomposition [ORP00]. It is necessary that the robot's camera does not move between the acquisition of the background scene and the image of the new object. To make the segmentation result more robust, disparity information is used and erroneously segmented parts from the distant background are removed. The visual appearance of the object is represented by SIFT features and Color Cooccurrence Histograms [CK99] which are extracted from the images after object segmentation. To keep the memory consumption of the object model moderate, the features are clustered and a sparse set of prototypical views is stored. The approach in [WIS+10] is chosen as a reference approach for the method presented in Chapter 5 of this thesis.

An issue which arises with the methods in [SGW+07] and [WIS+10] is that regions in the image which exhibit part the human body or the robot's arm have to be ignored during object acquisition. Steil et al. [SGW+07] use a skin color model for the human hand, which of course fails if the human's clothes are also visible. Welke et al. [WIS+10] include proprioceptive sensor information to determine the position of the robot's hand and arm in the image. This requires a very accurate camera-to-hand calibration. Another issue which arises with the approach in [WIS+10] is that the robot's hand occludes a large part of the object in certain poses. As a consequence no features can be computed in these regions.

3 View Synthesis from Unstructured Image-based Environment Representations

Image-based scene representations which incorporate explicit geometry information and camera pose data allow for camera motions on arbitrary smooth trajectories during acquisition. Thus, this type of scene representation is appropriate for mobile robots as a component of the internal environment model which preserves the natural appearance of the scene. The views which sample the appearance of the scene in terms of bunches of light rays at densely spaced discrete viewpoints are denoted by *reference views* throughout this thesis. The accurate estimation of the camera pose and the dense and exact reconstruction of the scene geometry are two challenges during the acquisition of the model. Both steps determine the quality of virtual views which are interpolated from reference views.

This chapter describes a system which is used for image-based scene modeling and rendering in Chapters 4, 5 and 6 of this thesis. The modules of the system provide functions for camera pose estimation, depth recovery and view synthesis. Since the depth estimation method is based on the minimization of a global energy function which is defined across the pixel grid of a reference view, it is computationally expensive and done in an offline step. The synthesis of virtual images from reference views is GPU-based and thus enables the robot to rapidly predict the appearance of the environment.

3.1 Camera pose estimation

For the estimation of the camera poses two alternative approaches are employed, depending on the robot platform used for model acquisition. The views in the image sequences captured by the Pioneer 3-DX platform (see Chapter 6) are localized using an image-based approach. In contrast, the laboratory environment of the "Cobot" platform (see Chapters 4 and 5) is covered by an optical tracking system, which is used for camera localization.

3.1.1 Image-based camera pose estimation

The image-based approach for camera localization applied in this thesis is presented in [MSSB09]. As input data the algorithm receives images from a stereo camera and provides the 6D poses (3D orientation + 3D translation) of the left camera, while using the images of the right camera for recovering the 3D position of features extracted in the left view. The pixel correspondences between the features in the left and right image are determined with subpixel accuracy. Since the algorithm is designed to work at frame rates of $25\,\mathrm{Hz}$ and higher and the 3D reconstruction of the features is time-consuming, it is only done for new, previously unseen features. To extract the features in the left images, the Kanade-Lucas-Tomasi (KLT) tracker [LK81, ST94] is used. For a fast feature tracking across several frames, the intensity gradients are only computed in patches around the feature locations and not in the

whole image as it is done in the original implementation.

The pose estimation method is based on the vision-based GPS (VGPS) approach in [BH04], which associates a reference coordinate system with the first captured image. The 6D poses of the images taken afterwards are computed with respect to this coordinate system. The initial reconstruction of the geometric structure of the features provides a set of N points ${}^0\mathbf{P}_i, i \in \{1, \ldots, N\}$. At the instant when the t-th image is captured, another point set ${}^t\hat{\mathbf{P}}_i, i \in \{1, \ldots, N\}, t \geq 1$ is predicted in [MSSB09], using the pose estimate for the $(t-1)$-th image. In an iterative procedure both the pose for the t-th view and the 3D positions of the points in ${}^t\hat{\mathbf{P}}_i$ are refined.

For the computation of the rotation between the two point sets, the positions of the centroids are subtracted from all points in each set, providing the point sets ${}^0\mathbf{P}'_i$ and ${}^t\hat{\mathbf{P}}'_i$. The inertia matrix subject to these two point sets is given by

$$ {}^t\mathbf{A} = \sum_{i=1}^{N} {}^t\hat{\mathbf{P}}'_i \, {}^0\mathbf{P}'^{\mathrm{T}}_i \tag{3.1} $$

In [AHB87], a solution for the rotation matrix is found from the singular value decomposition (SVD) of ${}^t\mathbf{A} = {}^t\mathbf{U} \, {}^t\mathbf{\Sigma} \, {}^t\mathbf{V}^{\mathrm{T}}$. The rotation matrix is then computed by

$$ {}^t\mathbf{R} = {}^t\mathbf{V} \, {}^t\mathbf{U}^{\mathrm{T}}. \tag{3.2} $$

The translation between the two point sets is given by

$$ {}^t\mathbf{T} = \frac{1}{N} \sum_{i=1}^{N} {}^t\hat{\mathbf{P}}_i - {}^t\mathbf{R} \cdot \frac{1}{N} \sum_{i=1}^{N} {}^0\mathbf{P}_i. \tag{3.3} $$

The rotation and translation of the two point sets is directly related to the rotation and translation of the camera in a static environment.

An important aspect of the algorithm in [MSSB09], which makes the estimation of the camera pose more robust, is a feature weighting strategy which attenuates the influence of features which are detected as outliers when fitting the two point sets. Each feature is weighted by a factor ${}^tw_i$ which decreases with the geometric error between the points in ${}^t\hat{\mathbf{P}}_i$ and the corresponding points in ${}^0\mathbf{P}_i$ rotated by ${}^t\mathbf{R}$. The weight equals 0 above a given threshold for the error. The modified inertia matrix from (3.1) then results in

$$ {}^t\mathbf{A}^{\mathrm{R}} = \sum_{i=1}^{N} {}^tw_i \, {}^t\hat{\mathbf{P}}'_i \, {}^0\mathbf{P}'^{\mathrm{T}}_i. \tag{3.4} $$

One of the strong points of the method in [MSSB09] is that the camera pose estimates are very accurate, which is achieved, i.a., by special mechanisms which ensure that the features are extracted in a wide area across the whole image. Furthermore, no artificial visual markers or any external reference systems are required. However, there are also several weaknesses. One issue is the localization of a camera image which is captured after all features and their 3D positions have been lost. A loss of features might happen in swift rotations of the robot or after the robot has been switched off. A solution to this problem is given in [MMBS09] where the pose of a new camera view is computed from SURF correspondences between the

new image and a camera image taken in the past and kept in memory. Here SURF stands for Speeded-Up Robust Features [BETG08]. However, the image from the memory must have been captured at a viewpoint close to where the new image is taken since SURF features are not invariant against viewpoint variations. Another issue results from dynamic scene parts like moving objects, humans etc. If these objects cover only a small region in the images, their features can be detected as outliers and receive a low weight during pose estimation. However, if they cover a large part of the image, the accuracy of the pose estimates drops. These issues can be tackled by using an active optical tracking system which provides 3D measurements with respect to a fixed coordinate frame which are not affected by the image content.

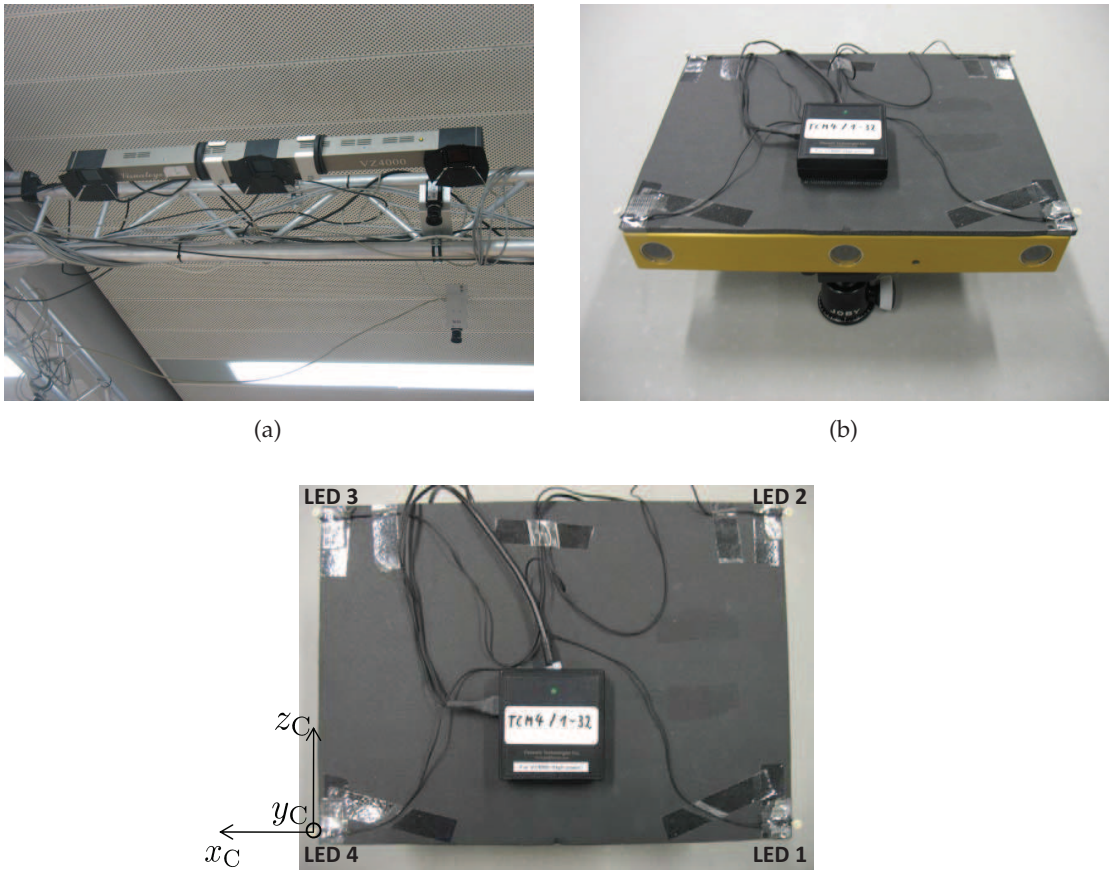
3.1.2 Camera pose estimation using active optical tracking systems

Figure 3.1(a) shows one of multiple Visualeyez$^{\mathrm{TM}}$ VZ 4000 trackers [pti] which are mounted on the ceiling of an indoor environment and capture the positions of LED markers attached to objects moving on the floor. These active-optical real-time 3D trackers are usually employed for human motion tracking. Due to their high accuracy in the range of $1\ \mathrm{mm}$, the systems can also be used for acquiring camera pose data during the acquisition of image-based environment representations. To this end, four LED markers are placed at the corners of a rectangular plate on top of the camera head, as depicted in Figure 3.1(b). The plate is strongly attached to the camera system in a way that it is approximately perpendicular to the cameras' image planes. The markers emit infrared light in a predefined pattern and are controlled via radiocommunication [1].

Figure 3.1(c) shows a top view on the plate and the local coordinate system which is defined for the camera head. As illustrated, the coordinate system has its origin in LED 4. The x_{C}-axis is parallel to the line which aligns the LEDs 2 and 3. The z_{C}-axis lies in the plane defined by the LEDs 1, 2 and 3 and is orthogonal to the x_{C}-axis. The y_{C}-axis is orthogonal to both the x_{C}-axis and the z_{C}-axis.

For the estimation of the 3D orientation of the camera head in the world coordinate system during robot motion the LEDs 1, 2 and 3 are used. The origin of the world coordinate system in the environment and the orientation of its axes have been defined before during the calibration of the tracking system. The computation of the camera head rotation requires the coordinates of the LEDs in the local coordinate system $x_{\mathrm{C}}y_{\mathrm{C}}z_{\mathrm{C}}$. To this end, the 3D positions of the LEDs are first measured in the world coordinate frame using the tracker, while keeping the camera head static over a time interval. In order to remove noise the measurements are averaged over time and the position of LED 4 (the origin of the local coordinate system) is subtracted from the positions of the other LEDs. Then the 3D positions of the LEDs 1, 2 and 3 are rotated so that the LEDs 1 and 2 lie on the x_{C}-axis and LED 3 in the $x_{\mathrm{C}}z_{\mathrm{C}}$-plane. In general, LED 3 will not exactly lie on the z_{C}-axis since, in practice, the straight line aligning the LEDs 1 and 2 is not exactly perpendicular to the straight line aligning the LEDs 2 and 3. The 3D point set with the positions of the LEDs 1, 2 and 3 in the local camera coordinate system is denoted by ${}^{0}\mathbf{P}_i,\ i \in \{1, 2, 3\}$.

[1] The Visualeyez$^{\mathrm{TM}}$ VZ 4000 tracker in Figure 3.1(a) and the Point Grey Bumblebee® XB3 camera system in Figure 3.1(b) are part of the CoTeSys Central Robotics Laboratory (CCRL), which is supported within the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see also `www.cotesys.org`.



(a) (b)

(c)

Figure 3.1: *(a) One of the trackers which capture the position of LED markers on the camera head from the ceiling. (b) LED markers in the corners of a rectangular plate on top of the camera head. (c) The local coordinate system of the camera head.*

When the robot moves and acquires an image sequence, the world coordinates of the LEDs 1, 2 and 3 captured by the tracker at the time instant when the t-th image is taken can be summarized in another 3D point set ${}^t\hat{\mathbf{P}}_i,\ i \in \{1,2,3\}$. To get the matrix ${}^t\mathbf{R}$ which describes the rotation between the local coordinate frame $x_C y_C z_C$ and the world coordinate frame $x_W y_W z_W$, the two point sets are fitted using the SVD of the inertia matrix computed from the two point sets, as described in Section 3.1.1 [AHB87, MSSB09]. The translation ${}^t\mathbf{T}$ of the camera head with respect to the origin of the world coordinate system is provided by the tracked 3D position of LED 4. The 6D pose of the camera head (3D orientation + 3D translation) is given by the 4×4-matrix

$$ {}^t\mathbf{M} = \begin{bmatrix} {}^t\mathbf{R} & {}^t\mathbf{T} \\ \mathbf{0}^T & 1 \end{bmatrix}. \tag{3.5} $$

The matrix is stored for each reference view in the image-based environment representation.

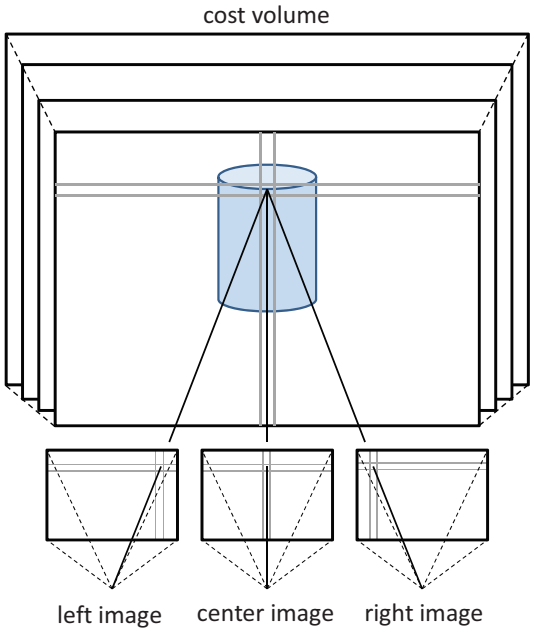


Figure 3.2: *Multiple depth hypotheses are tested in a plane sweep to find pixel correspondences for depth reconstruction.*

3.2 View-dependent geometric modeling

As already discussed in Section 2.2.3, view-dependent geometric scene models are often prefered over one global geometric model of the environment since they account for viewpoint-dependent depth perception of a point on a reflective surface. The image-based rendering approach in this thesis uses view-dependent geometry information in terms of dense depth maps. The depth maps are computed for each reference view using either multiple cameras in an image acquisition device (see Figure 3.1(b)) or other reference views stored in the environment representation. When a dynamic scene is captured, it is preferable to estimate the depth of the scene from images taken by a multi-camera device with synchronized cameras since then the images are acquired at the same time instant.

Figure 3.2 illustrates the search for pixel correspondences between the center image and the left or the right image. This is done by a plane-sweep technique similar to [Col96], which tests several depth hypotheses by computing the intensity differences between a pixel of the center image and pixels of the left and right image which correspond to it according to the depth hypothesis. Comparing the two intensity differences from the left and the right image, the smaller one is stored for each pixel of the center image and for each depth hypothesis, which results in a cost volume. The intrinsic and extrinsic camera parameters, which are required for the correspondence search, are determined beforehand by the calibration of the multi-camera system using [SSF+]. If the three views in Figure 3.2 are reference views from the environment representation, the extrinsic camera parameters are provided by their pose data.

The cost volume is the input data to a global energy minimization technique based on belief propagation as described in [FH06]. The pixel grid of the center image is treated as a Markov

Random Field and the most probable depth hypothesis is infered for each pixel, taking into account smoothness constraints between adjacent pixels.

3.3 View interpolation

Using the pose and depth information, new virtual images can be interpolated from the reference views at intermediate viewpoints. When the view synthesis module is initialized, a generic mesh is created and transfered to a vertex buffer in the memory of the graphics hardware. This mesh is used later for the representation of a view-dependent geometric model. The X_C, Y_C and Z_C coordinates of the mesh vertices are reconstructed in the camera coordinate system for each pixel with the image coordinates (u, v) by

$$\begin{pmatrix} X_\mathrm{C} \\ Y_\mathrm{C} \\ Z_\mathrm{C} \end{pmatrix} = \mathbf{K}^{-1} \cdot \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \quad \text{with} \quad u \in \{1, \ldots, w\} \text{ and } v \in \{1, \ldots, h\} \tag{3.6}$$

where w and h denote the image width and height, respectively. The matrix

$$\mathbf{K} = \begin{pmatrix} f_x & s & o_x \\ 0 & f_y & o_y \\ 0 & 0 & 1 \end{pmatrix} \tag{3.7}$$

contains the horizontal and vertical focal lengths (f_x, f_y) of the real camera used for the acquisition of the reference images as well as the horizontal and vertical pixel coordinates of its image center (o_x, o_y) and the pixel skew s. As (3.6) shows, the Z_C coordinate of all vertices equals 1, which means that the vertex array lies in a plane which is parallel to the image plane. In the index buffer of the graphics hardware, indices to mesh vertices are stored in triples defining a triangle structure. The triangle mesh is illustrated in Figure 3.3. Apart from the 3D positions the pixel locations (u, v) are stored as texture coordinates $(\frac{u}{w}, \frac{v}{h})$ with the vertices, which allows for mapping the reference image on the mesh. During the initialization of the view synthesis module seven instances of this generic mesh are created, i.e. one for each reference view selected for view interpolation.

3.3.1 View selection

Before a virtual image is rendered from the image-based environment representation, a subset of reference images is selected and loaded into the texture memory of the graphics hardware together with the associated depth and pose data. This view selection is necessary because the view-dependent geometric models are only locally valid. The same holds for non-Lambertian effects exhibited in the reference images. Furthermore, it would not make sense to use reference views covering a part of the environment which lies completely outside the field of view of the virtual camera.

To this end, all reference views in the representation are ranked in terms of a cost function. As illustrated in Figure 3.4, rays are defined within the viewing frustum of the virtual camera which align its optical center and one of the seven points $\mathbf{P}_i,\ i = 1 \ldots 7$, in its image plane

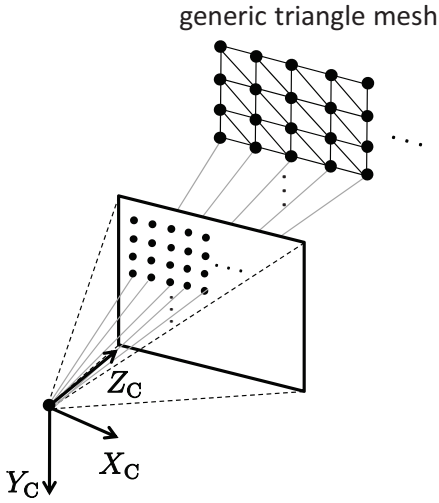


Figure 3.3: *Generation of a generic mesh for the representation of a view-dependent geometric model on the graphics hardware.*

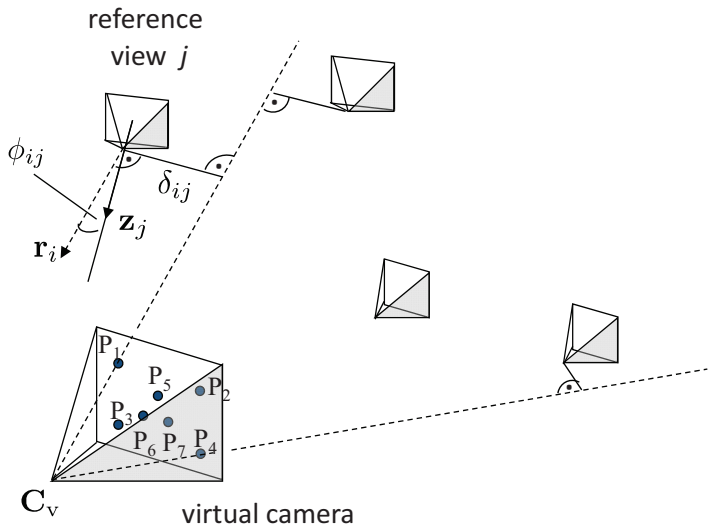


Figure 3.4: *For seven representative rays in the viewing frustum of the virtual camera the reference view with the smallest cost value is selected, respectively. The cost of a reference view is determined by its distance to the ray and its viewing direction.*

reconstructed in 3D space, respectively. In theory, the best reference view should be selected for all rays in the viewing frustum of the virtual camera. Since this is very tedious, the seven rays are used as representatives. As shown in Figure 3.4, three of them intersect the image plane of the virtual camera near the center and the other four near the corners. Thus, if reference views are selected which are close to the rays and whose viewing direction is similar, the whole image plane is covered.

The directions of the rays are described by the vectors $\mathbf{r}_i = \mathbf{C}_\mathrm{v} - \mathbf{P}_i$. $\mathbf{C}_\mathrm{v}$ is the optical center

of the virtual camera in world coordinates. The costs γ_{ij} of all reference views with respect to the seven rays depend on the distance measure δ_{ij} and the angle ϕ_{ij}.

$$\gamma_{ij} = \frac{\delta_{ij}}{\cos\phi_{ij}} = \frac{\|\mathbf{r}_i \times \mathbf{z}_j\|}{\frac{\mathbf{r}_i \cdot \mathbf{z}_j}{\|\mathbf{r}_i\|}}, \quad \mathbf{r}_i \cdot \mathbf{z}_j \neq 0. \tag{3.8}$$

In a nutshell, δ_{ij} is the orthogonal distance of the projection center of a reference camera to one of the rays. ϕ_{ij} is the angle between the viewing direction of a reference camera and one of the rays. $\mathbf{z}_j$ has unit length and denotes the direction of the optical axis of the j-th reference camera in world coordinates, which is equivalent to its viewing direction. $\mathbf{z}_j$ is retrieved from the pose data stored with the reference view. For each ray the reference view with minimum cost is selected. When the virtual camera is translated or rotated with the robot's motion, new reference views might be selected for view synthesis – the ones which are most suitable for the new viewpoint.

The view selection is similar to the strategy in [ESK03]. In constrast to the method in this section, Evers-Senne and Koch compute in [ESK03] the cost for selecting a reference view by the weighted sum of the distance δ_{ij} and the angle ϕ_{ij}. An open question is how to choose the weights. The weighting is not necessary if, as described here, the distance measure and the angular deviation are combined multiplicatively.

3.3.2 Texture blending

Rendering a virtual image is done in two passes. In the first pass a geometric model is reconstructed from the depth maps of the selected reference views, respectively. To this end, the 3D position of each vertex in the generic mesh in Figure 3.3 is scaled by the depth value d read from the corresponding pixel of the depth map and transformed to world coordinates by

$$\begin{pmatrix} X_\mathrm{W} \\ Y_\mathrm{W} \\ Z_\mathrm{W} \\ 1 \end{pmatrix} = {}^{t_j}\mathbf{M} \cdot \begin{pmatrix} d \cdot X_\mathrm{C} \\ d \cdot Y_\mathrm{C} \\ d \cdot Z_\mathrm{C} \\ 1 \end{pmatrix}. \tag{3.9}$$

${}^{t_j}\mathbf{M}$ denotes the pose matrix of the j-th selected reference view according to (3.5). The reconstruction of the vertices in world coordinates is done in parallel, and thus very fast, by a vertex shader program executed by the GPU. Using the texture coordinates which associate the pixels of the reference images with the mesh vertices, the reference images are mapped as textures on the geometry models. The projection of each textured geometry model on the image plane of the virtual camera results in seven warped reference views which are stored as textures in the GPU memory.

In the second pass, the virtual image is rendered by averaging the seven RGB color vectors from the warped reference images at each pixel.

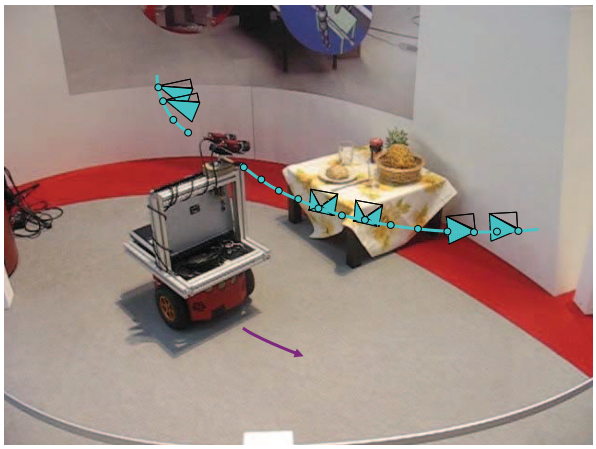


Figure 3.5: *Acquisition of an image sequence using a Pioneer 3-DX.*

3.4 Results

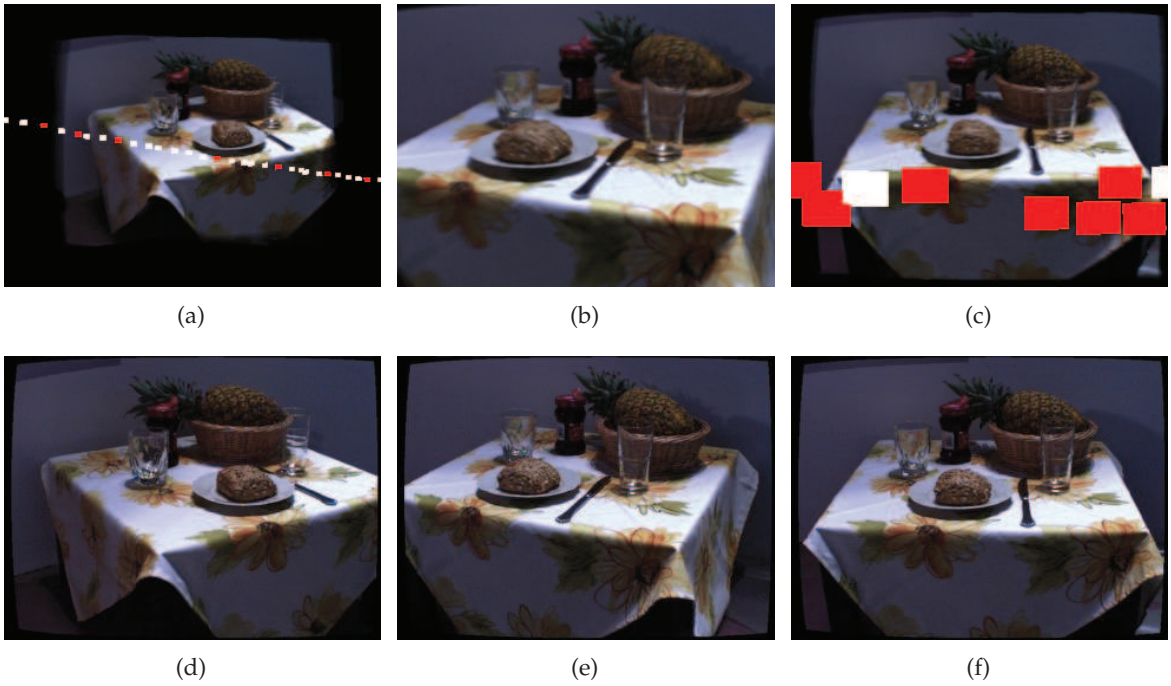
In experiments a stereo image sequence is acquired using the Pioneer 3-DX platform [mob] shown in Figure 3.5 which is equipped with a stereo camera [2]. The robot is controlled to move along a circular trajectory around a scene which contains household objects with complex appearance like glasses and knives. On a trajectory length of $1.8\ \mathrm{m}$, the robot captures 213 stereo images with a resolution of 640×480 pixels.

The left images of the stereo sequence are stored as reference images in the image-based scene representation. As described in Section 3.1.1, the poses of the left camera views are estimated with respect to the coordinate frame of the first acquired camera view, while the right views are used for the 3D reconstruction of the KLT features. Before the calculation of a depth map for each left view, the images are subsampled by a factor of 2 to reduce the computing time and the amount of memory required to store the cost volume.

The Figures 3.6(a), 3.6(b) and 3.6(c) show virtual images rendered at three different viewpoints. Figure 3.6(a) shows a distant virtual view of the scene and illustrates the reference image sequence in terms of red and white squares. The red squares indicate, as in Figure 3.6(c), the seven reference images chosen by the view selection method in Section 3.3.1 for the current pose of the virtual camera. A comparison between the virtual images and the real camera images in Figures 3.6(d), 3.6(e) and 3.6(f) shows that the view interpolation approach in Section 3.3 largely preserves the photorealism of the original images. The rendered images are nearly free of artifacts. Especially the virtual close-up view in 3.6(b) exhibits a realistic refraction of light by the left glass. In traditional geometry-based approaches for view synthesis the realistic rendering of such complex optical effects would require exhaustive raytracing techniques.

Apart from the quality of the virtual images, the execution time of the algorithm is analyzed. To this end, a sequence of 2620 virtual images is rendered on a Laptop with a Quad Core processor which works at a clock rate of $1.73\ \mathrm{MHz}$. The shader programs are executed by the GPU of a NVIDIA GeForce GT 435M. The motion of the virtual camera is controlled

[2] The Adept MobileRobots Pioneer 3-DX robot and the stereo camera head in Figure 3.5 have been provided by the German Aerospace Center (DLR).



(a) (b) (c)

(d) (e) (f)

Figure 3.6: *(a)-(c) Virtual images rendered at three different viewpoints. (d)-(f) One of the seven reference images used for the interpolation of the virtual images in (a)-(c).*

by keyboard and mouse commands in an interactive manner. Figure 3.7 shows the times for computing the virtual images of the sequence, which include the time for view selection, for loading depth maps and textures from the hard disk to GPU memory and the time of the two rendering passes. Over the whole sequence, an average computing time of $24\ \mathrm{ms}$ is measured for a frame. During the first part of the image sequence, until frame 1275, the virtual camera performs a predominantly translational motion near the sequence of the reference views, starting at the position of the first captured image. The increase of the computing time to $50\ \mathrm{ms}$ or higher is explained by the data transfer from the hard disk to the GPU when new reference views are selected. In the middle of the sequence, the virtual camera stops and performs a rotation around its vertical axis. Since here the image content changes very fast, new reference views have to be loaded permanently, which results in higher computing times sometimes reaching the interval between $250\ \mathrm{ms}$ and $300\ \mathrm{ms}$. When the virtual camera moves back again to its starting position, from frame 1760 on, the computing time drops again and mostly lies below $50\ \mathrm{ms}$.

3.5 Summary

This chapter describes an approach for the realistic visualization of real-world environments. Experimental results show that virtual images of an environment with complex appearance are nearly free of artifacts and can be rendered multiple times a second. The modules presented here are elementary for the interpolation of the probabilistic appearance prior in Chapter 4 and the synthesis of virtual images for the computation of illumination-invariant

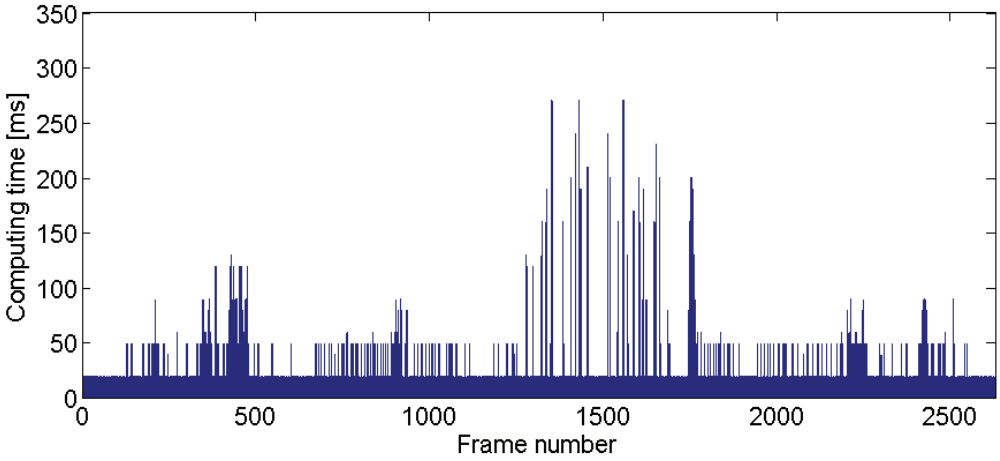


Figure 3.7: *The computing time for rendering the images of a sequence of virtual views.*

image-based environment representations in Chapter 6. As in other work [ESK03, ZKU+04], view-dependent geometric models are used together with the camera poses of the reference views to render novel images from pixel correspondences between the reference views. The approach in [ZKU+04] identifies and renders object boundaries in a separate pass and in [ESK03] neighboring depth maps are fused to get a consistent local geometric model for rendering. While these steps reduce rendering artifacts, they are also time-consuming and are omitted in the algorithm described in this chapter. Apart from the depth estimation module in Section 3.2 all modules of the system process data on average in less than $50\ \mathrm{ms}$. If a depth camera working with invisible structured light [SS03], as e.g. the Kinect [kin], is used for depth estimation, both the acquisition of the image-based environment representation and the interpolation of virtual images can be performed on the fly by a mobile robot.

4 A Probabilistic Appearance Representation and its Application to Surprise Detection

The ability to recall and predict the appearance of the environment from a percept history enables a cognitive robot to assess its current observation and to extract regions that convey novelty and thus are particularly interesting for task selection and task execution. Since there is strong evidence from literature that attention is driven not only top-down but also bottom-up from stimuli data (see Section 2.5), a representation which contains information about the luminance and chrominance of the environment facilitates rapid attentional selection, as tedious preprocessing of the currently observed image is not necessary. Hence, the cognitive robot can already filter relevant information from early stimuli before higher cognitive layers are reached. An image which is taken at a given time instant and stored in the robot's internal environment representation, as proposed in Chapter 3, only reflects the momentary appearance of the scene but does not tell how long the environment has been in the perceived state. The color value of an image pixel, e.g., does not reveal that the brown table in the middle of the kitchen is at its common position but that the spilled liquid on the floor is unusual. Hence, in order to assess the uncertainty of the currently perceived state of the environment, the robot has to evaluate a series of images taken over a time interval. The robot can then build an internal representation of the environment which reflects its belief in the hypothesis that the scene appears in a certain color [MS10].

To this end, a probabilistic appearance-based environment model is presented in this chapter, which represents the appearance of the scene at a series of densely-spaced viewpoints. The dense sampling of light ray intensities and colors is inspired by the image-based rendering approaches revisited in Section 2.2. However, in contrast to image-based representations, the pixels of a view do not store the luminance and chrominance captured by the robot's camera but the luminance and chrominance are treated as Gaussian-distributed random variables. The parameters of the Gaussian distributions are infered from the images the robot acquires near the viewpoints over time, using a Bayesian inference technique (see Section 2.1.2). Prior distributions whose hyperparameters are stored at the pixels of a view in the representation are adapted by the captured luminance and chrominance values, which results in a continuous update of the robot's internal environment model. The surprise measure in this chapter quantifies how much the prior at a pixel changes with new luminance and chrominance values, which is measured by the Kullback-Leibler divergence.

As the experiments in this chapter show, the proposed method for surprise detection reliably indicates image regions which exhibit novel changes in the robot's environment. The approach presented in this chapter is superior to less sophisticated methods like image differencing. Furthermore, it is investigated how robust the surprise measure is against a reduction of the number of reference views in the probabilistic appearance representation and a wider spacing between them.

4.1 A probabilistic appearance representation for cognitive robots

In the representation which is proposed in this thesis the luminance and chrominance values captured at a single pixel for a given viewpoint are modeled by Gaussian distributions

$$p\left(X_k \mid \mu_k, \lambda_k\right) = \left(\frac{\lambda_k}{2\pi}\right)^{\frac{1}{2}} \cdot \exp\left\{-\tfrac{1}{2}\lambda_k \left(X_k - \mu_k\right)^2\right\} \quad \text{with } k \in \left\{Y, C_\mathrm{b}, C_\mathrm{r}\right\}. \tag{4.1}$$

Three separate probability models for the luminance Y and the two chrominance components C_b and C_r are used. There is strong evidence that in the human visual system the luminance and chrominance information is similarly processed in decoupled pathways [EZW97]. The luminance Y and the chrominance components C_b and C_r are computed from the RGB values captured by the robot's camera using the irreversible color transform [TM02].

The parameter μ_k of the Gaussian distribution denotes the expected luminance or chrominance value and the parameter λ_k is the precision of the distribution, i.e., the reciprocal value of the variance. Hence, the larger the precision, the smaller is the uncertainty and the stronger is the belief that the environment appears in the expected luminance and chrominance. These parameters are updated with each new observation that the robot makes in the vicinity of the viewpoint.

Like in image-based representations with explicit geometry (see Section 2.2.3 and Chapter 3) a per-pixel depth map and the 6-Degree of Freedom (6DoF) pose of the robot's camera head with respect to a defined world coordinate system are stored at each viewpoint. Figure 4.1 illustrates the proposed probabilistic appearance representation.

4.1.1 Bayesian inference of model parameters

As discussed in Section 2.1, there are several ways to infer the parameters of the Gaussian distribution from the acquired luminance and chrominance samples. If a Maximum-Likelihood approach is applied, the mean can be estimated by the average of sequentially captured samples and the precision can be computed from the squared differences of the samples from the infered mean. This results in point estimates for the parameters describing one single Gaussian model, which is supposed to be the only valid model. In regions of the environment that the robot does not visit frequently, however, only a small set of luminance and chrominance samples is acquired around a given viewpoint. As pointed out in Section 2.1, the decision in favor of a specific probability model based on an insufficient amount of sample data is very unreliable. The Bayesian approach, in turn, takes into account that, apart from the most likely model, there might be other models that can be valid for the observed data and models this uncertainty using probability distributions over the model parameters.

One reason why a Gaussian model is used for the luminance and chrominance values at a pixel is that the Gaussian distribution belongs to the exponential family. Thus, there exists a joint conjugate prior for its parameters [Bis06]. This makes any further analysis like the comparison of posterior and prior distribution straight forward (see Section 4.2). The conjugate prior which is used for the Bayesian inference of both the mean and the precision of the

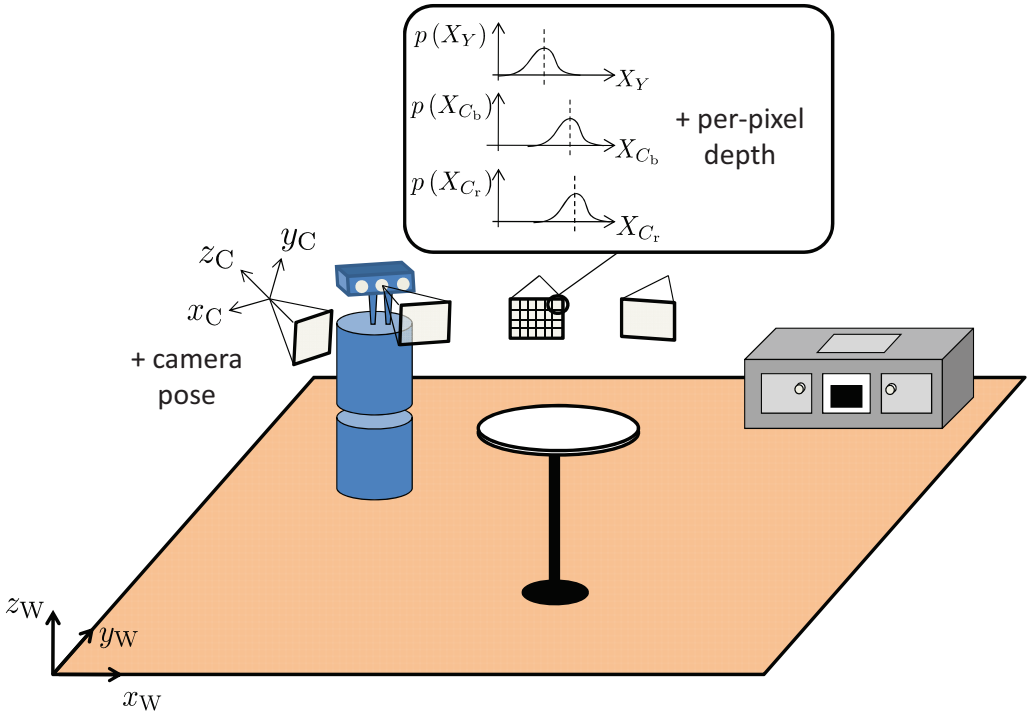


Figure 4.1: *The proposed appearance representation uses Gaussian models for the luminance and the chrominance of the environment at each pixel at a viewpoint. The Gaussian distributions are infered from observations along the robot's trajectory. The representation also includes a depth map and the pose of the robot's camera head for each viewpoint.*

Gaussian distribution is the normal-gamma distribution. It has the form

$$p_0(\mu_k, \lambda_k) = \frac{\beta_{0,k}^{\alpha_{0,k}}}{\Gamma(\alpha_{0,k})\sqrt{2\pi\sigma_{0,k}}} \cdot \lambda_k^{\alpha_{0,k}-\frac{1}{2}} \cdot \exp\left\{-\beta_{0,k}\lambda_k\right\} \cdot \exp\left\{-\frac{\lambda_k(\mu_k-\tau_{0,k})^2}{2\sigma_{0,k}}\right\} \tag{4.2}$$

with $k \in \{Y, C_b, C_r\}$ again. $\Gamma(\cdot)$ is the gamma function in (2.10). Since the form of the normal-gamma model is fully determined by its four hyperparameters $\alpha_{0,k}$, $\beta_{0,k}$, $\tau_{0,k}$ and $\sigma_{0,k}$, it is sufficient to store these four hyperparameters for a given pixel at a viewpoint. An example for a normal-gamma distribution in the luminance channel is shown in Figure 4.2.

When the robot makes a new observation $\mathbf{X}_{ob} = \{X_{ob,k}\}_{k=Y,C_b,C_r}$ at a viewpoint, the prior distribution in (4.2) is turned into a posterior distribution using Bayes' formula

$$p(\mu_k, \lambda_k \mid X_{ob,k}) \propto p(X_{ob,k} \mid \mu_k, \lambda_k) \cdot p_0(\mu_k, \lambda_k). \tag{4.3}$$

The posterior distribution in (4.3) is again a normal-gamma distribution with the hyperpa-

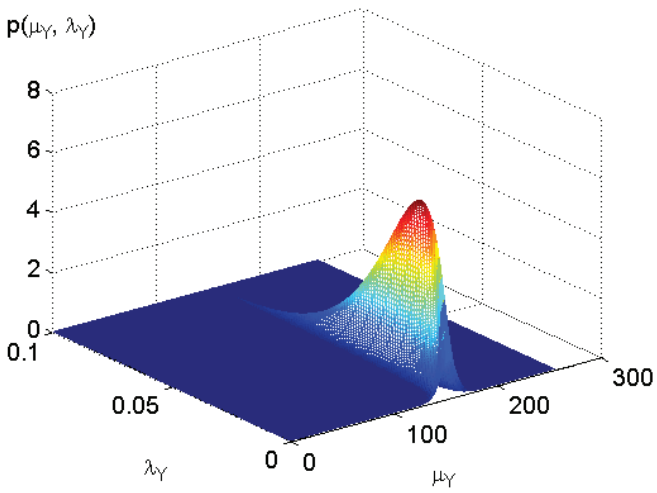


Figure 4.2: *An example of a normal-gamma distribution over the mean μ_Y and the precision λ_Y of the Gaussian model of the luminance channel. The expected luminance at this pixel is around 150.*

rameters

$$\alpha_k = \alpha_{0,k} + \frac{1}{2} \tag{4.4}$$

$$\beta_k = \beta_{0,k} + \frac{1}{2} \cdot \frac{(X_{\mathrm{ob},k} - \tau_{0,k})^2}{\sigma_{0,k} + 1} \tag{4.5}$$

$$\tau_k = \frac{\sigma_{0,k} \cdot X_{\mathrm{ob},k} + \tau_{0,k}}{\sigma_{0,k} + 1} \tag{4.6}$$

$$\sigma_k = \frac{\sigma_{0,k}}{\sigma_{0,k} + 1}. \tag{4.7}$$

These update equations correspond to the equations (2.22) - (2.25) for the special case that only one new sample is used for the parameter update ($N = 1$). Considering the update equation in (4.4) it shows that the parameter α_k can be interpreted as an indicator of how many samples have been acquired so far for inference. The parameter τ_k in (4.6) encodes the expected value of the luminance or chrominance at a pixel. Hence, the three two-dimensional arrays of the τ_k-values together are supposed to be close to a photorealistic virtual image in YCbCr color space. In (4.5) the parameter β_k contains a sum of squared differences between the new observed luminance or chrominance value and the corresponding expected value. The parameter σ_k in (4.7) determines the weight that a new luminance or chrominance value receives during the update of the expected value (see (4.6)). Considering the update equations in (4.4) - (4.7), two issues arise. First, the two parameters α_k and β_k increase towards infinity as the number of updates grows to infinity. Second, the parameter σ_k tends towards zero. This leads to the problem that after a couple of iterations the expectation τ_k does not change any more. However, if a new object with different color appears in the environment after a few observations, the expected appearance has to be adapted to the new scene. Furthermore, if the robot discovers during exploration new parts of the environment which have been occluded so far by objects, the appearance of the corresponding

region in the new observations has to be transfered to the internal representation, which is only achieved with values for $\sigma_{0,k}$ which are significantly larger than 0.

To prevent the unbounded growth of α_k and β_k, a forgetting factor $f < 1$ is introduced in [IB05] where similar update equations have been derived for a gamma model as a prior over the parameter of the Poisson-distributed neural firing rates. Furthermore, in order to ensure rapid adaptation to unexpected changes, the parameter $\sigma_{0,k}$ is modified before the Bayesian update as

$$\sigma_{p,k} = \min\left(\max\left(\sigma_{0,k} \cdot \exp\left\{\frac{\xi}{M_{01}}\right\}, \sigma_{\min}\right), \sigma_{\max}\right) \tag{4.8}$$

where ξ is a constant. M_{01} denotes the first-order moment of the prior normal-gamma distribution with respect to the precision

$$\begin{aligned} M_{01} &= \int_{\lambda_k=-\infty}^{+\infty} \int_{\mu_k=-\infty}^{+\infty} \lambda_k \cdot p_0\left(\mu_k, \lambda_k\right) \, \mathrm{d}\mu_k \, \mathrm{d}\lambda_k \\ &= \frac{\alpha_{0,k}}{\beta_{0,k}} \, . \end{aligned} \tag{4.9}$$

The reciprocal value of M_{01} is the expected uncertainty about the appearance of the environment at a given pixel. It controls $\sigma_{p,k}$ and so the update of the expectation $\tau_{0,k}$. If a change occurs in the environment, new observations deviate from the prior expectation and the robot gets unsure about the true appearance. In this case, the internal representation has to be updated rapidly. Furthermore, (4.8) prevents the parameter $\sigma_{p,k}$ from growing beyond an upper threshold $\sigma_{\max}$ and from falling below a lower threshold $\sigma_{\min}$ near 0. The update equations (4.4) - (4.7) then become

$$\alpha_k = f \cdot \alpha_{0,k} + \frac{1}{2} \tag{4.10}$$

$$\beta_k = f \cdot \beta_{0,k} + \frac{1}{2} \cdot \frac{\left(X_{\mathrm{ob},k} - \tau_{0,k}\right)^2}{\sigma_{p,k} + 1} \tag{4.11}$$

$$\tau_k = \frac{\sigma_{p,k} \cdot X_{\mathrm{ob},k} + \tau_{0,k}}{\sigma_{p,k} + 1} \tag{4.12}$$

$$\sigma_k = \frac{\sigma_{p,k}}{\sigma_{p,k} + 1} \tag{4.13}$$

Eq. (4.9) shows that the multiplication of both $\alpha_{0,k}$ and $\beta_{0,k}$ with the forgetting factor f does not change the first-order moment of the prior distribution with respect to λ_k. In this thesis, a forgetting factor of 0.8 is used.

After the Bayesian update, the prior distribution at the robot's current viewpoint is replaced by the new posterior distribution, which is stored in the representation and serves as a new prior for future observations.

4.1.2 View interpolation

When a robot returns to a part of the environment that it has visited before it will never exactly go along the same trajectory and never make its observations at the same viewpoints as before. Hence, to retrieve the prior distributions for the pixels of a currently captured image the parameters of the probability model in (4.2) have to be interpolated from nearby views

from the internal representation. Section 4.1.1 points out that the robot stores the infered posterior distributions at a viewpoint in terms of two-dimensional arrays for each luminance and chrominance channel. Thus, the robot can later retrieve them from memory like images in order to interpolate the prior for the current viewpoint. The array of parameters which are stored at a viewpoint in the representation will be denoted by *reference parameter images* and the array of parameters that is interpolated at the current position will be denoted by *virtual parameter image*.

As in Chapter 3, the poses of the reference views and view-dependent geometric models of the scene are used for the interpolation of virtual parameter images. To determine the pose of the camera head, multiple active-optical real-time 3D trackers, as shown in Figure 3.1(a), are employed. The pose with respect to a fixed world coordinate frame $x_{\mathrm{W}} y_{\mathrm{W}} z_{\mathrm{W}}$ (see Figure 4.1) is estimated using the method in Section 3.1.2. View-dependent range information is recovered by calculating dense depth maps as explained in Section 3.2. The depth maps are estimated from three images taken by the camera unit simultaneously.

Each time before a new virtual parameter image is rendered a subset of all reference views is selected according to the method in Section 3.3.1. Finally, the virtual parameter image is computed by reconstructing the local geometry of the scene and averaging the warped reference parameters at each pixel. The interpolated values at a pixel of the virtual parameter image correspond to the parameter values $\alpha_{0,k}$, $\beta_{0,k}$, $\tau_{0,k}$ and $\sigma_{0,k}$ of the normal-gamma prior in (4.2).

4.2 Computation of surprise maps

The probabilistic appearance representation in Section 4.1 provides a framework for the detection of surprising events and for attentional selection. The posterior distribution obtained by (4.10) - (4.13) expresses the robot's belief in a hypothesis about the appearance of the environment after a new observation. If this new observation drastically changes the belief the robot had before this observation, the robot gets surprised.

A formal way of describing Bayesian surprise in terms of how much a new observation changes the robot's prior belief is provided by the Kullback-Leibler divergence [SSH95, IB09]. The Kullback-Leibler divergence of two normal-gamma distributions

$$\mathcal{KL}\left(p\left(\mu_k, \lambda_k\right); p_0\left(\mu_k, \lambda_k\right)\right) = \iint p\left(\mu_k, \lambda_k\right) \ln \frac{p\left(\mu_k, \lambda_k\right)}{p_0\left(\mu_k, \lambda_k\right)} \, \mathrm{d}\mu_k \mathrm{d}\lambda_k \tag{4.14}$$

can be written in a closed form [MS10]. This is very convenient for technical implementations and rapid computation of per-pixel surprise maps on graphics hardware. In the luminance and chrominance channels of the new observation, the surprise values $S_k,\ k \in \left\{Y, C_{\mathrm{b}}, C_{\mathrm{r}}\right\}$ at a given pixel are then computed by

$$S_k = \mathcal{KL}\left(p\left(\mu_k, \lambda_k\right); p_0\left(\mu_k, \lambda_k\right)\right) = T_1 + T_2 + T_3 + T_4 + T_5, \tag{4.15}$$

where

$$T_1 = \ln\left(\frac{\beta_k^{\alpha_k}\cdot\Gamma(\alpha_{k,0})\cdot\sqrt{\sigma_{k,0}}}{\beta_{k,0}^{\alpha_{k,0}}\cdot\Gamma(\alpha_k)\cdot\sqrt{\sigma_k}}\right) \tag{4.16}$$

$$T_2 = (\beta_{k,0}-\beta_k)\cdot\frac{\alpha_k}{\beta_k} \tag{4.17}$$

$$T_3 = (\alpha_k-\alpha_{k,0})\left[\psi(\alpha_k)-\ln(\beta_k)\right] \tag{4.18}$$

$$T_4 = \frac{1}{2\sigma_{k,0}}\left[\left(\tau_{k,0}^2-2\tau_{k,0}\tau_k+\tau_k^2\right)\cdot\frac{\alpha_k}{\beta_k}+\sigma_k\right] \tag{4.19}$$

$$T_5 = -\frac{1}{2} \tag{4.20}$$

In (4.16) and (4.18), $\ln(\,\cdot\,)$ denotes the natural logarithm and $\psi(\,\cdot\,)$ in (4.18) is the digamma function, as given in (2.15). The surprise values which are computed in the luminance and chrominance channels are finally combined to a total surprise score by

$$S = S_Y + S_{C_\mathrm{b}} + S_{C_\mathrm{r}}\,. \tag{4.21}$$

Image regions which exhibit large surprise values convey much novelty over the internal representation and can be used in order to guide the robot's attention.

For a fast generation of surprise maps, the Kullback-Leibler divergence in (4.15) is computed for each pixel in parallel by a pixel shader program executed by a GPU. Since shader languages like the OpenGL Shading Language (GLSL) or C for graphics (Cg) do not provide functions to compute gamma and digamma functions directly, approximations are used [MMBS09]. The gamma function in (4.16) is approximated by Stirling's Series

$$\Gamma(z) \approx \sqrt{\frac{2\pi}{z}}\cdot\left(\frac{z}{e}\right)^z\cdot\exp\left(\frac{1}{12z}-\frac{1}{360z^3}+\frac{1}{1260z^5}\right) \tag{4.22}$$

where $e = 2.71828\ldots$ is the Euler's number. Likewise, an approximate value for the digamma function in (4.18) is given by

$$\psi(z) \approx -\frac{1}{z}-\gamma+\sum_{n=1}^{5}\left(\frac{1}{n}-\frac{1}{z+n}\right) \tag{4.23}$$

where $\gamma = 0.57721\ldots$ denotes the Euler's constant.

4.3 Experimental results

To evaluate the computation of surprise based on the proposed probabilistic appearance representation a long image sequence $\mathcal{I}_1$ with 1283 frames is captured, using the mobile platform "Cobot" shown in Figure 4.3(a) [RRK+10] [1]. The robot is equipped with the Point Grey Bumblebee® XB3 camera system [pgb] depicted in Figure 3.1(b). After an image is taken, it

[1] The robot platform in Figure 4.3(a) is part of the CoTeSys Central Robotics Laboratory (CCRL), which is supported within the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see also `www.cotesys.org`.

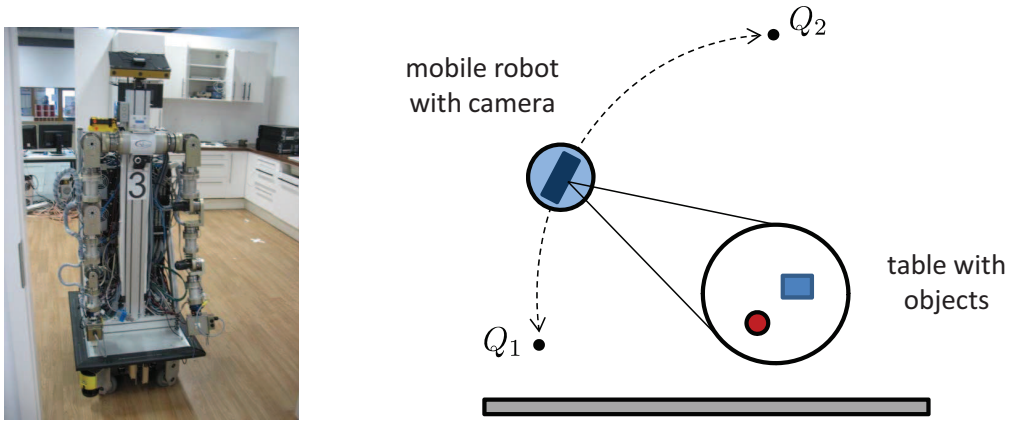


Figure 4.3: *(a) The mobile platform "Cobot" used for image acquisition. (b) During the acquisition of the image sequence $\mathcal{I}_1$ the robot moves multiple times from point Q_1 to point Q_2 along an approximately circular trajectory. The trajectories between the two points are similar but never identical.*

is immediately downsampled to a resolution of 320×240 pixels. The images are captured at a frame rate of 7 fps. While the estimation of the camera head's poses is performed during image acquisition in real-time at 25 Hz, the computation of the depth maps is done off-line, as described in Section 3.2. To this end, the images are saved on a harddisk. The interpolation of the prior parameters, the inference of the posterior distributions according to (4.10) - (4.13) and the computation of the surprise maps in (4.15) are performed by the GPU of a NVIDIA GeForce GTX 275. For these steps an average execution time of 120 ms is measured.

The robot starts at point Q_1 in Figure 4.3(b) and is controlled to go along a circular trajectory with its camera head looking towards the center of the circle. When it reaches Q_2 it stops and immediately goes back on the circle to Q_1. Arrived at Q_1 again, it repeats the motion multiple times. At the turning points the robot continues the data acquisition so that the image sequence $\mathcal{I}_1$ is not interrupted. Each time it reaches Q_1 the scene is changed by a human who adds and removes objects. Thus, the acquisition of the image sequence $\mathcal{I}_1$ can be divided into several phases (A to D) which are separated by events (X_1 to X_3). The phases and events are described in Table 4.1.

The environment representation consists of reference parameter images infered at 200 dense viewpoints around the scene. During the first run of the robot from Q_1 to Q_2 an initial set of 200 reference parameter images is stored. The parameters of the normal-gamma prior distribution for the inference of the first reference parameter image are chosen as $\alpha_{0,k} = 1$, $\beta_{0,k} = 1$, $\tau_{0,k} = 0$ and $\sigma_{0,k} = 5$ with $k \in \{Y, C_\mathrm{b}, C_\mathrm{r}\}$. All other reference parameter images in the representation are infered during the first run by using the robot's observation and a prior which is interpolated from reference parameter images at nearby viewpoints. In all other runs along the trajectory the latest reference parameter image selected for view interpolation (see Section 3.3) is replaced by the parameter image of the posterior distributions at the robot's current viewpoint. The depth map and pose matrix associated with the latest selected reference parameter image are replaced by the depth map and the pose at the robot's current viewpoint as well.

Table 4.1: *The acquisition of the image sequence $\mathcal{I}_1$ can be divided into several phases. In each phase the robot moves from Q_1 to Q_2 and back.*

Phases and Events	Frames	Description
A	1 - 339	Robot acquires reference model of the scene.
X_1	325 - 358	Human adds glass.
B	340 - 674	Robot captures images of the scene with the new glass.
X_2	657 - 686	Human adds a black cup.
C	675 - 979	Robot captures images of the scene which now also contains the new cup.
X_3	962 - 985	Human removes cup.
D	980 - 1283	Robot captures images of the scene. Glass is the only additional object.

4.3.1 Evaluation of the Bayesian surprise measure based on the probabilistic appearance prior

Figures 4.4(c), 4.5(c) and 4.6(c) illustrate three surprise maps which are computed in the phases B, C and D, respectively. In Figure 4.4(a), the image captured by the robot at frame 465 is shown. Figure 4.4(b) illustrates the values of $\tau_{k,0}$ which are interpolated at the robot's viewpoint from the internal representation. The parameters $\tau_{k,0}$, which are stored in YCbCr domain, are transformed to RGB domain in order to facilitate a better comparison to the captured image. As described in Section 4.1.1, the parameters $\tau_{k,0}$ encode at each pixel the luminance and chrominance of the scene which the robot expects at its viewpoint. Figure 4.4(b) shows that the appearance representation proposed in this thesis enables the robot to predict a virtual image with high realism.

The surprise map in Figure 4.4(c) clearly indicates the glass as a novel object, which has been added to the scene at the beginning of phase B and is not contained in the internal representation. Figure 4.4(d) depicts the sum of the parameters $\beta_0^{\Sigma} = \sum_{k\in\{Y,C_\mathrm{b},C_\mathrm{r}\}} \beta_{k,0}$, which express the uncertainty of the appearance at each pixel (see (4.9)). The figure shows that the values are relatively low over the image, which means that the robot is quite sure about the appearance of the scene. There are slightly elevated values around the edges of the objects. This is because small pose inaccuracies due to tracker noise can lead to small shifts of the object edges in the observation and the predicted image during phase A.

Figure 4.5(a) shows an observation of the robot in phase C and Figure 4.5(b) the appearance of the scene which the robot expects from the internal representation. The glass, which has been a novel object in phase B, is already shown in this virtual image. The surprise map in Figure 4.5(c) shows that the black cup added by the human at the beginning of phase C conveys a lot of novelty. At frame 1010 in phase D the robot makes the observation in

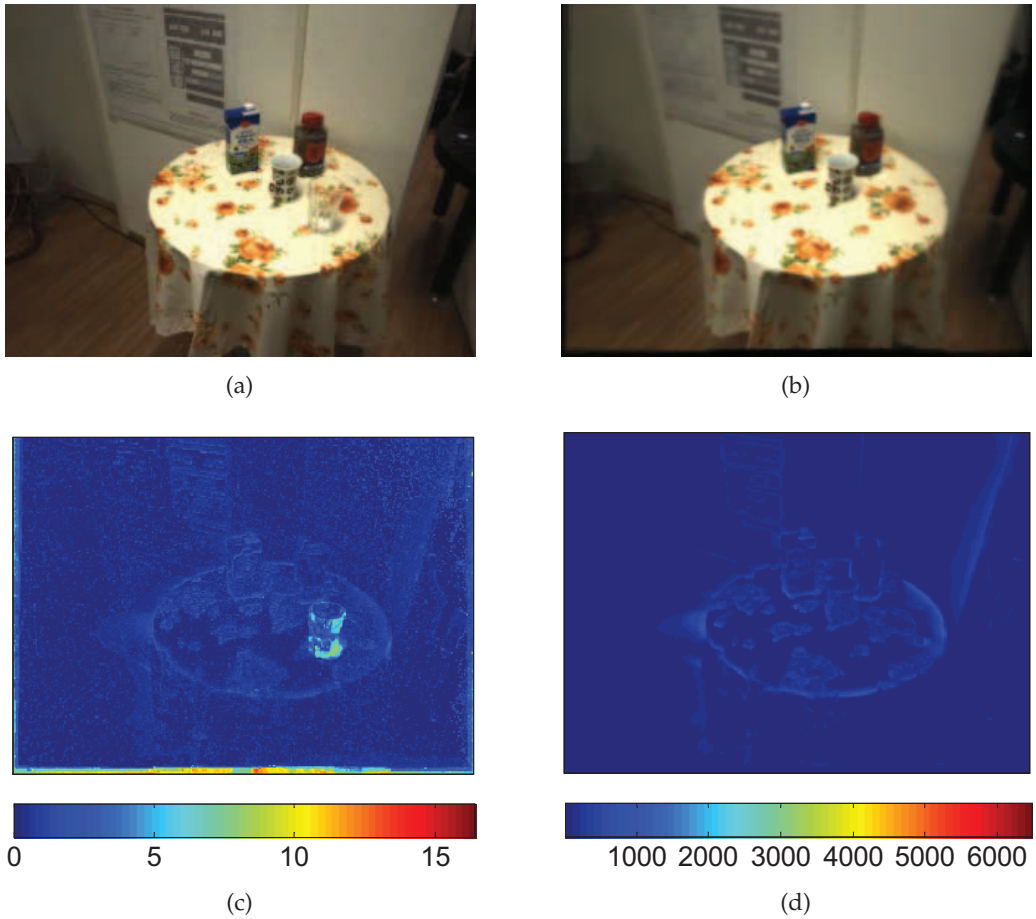


Figure 4.4: *(a) Frame 465 of the image sequence $\mathcal{I}_1$. (b) The parameters $\tau_{0,k}$ correspond to the robot's expected appearance. For illustration the values of $\tau_{0,k}$ are transformed to RGB domain. (c) The surprise map indicates the glass as a novel object. (d) The parameters $\beta^{\Sigma}_{0,k}$ show that the robot's uncertainty about the appearance is low across the image.*

Figure 4.6(a) which shows that the cup has been removed again by the human. Although the cup is still contained in the internal representation, as depicted in Figure 4.6(b), the robot is only little surprised that it has disappeared. Since the sudden appearance of the cup at the beginning of phase B has aroused a large stimulus difference in the luminance channel, the robot is still unsure about the true appearance in that region (large values of β^{Σ}_{0} in Figure 4.6(d)). The robot expects low luminance values but the infered Gaussian model has a small precision. Hence, large luminance values, which are captured from the table cloth in the new observation, are unlikely but still possible. The sum of the parameters β^{Σ}_{0} in Figure 4.6(d) in the region of the glass indicates that the proposed environment representation is able to store the appearance of complex transparent objects with a relatively low uncertainty.

Next, a quantitative evaluation of the surprise maps over the whole image sequence $\mathcal{I}_1$ is presented. To this end, polygons which encompass the regions of the glass and the cup are manually drawn, starting with the frame in which they appear on the table for the first time. These polygons define masks indicating image regions which are supposed to evoke high

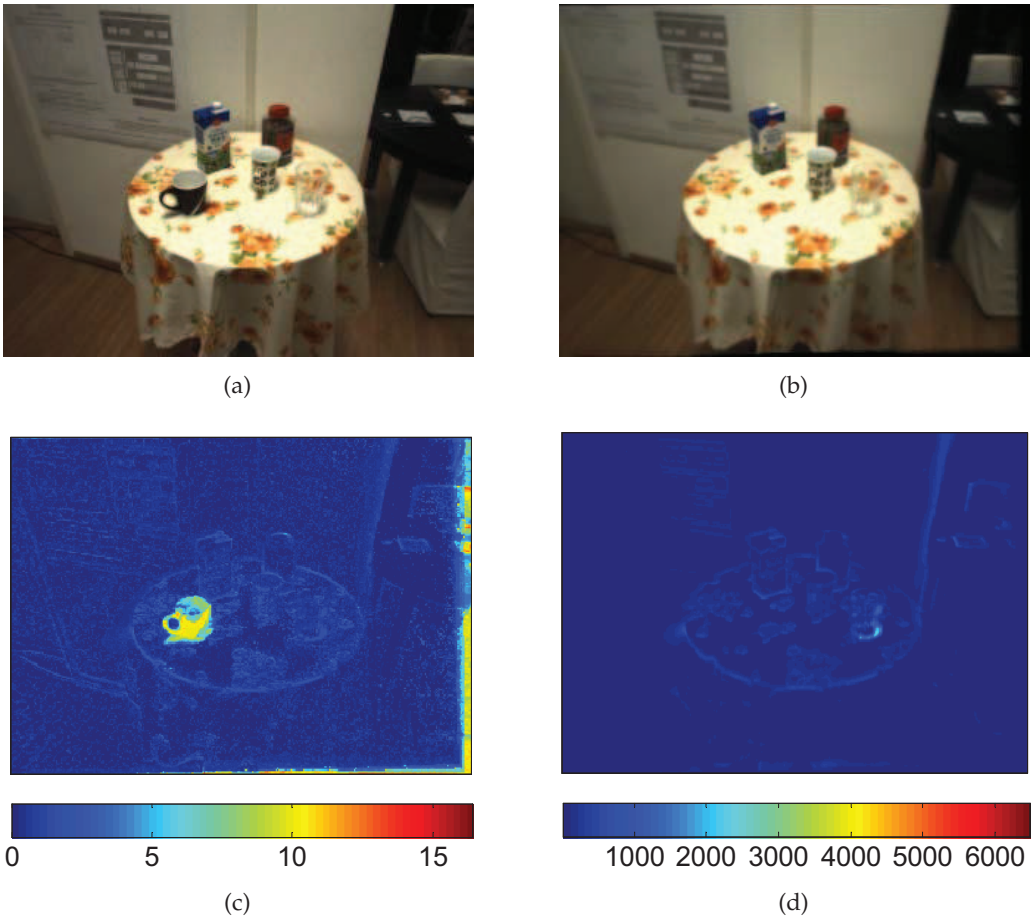


Figure 4.5: *(a) Frame 800 of the image sequence $\mathcal{I}_1$. (b) The parameters $\tau_{0,k}$ represent the robot's expected appearance. For illustration the values of $\tau_{0,k}$ are transformed to RGB domain. (c) The surprise map indicates the cup as a novel object. (d) The parameters $\beta^{\Sigma}_{0,k}$ show that the robot's uncertainty about the appearance is low across the image.*

surprise values. Figure 4.7 shows the masks tinged in blue color. The mask in Figure 4.7(c) indicates the region on the table where the cup has been before it is removed. It is used to measure the robot's surprise about the missing cup. To remove noise from the surprise maps the values are averaged over 4×4-blocks. Figure 4.8(a) shows the maximum values which are measured in a block within the masks that indicate the glass and the cup, respectively. The measurements are started at frame 340, after the glass has been put on the table (green curve). The reason for the drop of the surprise values in the region of the glass at the beginning of phase B is that the robot coming from point Q_2 reaches the turning point Q_1. Since in Q_1 several images are captured at the same viewpoint, the surprise values decrease rapidly. As soon as the robot starts moving again towards Q_2, the surprise values increase since the reference parameter images along the trajectory still represent the scene without the glass. After all reference parameter images have been updated when the robot reaches Q_2 at frame 520, the surprise values decrease since the scene does not contain any novelty along the way back to Q_1.

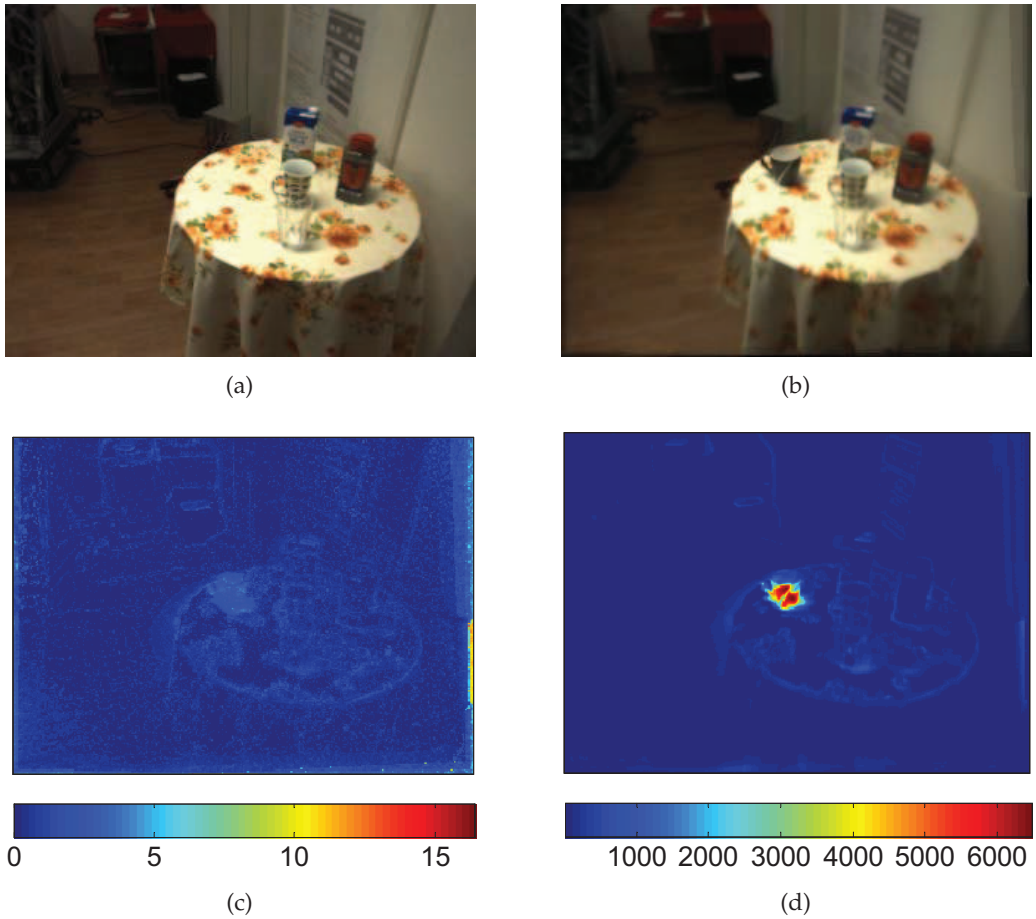
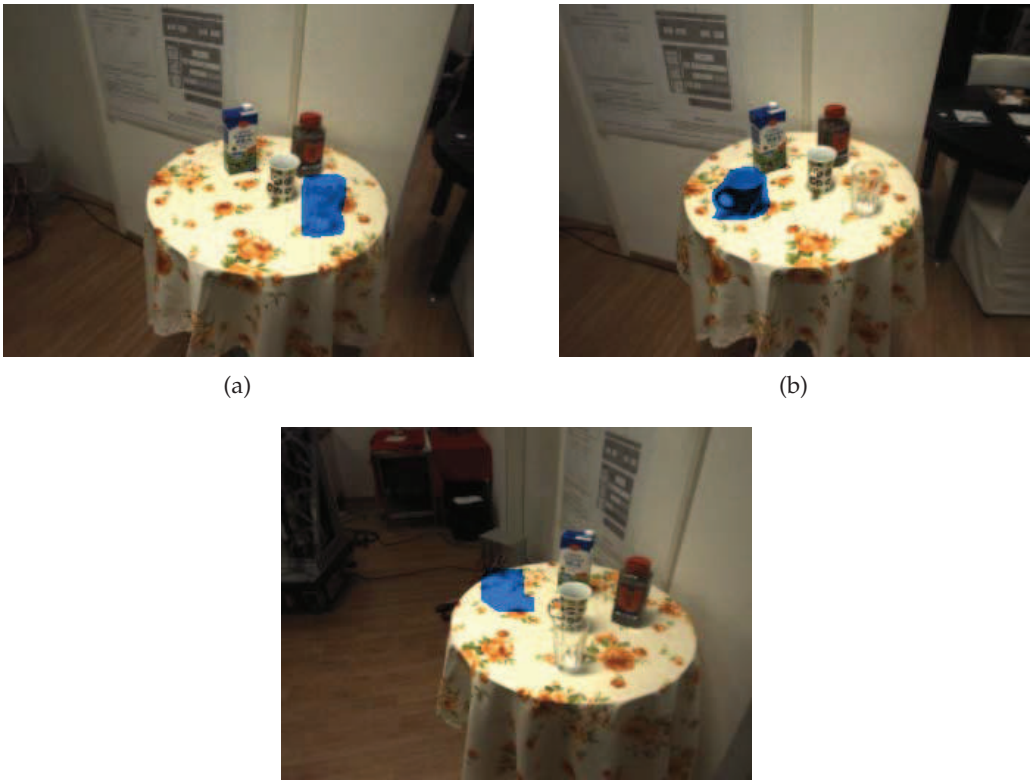


Figure 4.6: *(a) Frame 1010 of the image sequence $\mathcal{I}_1$. (b) The parameters $\tau_{0,k}$ correspond to the robot's expected appearance. For illustration the values of $\tau_{0,k}$ are transformed to RGB domain. (c) The surprise map shows only slightly elevated values in the region of the missing cup. (d) The parameters $\beta^{\Sigma}_{0,k}$ show a region of high uncertainty where the cup has been removed.*

The measurements of the maximum surprise values in a block within the region of the cup are started at frame 675 as soon as it is on the table (orange curve). Figure 4.8(a) shows that the cup evokes larger surprise values than the glass does at the beginning of phase B. This is because the stimulus difference between the bright table cloth and the dark cup is higher. The large surprise values hold on along the trajectory until the robot reaches Q_2. There, as in the case of the glass, the surprise values drop since the cup then is no longer novel for the robot. When the cup is removed again around frame 980 the surprise values increase but do not reach as high values as at the beginning of phase C. Hence, as already noted before, the removal of the cup is not as surprising for the robot as its addition since the robot has already seen the table without the cup before.

A similar behavior can be found for the mean surprise values computed from all 4×4 blocks within the masks that indicate the glass and the cup (see Figure 4.8(b)). The surprise values averaged over the whole region of the glass are of course much lower than the maximum



(a) (b)

(c)

Figure 4.7: *The image regions showing the glass in (a) and the cup in (b) are manually labeled to create a mask for the evaluation of the robot's surprise about these objects. In (c) the region where the cup has been is labeled to measure the robot's surprise about the missing cup.*

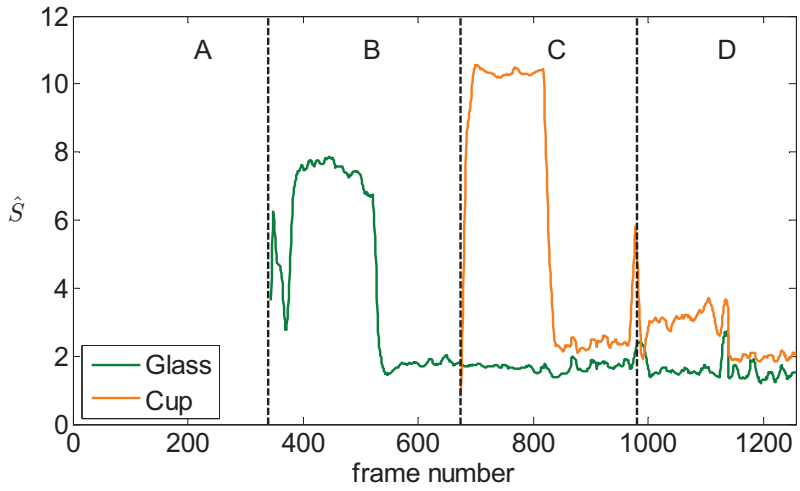
surprise values because the stimulus difference is very low at sites where the glass hardly refracts the light. However, the values during the first run from Q_1 to Q_2 are still higher than in all following runs when the novelty of the glass has gone.

The acquisition of the image sequence $\mathcal{I}_1$, the computation of surprise maps and the update of the internal representation is illustrated in the video on [sur].

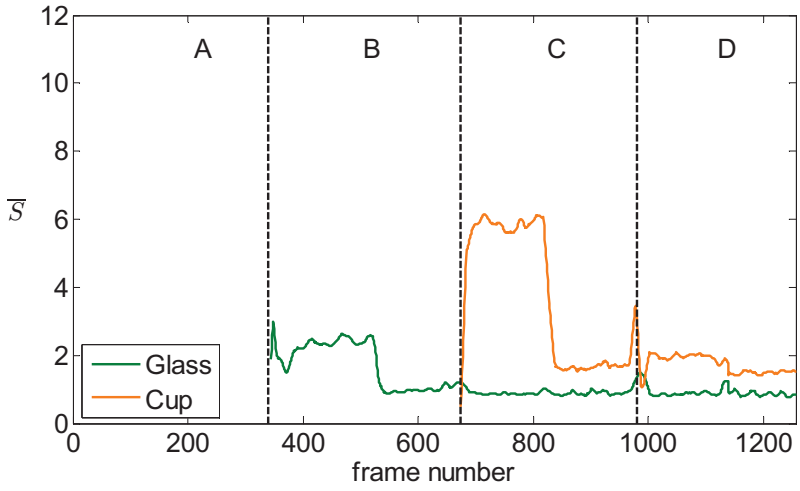
4.3.2 Comparison to change detection using image-based representations

For visual search tasks, the relationship of the surprise values within the region of an object of interest to the surprise values in the rest of the map is important. The surprise values within the region have to be higher than outside so that the attention of the robot is directed to the novel object. To evaluate this, a novel measure named *Attentional Selectivity (AS)* is introduced and calculated as follows [MS10]

$$\mathrm{AS} = 10 \log_{10} \left(\frac{\overline{S}}{\hat{S}_{\mathrm{out}}} \right) \ \mathrm{dB}, \tag{4.24}$$



(a)



(b)

Figure 4.8: *(a) The maximum surprise value of a 4×4-block within the regions of the glass (green) and the cup (orange). (b) The average surprise value over all 4×4-blocks within the regions of the glass (green) and the cup (orange). In both cases there are high values during the robot's first run from Q_1 to Q_2 after the new object has been put on the table.*

where $\overline{S}$ is the surprise value averaged over all 4×4-blocks within the region of the object of interest. $\hat{S}_{\text{out}}$ denotes the maximum value of all blocks outside the region. Blocks near the borders of the surprise map are excluded from the analysis since there the virtual images which predict the appearance of the scene often do not contain any information (cf. right border of Figure 4.5(b)). This automatically leads to high surprise values (cf. right border of Figure 4.5(c)). In practice, the robot can always determine the distance of a surprising block to the borders of the image and give the blocks near the borders a lower priority for attentional selection.

Furthermore, another measure named *Peak Attentional Selectivity (PAS)* is introduced and computed as

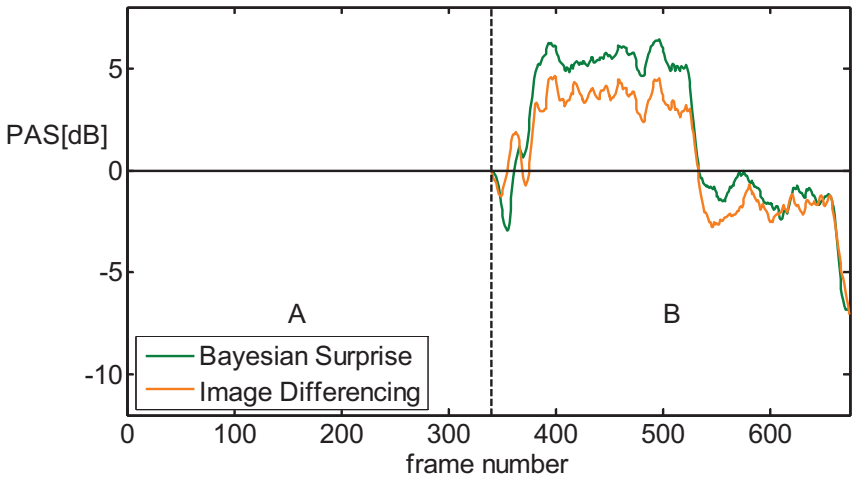
$$\mathrm{PAS} = 10 \log_{10} \left(\frac{\hat{S}}{\hat{S}_{\mathrm{out}}} \right) \ \mathrm{dB}, \tag{4.25}$$

where $\hat{S}$ denotes the maximum surprise value of all blocks within the region of the object of interest.

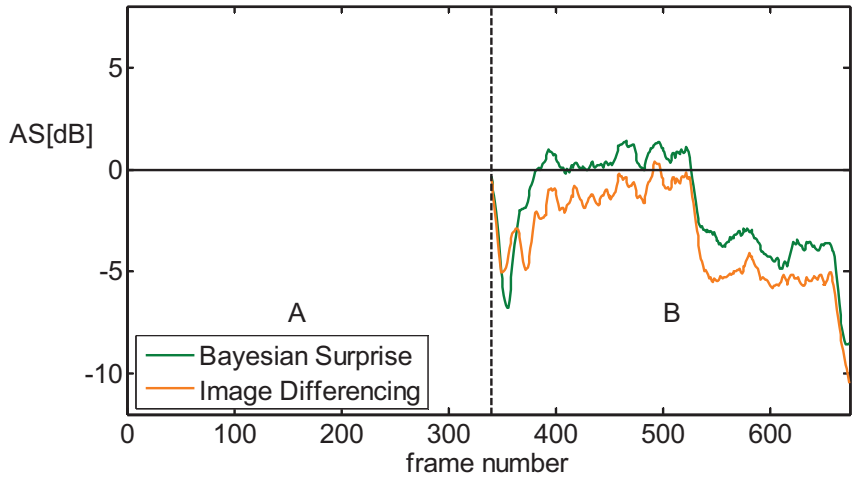
In the following analysis the AS and the PAS are evaluated with respect to the regions of the glass and the cup over the image sequence $\mathcal{I}_1$. For comparison, the AS and PAS obtained by the method *Image Differencing* are also evaluated. Here, the appearance of the environment is stored in terms of a (non-probabilistic) image-based representation and a virtual image is interpolated at the current viewpoint of the robot from nearby reference images, as described in Chapter 3. The image-based representation is updated by replacing the latest selected reference image by the current observation. The absolute differences between the luminance and chrominance values in a new observation and in the predicted image are used in order to detect changes between the two images. The AS and PAS are computed using (4.24) and (4.25), while replacing the surprise values with the corresponding values of the sum of absolute differences over the luminance and chrominance channels. The difference to the appearance representation presented in this chapter is that the robot only stores deterministic snapshots of the environment and does not hold any information about the history and the uncertainty of the appearance.

Figure 4.9(a) shows the PAS values obtained by Bayesian Surprise and Image Differencing for the glass region in phase B. On the robot's way from Q_1 to Q_2 the PAS is above 0 dB and is higher for Bayesian Surprise. When the robot returns to Q_1, the PAS drops below 0 dB in both cases, which means that the region of the glass does not contain a block which is more surprising than the blocks in the rest of the map. Figure 4.9(b) shows that the AS obtained by Image Differencing hardly gets over 0 dB between the frames 340 and 520, whereas the AS obtained by Bayesian Surprise clearly does. That means that according to Image Differencing, the glass is not more novel than other parts in the image, which is not the case during the first run of robot from Q_1 to Q_2. The explanation for this is shown in Figure 4.10. Here, the edges of the objects in the scene exhibit elevated absolute difference values in Figure 4.10(d), which are due to pose inaccuracies. This type of noise decreases the AS. In contrast, the surprise values around the object edges in Figure 4.10(c) are relatively low, since the probabilistic appearance representation establishes small regions of uncertainty around the object borders. The drop of the AS towards the end of phase B is due to the human person which is about to put the cup on the table and enters the image from the left (see Figure 4.11). Both methods detect the human as a novelty and provide surprise/difference values which are higher than the ones within the region of the glass.

Figure 4.12(a) compares the PAS values obtained by Bayesian Surprise to the PAS values obtained by Image Differencing with respect to the region of the cup. Due to the large stimulus difference, which causes large absolute difference values especially in the luminance channel, the PAS obtained by Image Differencing is higher than the PAS obtained by Bayesian Surprise. However, both the PAS and the AS values in Figure 4.12(b) lie above 0 dB so that the cup is clearly detected as a novel object. In phase D, it is shown that both the PAS and the AS values obtained by Image Differencing do not fall below 0 dB, whereas the AS obtained by Bayesian Surprise does. Thus, in case of Bayesian Surprise, the robot would briefly be



(a)



(b)

Figure 4.9: *The region of the glass is evaluated. Both the PAS values in (a) and the AS values in (b) are higher for Bayesian surprise than for Image Differencing. In (b), the AS values below* 0 dB *obtained by Image Differencing during the robot's run from* Q_1 *to* Q_2 *wrongly show that the glass does not convey more novelty than the rest of the scene. In contrast, Bayesian surprise correctly detects the glass as a novel object.*

astonished about the missing cup but on average this image region is not more surprising than others because the robot has seen the table cloth before phase C. In contrast, Image Differencing detects high novelty in the region of the missing cup. This could deteriorate the attentional selection of other image regions which might contain new objects that the robot has not seen before.

Furthermore, the effect of a reduction of the number of reference parameter images in the probabilistic appearance representation on the peak attentional selectivity is investigated.

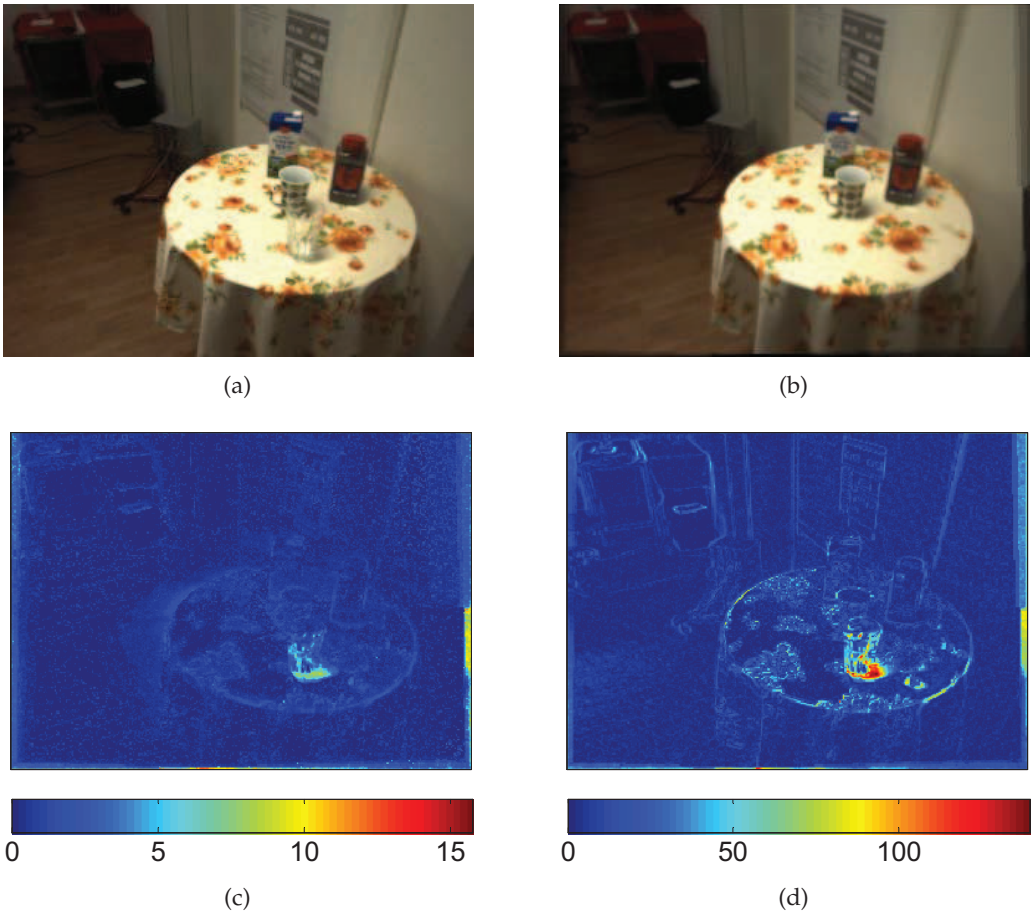
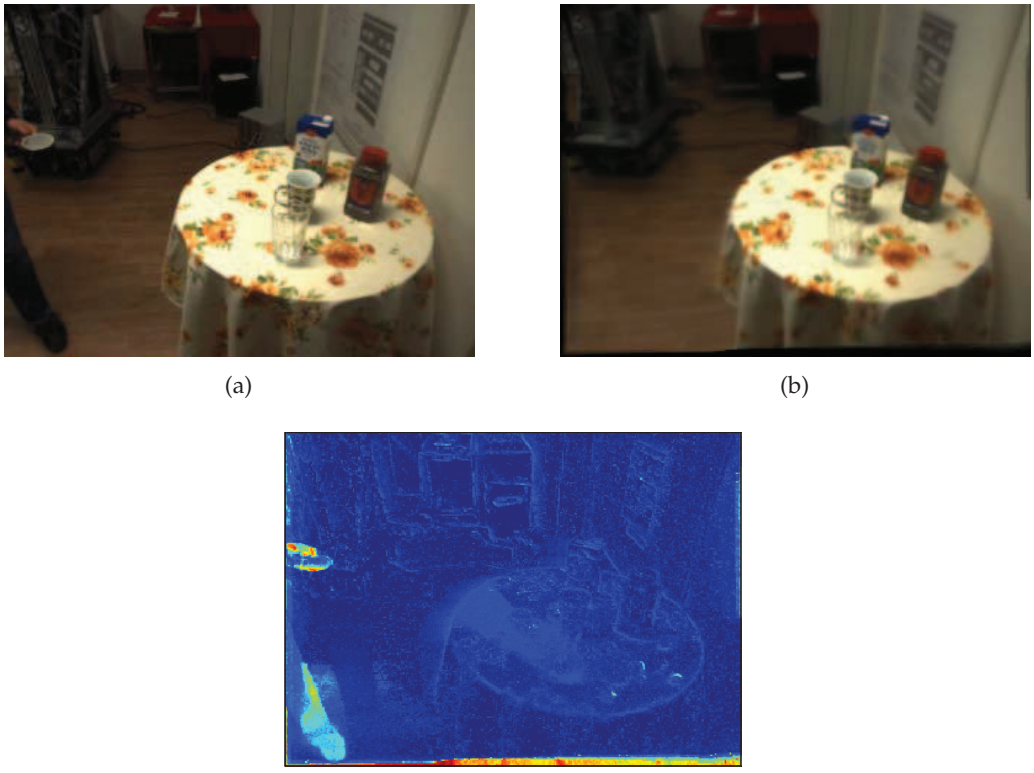


Figure 4.10: *(a) Frame 400 of the image sequence $\mathcal{I}_1$. (b) The virtual image which is interpolated from the robot's image-based representation and hence predicts the appearance of the scene. (c) The Bayesian surprise values are higher in the region of the glass than in the rest of the map. (d) The map obtained by image differencing is sensitive to pose inaccuracies and shows false positives near the edges of the objects.*

Figure 4.13 compares the PAS values in phase B obtained by Bayesian Surprise to the PAS values obtained by Image Differencing for different viewpoint densitites. The curves denoted by "(F)" are obtained using the complete environment representation, which consists of 200 reference parameter images as described before. The addition "(R2)" denotes PAS curves obtained from an environment representation with 100 reference parameter images and with a distance between two neighboring views which is twice as large as in the case "(F)". Finally, the curves with the addition "(R4)" are obtained from an environment representation which contains only 50 reference parameter images whose spacing is four times as large as in the case "(F)". The curves in Figure 4.13 show that the PAS values obtained by Bayesian surprise using an environment model with a number of views reduced by a factor of 2 are still as high or even a little bit higher than the PAS values obtained by Image Differencing using the complete environment model. If the number of views is reduced by a factor of 4, the PAS values drop for both metrics. Hence, in phase B, the surprise metric behaves more robust with respect to a lower number of reference parameter images in the



(a) (b)

(c)

Figure 4.11: *The human who is about to put a cup on the table in frame 662 (a) is not expected by the robot (b) and thus causes high surprise values near the left border (c).*

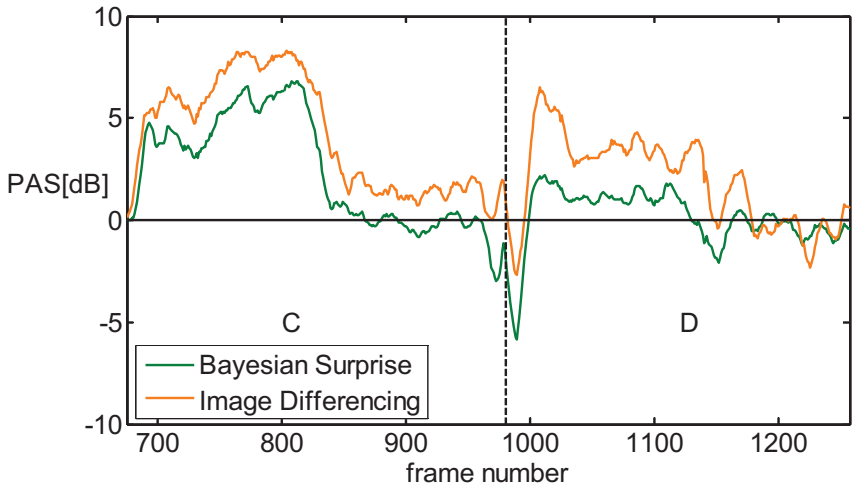
environment representation. In general, however, the optimal number of reference views in the representation always depends on the complexity of the scene [CTCS00].

4.4 Discussion

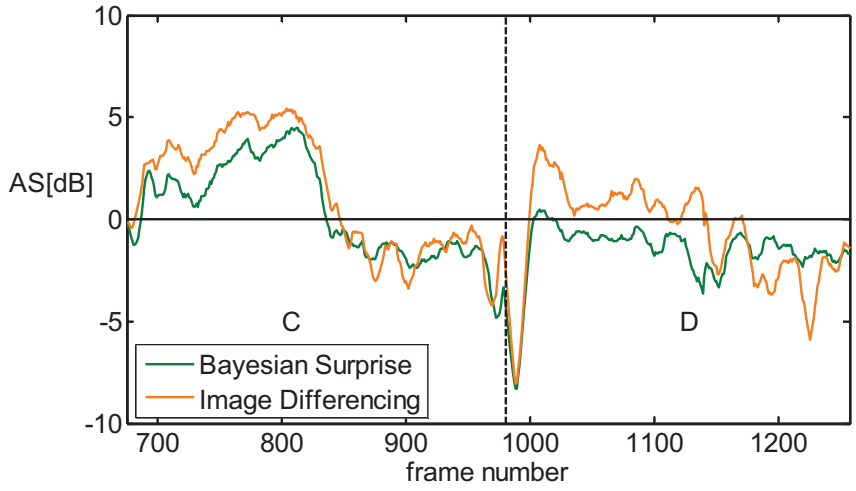
This section points out the limitations of the appearance-based modeling approach presented in this chapter and discusses the insights gained from the experimental results.

Limitations

One of the major limitations of the approach is that the robot's motion is constrained to an area which is covered by the optical tracking system. Outside this area an estimation of the camera head's pose is not possible. Occlusions of the LEDs in case of an extreme tilt of the camera head also pose problems. While an erroneous or missing pose of a single LED can be recovered using the redundancy of the others, the rotation of the camera head cannot be estimated any more if two or more LEDs are not tracked. An alternative would



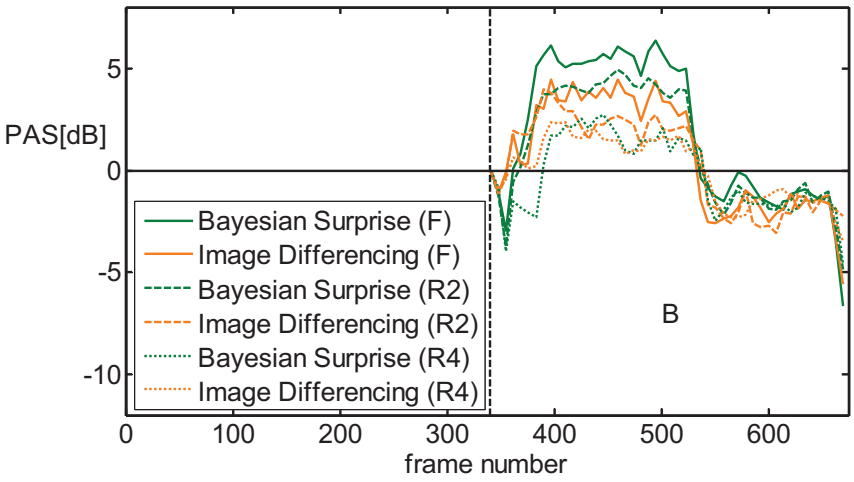
(a)



(b)

Figure 4.12: *The region of the cup is evaluated. Both the PAS values in (a) and the AS values in (b) are higher for Image Differencing than for Bayesian surprise. However, during the robot's first run from Q_1 to Q_2 the PAS and AS values obtained by Bayesian surprise are clearly above* 0 dB. *The high values of the PAS and AS obtained by Image Differencing at the beginning of phase D lead to a strong attentional selection of the region of the missing cup. However, the absence of the cup does not convey much novelty since the robot has seen the table without the cup before phase C. This is reflected by the lower AS values obained by Bayesian surprise.*

be to use visual localization methods to get the orientation and position of the camera, as described in Section 3.1.1. These methods extract a set of natural features and keep it up to date when the scene changes. However, a purely vision-based localization of the camera head is challenging due to the highly dynamic environment of the robot and the estimated



(a)

Figure 4.13: *Comparison of the performance of Bayesian surprise and Image Differencing with respect to the attentional selectivity when the number of reference views in the environment representation is reduced. The addition "(F)" refers to the environment representation which contains all 200 reference views. The additions "(R2)" and "(R4)" refer to environment models with a number of reference views reduced by a factor of 2 and 4, respectively.*

pose usually drifts from the actual pose due to error accumulation. Pose estimation from odometry data obtained from the mobile robot's wheel movements also suffers from drift and requires a very accurate estimation of the transformation between the coordinate system of the robot platform and the coordinate system of the camera.

Another issue of the appearance representation presented in this chapter is that the memory consumption of the model grows steadily in case of large-scale environments due to the large amount of densely spaced reference views. To adress this issue, compression schemes which have been introduced in the context of image-based representations can be applied. Since the parameters $\alpha_{0,k}$ in (4.2) are independent of the acquired image data (see (4.10)) their values are usually uniform across large parts of the reference parameter images. Hence, due to the strong correlations between the parameter values of neighboring pixels a high compression gain can be expected.

Furthermore, as the detection of novel changes in the environment is purely based on visual stimuli, the surprise values are low if the luminance and chrominance differences caused by the scene change are small.

Comparative evaluation

The comparison between the surprise measure based on the probabilistic appearance prior and a simple differencing of the currently captured and the predicted image shows that the proposed surprise measure is superior in terms of novelty detection. In case of novel transparent objects in a cluttered environment the image differencing method provides an

average novelty measure inside the image region of the new object which is not higher than outside this region. The appearance representation in this chapter differs from image-based representations known from computer graphics in that it explicitly models the uncertainty of the luminance and chrominance of the environment. Small displacements between corresponding intensity gradients, e.g. at object edges, in subsequently captured images lead to an elevated uncertainty. An appearance prior with higher uncertainty damps the surprise score since new luminance and chrominance values are inherently expected to lie in a broader range. Hence, the removal of a new object also leads to small surprise values. Reducing the density and augmenting the spacing between the reference views in the representation leads in general to worse attentional selectivity, since the interpolation of virtual reference parameter images from an undersampled appearance representation entails artifacts. However, the PAS drops faster for image differencing than for the surprise detection technique introduced in this chapter.

The environment representation in [RD09] is feature-based and thus provides a more compact model of the robot's surroundings which scales well with large environments. Ranganathan and Dellaert show in [RD09] results with respect to the surprise-based detection of novel navigation landmarks in hallways. However, their experiments do not include the detection of novel transparent objects in the environment. Itti and Baldi present in [IB09] a neurologically inspired approach for surprise detection. However, they do not include geometric information about the environment and the camera motion between subsequently captured views. Hence, pixel correpondences for visual stimuli across the image sequences are ignored.

4.5 Summary

This chapter presents a probabilistic appearance representation which represents the appearance of the scene at a densely spaced series of viewpoints. The luminance and chrominance of the scene are considered as random variables and modeled at each pixel of a viewpoint in terms of Gaussian distributions. While moving through the environment, the robot infers the parameters of joint prior distributions over the mean and the precision of the Gaussian models from the captured images. The robot stores belief distributions over hypotheses about how the scene is expected to look like at a given viewpoint and how large the uncertainty of the appearance is. During the computation of suprise maps the robot assesses how much the luminance and chrominance values in a new captured image change this prior belief distribution. Luminance and chrominance values which cause a strong change lead to high surprise values. To retrieve a prior belief distribution in a continuous viewpoint space, the robot interpolates the prior belief distributions from nearby reference views using a view synthesis technique which incorporates explicit geometry information.

The main contribution of this chapter is an internal environment representation for cognitive mobile robots which enables them to make photorealistic predictions of the appearance of the scene. The mathematical framework addresses both the update of the environment representation and the computation of surprise maps. The high realism of the environment model allows for the detection of image regions with high surprise level at a very early stage, as the predicted appearance can be directly related to the appearance captured by the camera. Hence, neither a costly preprocessing of the captured image nor the extraction of local

image features is necessary for attentional selection.

As shown in Chapter 5, surprise detection can be applied to the acquisition of visual object representations. The robot can learn the appearance of new unknown objects in the environment by storing selected image features inside the focus of attention, which is defined by the image region with highest surprise values.

5 Surprise-driven Acquisition of Object Representations

The environment representation in Chapter 4 provides an image-based 3D model of the robot's surroundings, which is complete within the viewpoint space covered during acquisition, since it is infered from whole images and not from sparse features. It informs the robot about the location of the objects in the environment and serves as a reference for the detection of novel objects or objects which are unexpectedly displaced or missing. However, the representation does not segment the objects in the environment and does not hold any information about their identity. The identification of given objects in the environment is an essential capability of a cognitive robot for reasoning during the execution of tasks on a higher cognitive level. Object descriptions based on raw luminance and color stimuli in general provide unsatisfactory results during recognition since this type of representation is not invariant to changes in illumination. Furthermore, object representations based on color histograms are often not distinctive enough since objects with different appearance can exhibit similar histograms. In turn, it has been shown that object descriptions based on a set of local image features allow for reliable object recognition [Low04, BETG08].

The Speeded-Up Robust Features (SURF) in [BETG08] are detected at salient image locations which provide high responses when the image is convolved with a Difference-of-Gaussians (DoG) filter. This keypoint detection is done on various scales of the image. At each keypoint a 64-dimensional vector is computed from Haar wavelet responses within an image patch around the keypoint whose size depends on the scale. This vector provides a feature descriptor which is invariant to image rotation and to illumination changes which do not change the gradient structure of the image, i.e. do not lead to shadow edges etc. The extraction of features on multiple scales ensures the recognition of the object independently of its distance from the camera.

Feature-based object recognition requires a database which contains reference features associated with each known object. One way to build such a database would be to take images of an object in front of a uniform background and store the computed features in the database. Another way would be to manually label the image regions of objects of interest. However, these processes are unnatural and tedious and a cognitive robot should be able to autonomously acquire representations of unknown objects.

Surprise, which informs the robot about novelty in the environment, can be used as a cue to direct the focus of attention of the robot to image regions which contain new objects. In case of unknown objects, the robot can selectively extract features within its focus of attention and store them in the database. This chapter describes an algorithm for such a learning procedure [MS11] and analyzes the recognition performance using the autonomously acquired feature database. The performance is compared to the performance of a reference approach.

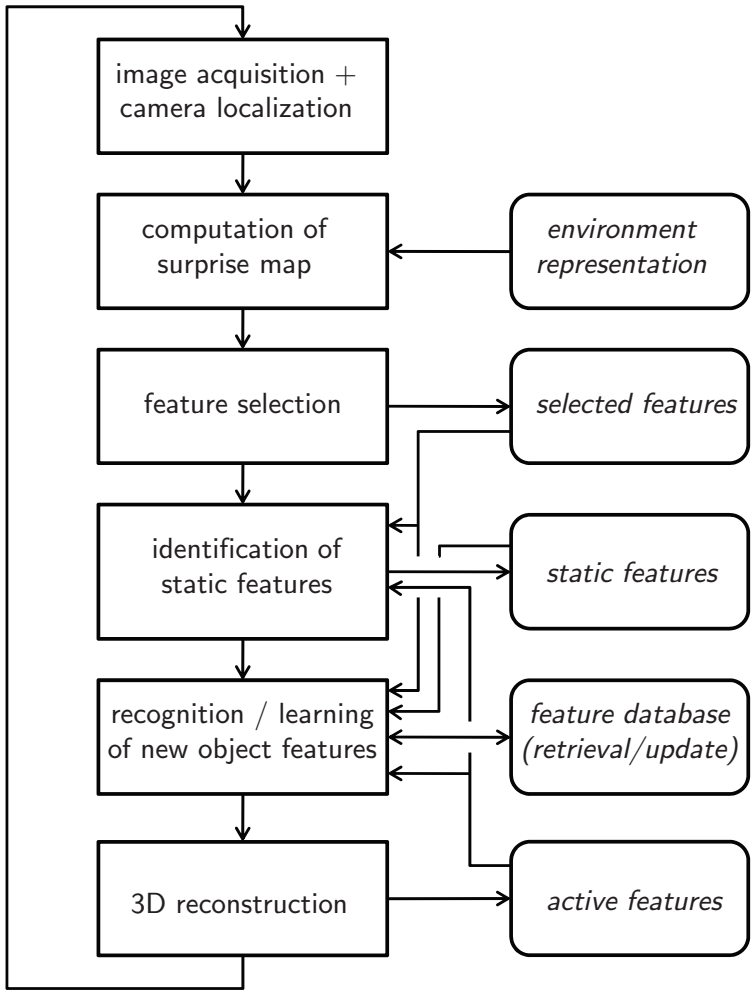


Figure 5.1: *An overview of the steps performed by the proposed algorithm for the surprise-driven autonomous acquisition of object representations.*

5.1 Description of the algorithm

Figure 5.1 shows an overview of the algorithm for the autonomous acquisition of object representations. The algorithm assumes that the robot is equipped with a multi-camera system, similar to the one in Figure 3.1(b). The image acquired by the center camera provides features for the generation of object representations. The left and the right camera images are used for the reconstruction of the 3D positions of the features. Besides, the algorithm could of course be easily adapted to a stereo camera system. For each captured image the algorithm performs the steps "Computation of surprise map", "Feature selection", "Identification of static features", "Recognition / learning of new object features" and "3D reconstruction of features". In the following, these steps are explained in detail.

Computation of surprise map

It is assumed that the robot has acquired a probabilistic appearance representation of its environment, as described in Chapter 4, before it begins to learn new objects. Hence, the environment is familiar and the robot is assumed to be able to predict its appearance with a low uncertainty. When a new image is captured, the pose of the camera is estimated and a surprise map is computed, using the environment model as prior information. As described in Chapter 4, the prior distributions in (4.2) have to be interpolated for the update of their hyperparameters, since the robot usually does not take new images exactly at the reference viewpoints. A closing and an opening operation [Ser82] are applied to the surprise map in order to widen regions of high surprise values and to remove noise.

Feature selection

The robot uses only features from the image region with the highest surprise values for further processing. Hence, it first determines the pixel with the maximum surprise score in the surprise map. This pixel is used as a seed point for the flood filling algorithm in [Bra00], which grows a region around the seed point until it encounters pixels in the surprise map whose values lie below the threshold $\theta_S = \mu_S + \xi \cdot \sigma_S$. μ_S is the mean value computed from all values in the surprise map and σ_S is the standard deviation from the mean value. In this work ξ is set to $\xi = 2$. The flood filling algorithm returns a bounding rectangle which encompasses the contour of a coherent region of high surprise. SURF features are then extracted in the part of the captured image which lies inside the bounding box. To this end, the OpenSURF library [Eva09] is used. The descriptor of the SURF features is computed as proposed in [AKB08] and is a modified version of the descriptor proposed in [BETG08] (M-SURF). The obtained set of features $\mathcal{F}$ is denoted by *selected features*.

Identification of static features

The algorithm described in this chapter is supposed to acquire representations only of novel objects which remain on a static position in the environment. Dynamic objects, and body parts of humans which are also surprising when they unexpectedly enter the field of view of the robot's camera should be ignored by the algorithm. Hence, the algorithm needs a method to determine if the selected features come from a static or a dynamic object. A selected feature $f \in \mathcal{F}$ belongs to the set of *static features* $\mathcal{SF}$ if there is an active feature $g \in \mathcal{AF}$ which fulfills the two conditions

$$\left\|\mathbf{p}_f - \mathbf{p}'_g\right\| < \delta \tag{5.1}$$

and

$$\left\|\mathbf{d}_f - \mathbf{d}_g\right\| \leq \left\|\mathbf{d}_g - \mathbf{d}_m\right\| \tag{5.2}$$

An active feature is a feature whose 3D position has been reconstructed in a previously captured image (see Step "3D reconstruction of features" below). The two conditions imply that the image location of a static feature can be predicted from its 3D position and the camera poses over a sequence of images and that the feature descriptor does not change much. Hence, the first condition the active feature must fulfill in (5.1) is that the Euclidean distance between the keypoint $\mathbf{p}_f$ of the selected feature and the pixel position of the active feature

after reprojection into the current image $\mathbf{p}'_g$ is smaller than a threshold δ. This condition takes into account slight inaccuracies of the camera pose and the resulting reprojection error of an active feature's 3D position. In this work δ is chosen as $\delta = 2$ pixels. The second condition in (5.2) requires that the Euclidean distance between the SURF descriptor $\mathbf{d}_f$ of the selected feature and the SURF descriptor $\mathbf{d}_g$ of the active feature be at most as large as the Euclidean distance between $\mathbf{d}_g$ and the descriptor $\mathbf{d}_m$. $\mathbf{d}_m$ is the descriptor of the SURF feature in the left or right camera image which has been found to be the best match during 3D reconstruction. If the frame rate of the camera is sufficiently high, the baseline between two subsequently captured views is smaller than the baseline of the views used for stereo reconstruction. Hence, considering that SURF features are only little variant to viewpoint changes, the descriptors of features in subsequently captured images are supposed to be more similar than the descriptors of features which are matched during stereo reconstruction.

Recognition / learning of new object features

The robot always tries to recognize an already known object from the selected features using its feature database. The feature database consists of a single set of SURF features $\mathcal{DF}$ which represent a set of known objects $\mathcal{O}$. In addition, the database contains for each known object $k \in \mathcal{O}$ a vector which stores for each feature $h \in \mathcal{DF}$ a number $n_{h,k}$ indicating how many times the feature has been detected on the object in past observations since its first appearance. Thus, the relative frequency $p_{h,k}$, which is the ratio of the number of detections for a feature on a given object with respect to the number of detections for this feature on all objects, is computed by

$$p_{h,k} = \frac{n_{h,k}}{\sum_k n_{h,k}} \tag{5.3}$$

The algorithm assigns a feature $h \in \mathcal{DF}$ to an object $k \in \mathcal{O}$ by looking for the highest relative frequency $\max_k (p_{h,k})$.

For each known object k a variable N_k counts the number of matches between the database features which belong to the object and the selected features. A selected feature matches a feature from the database if the Euclidean distance between their descriptors is minimum with respect to other selected features and if the distance to the descriptor of the second best selected feature is smaller by a factor of at least 0.65. The object class c is then retrieved from the database by looking for the maximum value of all N_k.

$$c = \left\{ k \in \mathcal{O} \mid \max_k (N_k) \right\} \tag{5.4}$$

Each feature match also increases the variables $n_{h,M}$ by 1 where M refers to the most probable object class for the database feature h. If $N_k = 0 \; \forall k \in \mathcal{O}$, the object is not recognized from the database. The algorithm then tries to recognize the object from the active features in a similar way as it does using the database features.

If the algorithm cannot find any matches between the selected features and the active features either, an unknown object is detected in the bounding box. Then a new object class v is created in the database, if the condition

$$\exists b \in \mathcal{SF} \; : \rho_S > T \tag{5.5}$$

is fulfilled. ρ_S is the percentage of the pixels within the image region of the descriptor of the static feature b whose surprise values are above the threshold θ_S in Step "Feature selection" (see above). In this work we choose $T = 0.8$. Only static features $\mathcal{SF}_{\text{db}} \subseteq \mathcal{SF}$ which fulfill the condition in (5.5) are then added to the set of features $\mathcal{DF}$ in the database ($\mathcal{DF} := \mathcal{DF} \cup \mathcal{SF}_{\text{db}}$). Thus, this condition prevents that feature descriptors computed from an image patch which contains too much background are added to the database. The variables $n_{h,k}$ are initialized for all new features in the database as

$$n_{h,k} = \begin{cases} 1 & k = v \\ 0 & k \in \mathcal{O} \setminus \{v\} \end{cases} \quad \forall h \in \mathcal{DF} \cap \mathcal{SF}_{\text{db}} \tag{5.6}$$

If a known object is recognized in (5.4) and unknown static features $\mathcal{SF}_{\text{db}} \subset \mathcal{SF}$ which fulfill the condition in (5.5) are detected, they are also added to the set of features in the database. The variables $n_{h,k}$ are then initialized for the new features in a similar way

$$n_{h,k} = \begin{cases} 1 & k = c \\ 0 & k \in \mathcal{O} \setminus \{c\} \end{cases} \quad \forall h \in \mathcal{DF} \cap \mathcal{SF}_{\text{db}} \tag{5.7}$$

3D reconstruction of features

If there are selected features which are not classified as static features and which are not in the database but which have a match in either the left or the right image, their 3D position in the environment is reconstructed. Thereby, the algorithm looks for matches on the epipolar lines in the left and right image. The criterion for the descriptor distances, which has to be fulfilled for a feature match, is the same as in Step "Recognition / learning of new object features". The selected features whose 3D position is known are called *active features*. They are kept in working memory only when the robot learns a new object and are not permanently added to the database. The image patches used for the computation of their descriptors do not have to fulfill the condition in (5.5) and thus can contain background. Active features are accumulated over the captured image sequence and are used to determine static features (see Step "Identification of static features") or to recognize the new object and its surroundings (see Step "Recognition / learning of new object features") when this is not possible using the database features. Hence, active features make the acquisition of object representation more robust.

When no new object is present in the captured images, the bounding box defining the focus of attention is usually very small and its position is guided by noise. Hence, the algorithm in general does not extract features within the focus of attention in this case. Once the robot cannot extract any features from the bounding box over a period of 10 images, the active features are discarded, since it is assumed that the new object has been removed again from the scene.

5.2 Experimental results

To evaluate the algorithm proposed in this chapter three image sequences $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$ are captured, using the robot platform "Cobot" depicted in Figure 4.3(a). The images are taken at

a resolution of 1280×960 at a rate of about 7 frames per second and subsampled to 320×240 pixels for further processing. The probabilistic environment model is infered from images taken of a scene which consists of a table with some prior objects on it. The objects, the table cloth and the background are strongly textured so that the image features describing the appearance of a new object to be learned have to be selected carefully among all image features extracted from the cluttered scene. The environment representation encompasses 180 views which are densely distributed between the points Q_1 and Q_2 in Figure 4.3(b) on a trajectory length of $1.6\,\mathrm{m}$. The Gaussian distributions at the pixels of each reference view are infered from 360 subsampled images of the first sequence $\mathcal{I}_1$, which is acquired during one run of the robot from Q_1 to Q_2 and back. For each of these images a depth map is computed, as described in Section 3.2. The camera pose is estimated for each image in $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$ using the optical tracking system in Section 3.1.2.

During the acquisition of the next image sequence $\mathcal{I}_2$ three objects are sequentially presented to the robot. During the first run from Q_1 to Q_2 a cup ("Object 1") is added to the objects on the table by a human. When the robot arrives at Q_2, the human removes the cup and adds a biscuit box ("Object 2"). Next, when the robot goes back and arrives at Q_1, the human removes the biscuit box and adds an ice tea box ("Object 3"). In its final run, the robot goes again to Q_2. In total, 600 images are taken of the scene with the different objects.

Figures 5.2(a) to 5.2(c) show the three objects from different viewpoints along the way between Q_1 and Q_2. The expected appearance which the robot interpolates from reference views in its environment representation is shown in Figures 5.2(d) to 5.2(f). The images show that the expected appearance, which corresponds to the hyperparameters τ_k of the prior distribution in (4.2), provides a photorealistic reference of the environment as it has been captured in image sequence $\mathcal{I}_1$. Figures 5.2(g) to 5.2(i) show the surprise maps which are computed from the interpolated prior in (4.2) and the posterior in (4.3). The surprise maps clearly show a region of high surprise values at the positions of the new objects. Figures 5.2(a) to 5.2(c) depict the bounding rectangles which are obtained by the flood filling algorithm described in Section 5.1 and define the region where the SURF features are selected. In case of the cup and the red biscuit box, the bounding box encompasses the whole object. In case of the ice tea box, the bounding box surrounds most of the lower part of the object. As the surprise map in Figure 5.2(i) shows, there is a line with relatively low surprise values where the table border in the background is expected. This terminates the flood filling.

Furthermore, it is investigated how well the surprise maps guide the position of the bounding box in sequence $\mathcal{I}_2$ and direct the robot's attention towards the novel objects in the scene. Figures 5.3(a) and 5.3(b) show the horizontal (x) and vertical (y) pixel position of the center of the bounding box in the acquired images computed in the step "Feature selection" of the algorithm in Section 5.1. The origin of the pixel coordinate system is here in the upper left corner of an acquired image. Figures 5.3(a) and 5.3(b) also indicate the ground truth values which are obtained by manually labeling bounding boxes which completely and tightly include the objects in the images. The ground truth is determined for those frames which are captured when the objects stand completely still on the table. As the plots show, the measured x-positions are in most frames very close to the corresponding ground truth values. The fluctuations between frame 250 and 300 happen because, in some frames, the surprise region is only grown over the upper left part of Object 2 and the bounding box does not cover the whole biscuit box. The y-positions show a similar behavior for Object 1 and Ob-

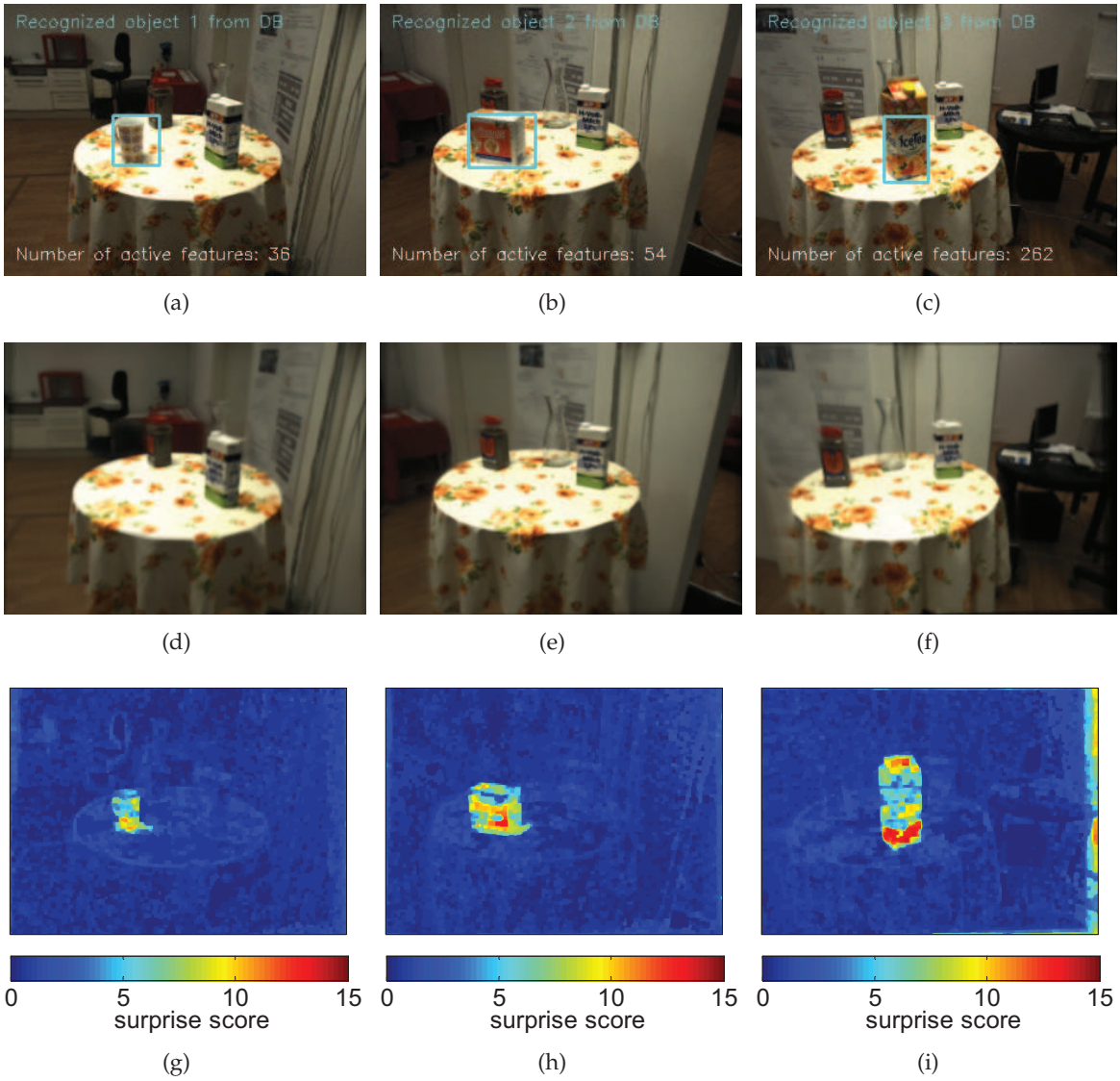


Figure 5.2: *(a)-(c): Images of the sequence $\mathcal{I}_2$ captured by the robot during the acquisition of the object representations. They show a cup, a red biscuit box and a large ice tea box which a human added to the scene. The images also show the cyan bounding box which is computed by the flood filling algorithm. (d)-(f): The expected appearance computed from the internal environment representation of the robot for the three viewpoints in (a) to (c). (g)-(i): The surprise maps computed by the robot clearly indicate the new objects in the scene.*

ject 2. When Object 3 is acquired, however, the measured y-coordinate of the bounding box differs from the ground truth data by approximately 20 pixels. In those frames where the y-coordinate lies above the ground truth the lower part of the ice tea box is covered (see Figure 5.2(c) for an example) and in those frames where the y-coordinate lies below the upper part is covered by the bounding box. Nevertheless, a sufficient number of features can still be extracted within the bounding box for object learning, as Figure 5.3(c) shows. Figure 5.3(c) also shows that when an object is present many more features are extracted within

the bounding box than when the scene does not contain any new object because in that case the bounding boxes are very small. The lack of features in the periods when the objects are changed leads to the deletion of all active features (see Step "3D reconstruction of features" of the algorithm in Section 5.1).

To evaluate the quality of the feature selection and learning steps of the algorithm proposed in this chapter, the recognition performance is investigated when the acquired objects are placed in a different part of the laboratory with different background and lighting conditions. Figure 5.4 shows two images taken from the Objects 2 and 3. These are two images from the sequence $\mathcal{I}_3$ which the robot takes on an approximately circular trajectory around the scene between two turning points, similar as depicted in Figure 4.3(b). During the first run of the robot, Object 3 is presented and during its way back Object 2 is placed into the scene. The squares in cyan color illustrate matches between features extracted from the image and features stored in the database.

Figure 5.5(a) shows the number of features in the images of $\mathcal{I}_3$ which can be matched with a feature in the database, respectively. As explained in Section 5.1, a feature in the database is assigned to the object for which the relative frequency $p_{h,k}$ in (5.3) is highest. The ground truth in Figure 5.5(b) indicates in black which objects are indeed visible in the images of $\mathcal{I}_3$. A comparison shows that mostly features are detected which belong to the object in the scene. In rare cases, features are wrongly matched with features from an object which is not visible in the frame.

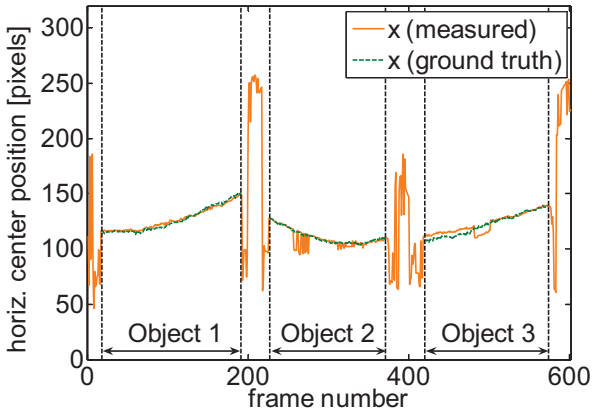
The results obtained from the approach presented in this chapter are compared to the results obtained from the approach in [WIS+10], using the same image sequences $\mathcal{I}_1$ to $\mathcal{I}_3$ and the same parameterization for the computation of the SURF features. The method in [WIS+10] computes an Eigenbackground image, as described in [ORP00], which is used for the segmentation of the new object. A representation of the object is then acquired by extracting features from an image in which the background is set to a uniform color (e.g. gray). Since in the experimental setup described in this work the camera is not static, the Eigenbackground for an image in $\mathcal{I}_2$ is computed from seven warped reference images from $\mathcal{I}_1$. The reference images are chosen to be the closest ones in space with respect to the image in $\mathcal{I}_2$, according to the view selection method in Section 3.3.1. While the segmentation of the new object in general performs well using this method, the recognition of the Objects 2 and 3 in $\mathcal{I}_3$ is worse compared to the proposed method. Figure 5.5(c) shows that in general more features are matched on the two objects, which is also due to the higher amount of features in the database. However, many false matches are found in the images.

Furthermore, the receiver operating characteristic (ROC) curve is evaluated for the approach proposed in this chapter and the approach in [WIS+10]. A ROC curve is obtained by plotting the true positive rate (TPR) versus the false positive rate (FPR), where the TPR describes the proportion of true positive (TP) feature matches among the union of true positive and false negative (FN) feature matches

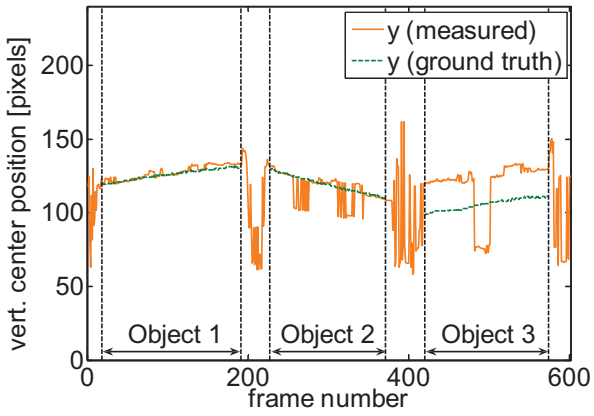
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{5.8}$$

The false positive rate (FPR) describes the proportion of false positive (FP) feature matches among the union of false positive and true negative (TN) feature matches

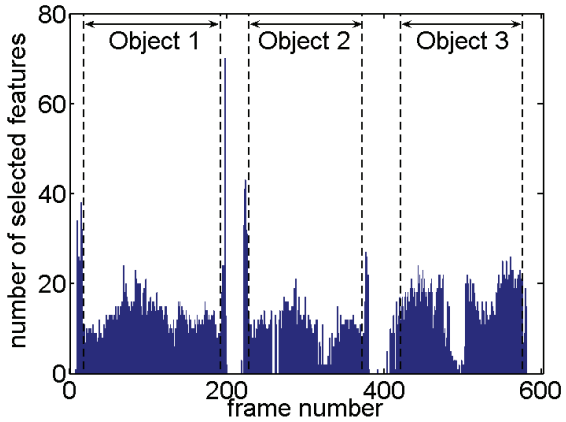
$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{5.9}$$



(a)

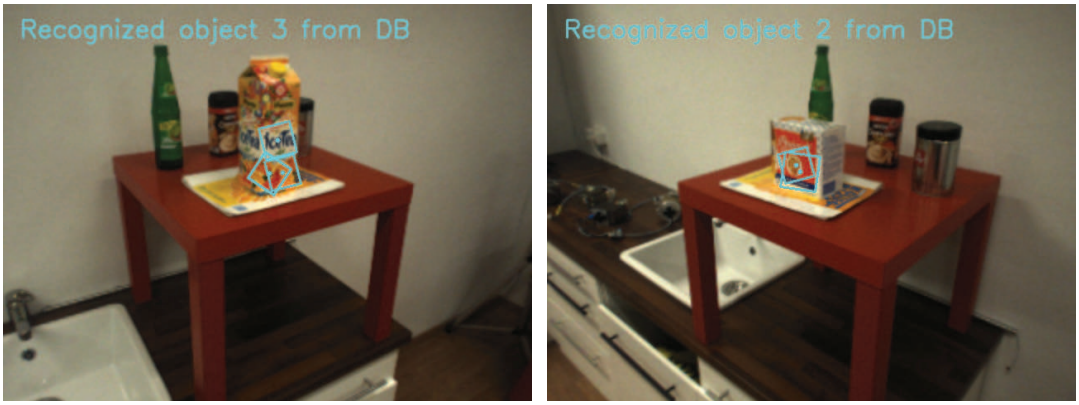


(b)



(c)

Figure 5.3: *(a)-(b): The horizontal (x) and vertical (y) pixel position of the bounding box center in the captured images of sequence $\mathcal{I}_2$. The plots show both the measured data and the ground truth. (c): The number of features which the robot selects within the bounding box in the frames of $\mathcal{I}_2$.*



(a) (b)

Figure 5.4: *Two images from sequence $\mathcal{I}_3$ which show the ice tea box and the biscuit box in front of a different background. The cyan squares in the images visualize SURF descriptors which are successfully matched with descriptors from the database.*

between image and database. The two ROC curves are shown in Figure 5.5(d). The pairs of TPR and FPR values along the curves are obtained by varying a threshold for the number of features in the image which have to match features of a given object in the database for successful object recognition. It is shown that the method in Section 5.1 achieves a maximum TPR of 81.5% at a very low FPR of 0.3%. The approach in [WIS+10], in comparison, reaches a TPR of 81.5% at a FPR of 7.2%. As the plot in Figure 5.5(d) shows, higher TPRs can be achieved by [WIS+10] but at the cost of rapidly increasing FPRs.

5.3 Discussion

This section points out the limitations of the approach for surprise-driven acquisition of object representations presented in this chapter and discusses the insights gained from the experimental results.

Limitations

As the approach in [WIS+10], the approach proposed in this chapter is not able to detect an object and segment it from the environment if the appearance of the object is very similar to the appearance of the background in that region. Small luminance or chrominance changes with respect to the expected appearance cause small surprise values which do not differ much from the surprise values in the rest of the image. Hence, a prominent region of surprise with values which lie above the threshold θ_S in the step "Feature selection" of the algorithm in Section 5.1 cannot be found. Here, geometry information would help in the detection of the object.

Geometry data would also be helpful to distinguish if surprise is detected because a new object has been added to the scene or if surprise has been detected because an object has been removed from the scene. The latter case is not considered by the algorithm in Section 5.1 and

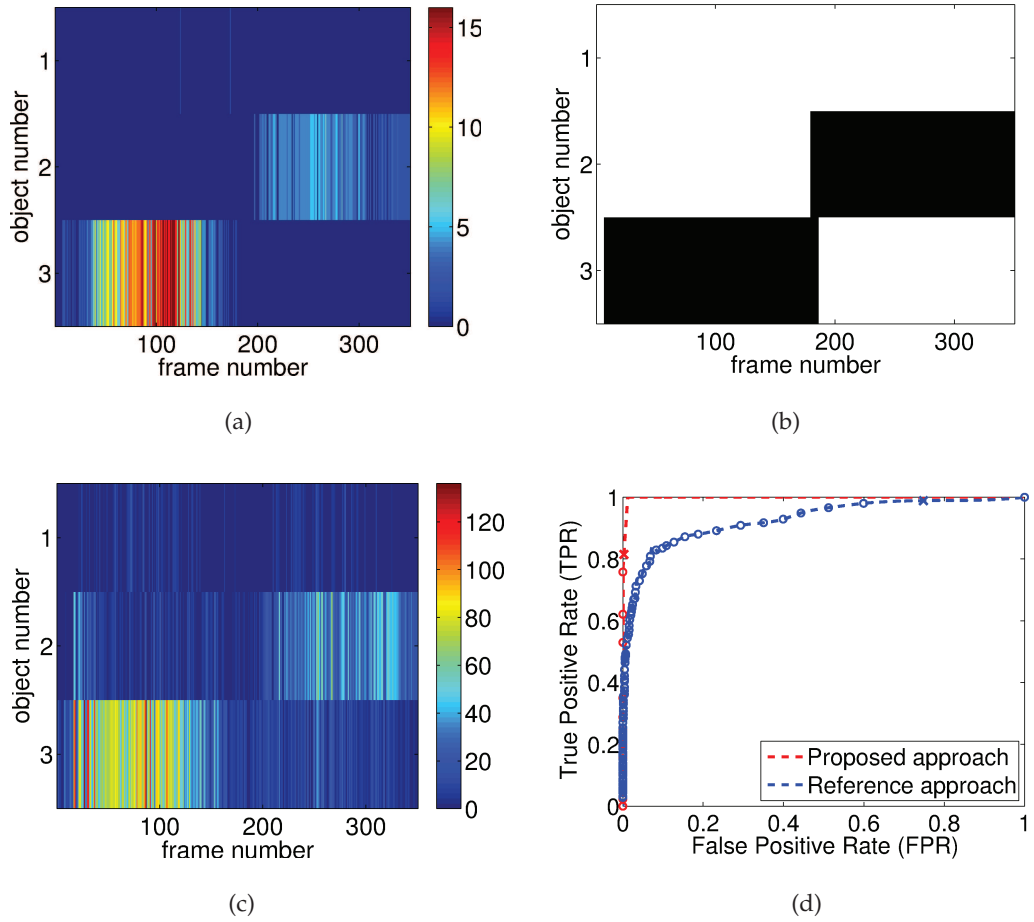


Figure 5.5: *(a): The number of matches between features extracted in the images of $\mathcal{I}_3$ and features from the database using the object representations obtained by the algorithm proposed in this chapter. The features are assigned here to the object for which the probability that the feature belongs to it is highest. (b): Ground truth which indicates in black which objects are in fact visible in the images. (c): The number of matches between features extracted in the images of $\mathcal{I}_3$ and features from the database using the object representations obtained by the algorithm in [WIS+10]. (d): A comparison of the receiver operating characteristic of the proposed approach and the approach in [WIS+10].*

as a consequence features extracted from the background would be extracted and possibly learned as a new object representation. With the aid of geometry data, the algorithm can check if the mean scene depth is smaller in the region of high surprise than predicted from the environment model or larger. If it is smaller, a new object has been added. In constrast, if it is larger, an object has been removed, which means that the algorithm must not learn any new features.

A further limitation is that SURF features can only be detected on object surfaces with prominent intensity gradients. Objects which do not exhibit any texture like unicolor cups do not provide any keypoints which are interesting for the SURF algorithm. On unicolor objects keypoints are detected at the borders if there is an intensity gradient with respect to

the background. However, the descriptors of these features are not added to the database by the proposed algorithm since the image patches from which they are computed contain too much background. It would not make sense to include these features in the object representation since they would not be detected on the object again if the background changes. Shape-based descriptions provide representations which are more appropriate for this type of objects.

Comparative Evaluation

The approach in [WIS+10] allows for the acquisition of images from many different viewpoints around the object since the robot's arm can rotate the object in manifold ways. In contrast, the viewpoint space in the experiments in Section 5.2 is limited. Images which show the bottom side of the object cannot be captured unless the object on the table is flipped by the human during the acquisition of the representation. The viewpoint space, however, can be extended both for the acquisition of the probabilistic appearance representation and for the acquisition of the object representation by employing a fully humanoid robot platform, as used in [SLL+07].

After the acquisition of representations of the three objects in Section 5.2 the database comprises 5531 SURF descriptors using the approach in [WIS+10] and 88 SURF descriptors using the approach proposed in Section 5.1. Since, the approach in [WIS+10] extracts features from an image which exhibits the segmented object in front of a uniform background and not from a selected image region, the database contains many features. The keypoints of the features are detected on the object, however, the image patches used for the computation of the descriptors can contain much of the uniform background. This holds especially if the features are extracted on a large scale or at the border of the object to the uniform background. It is unlikely that these features later match features extracted from a camera image for object recognition, unless the object is situated in front of a uniform background, which is a rare case in natural environments. Furthermore, if the object is acquired while the robot holds it in its hand, there are always parts of the object which are occluded by the hand. In these image regions no features can be extracted. Besides, keypoints are detected at the border between a region where the object is visible and an image region where the object is occluded and where, hence, the color has been set to the background color. These intensity contrasts are unnatural.

As the experimental results in Section 5.2 show, much more features can be matched during object recognition if the database is created using the method in [WIS+10]. However, there are also many false matches due to the large amount of features which are not very descriptive with respect to the appearance of the objects. The careful selection of features in the approach proposed in this chapter keeps the false positive rates comparably low.

5.4 Summary

This chapter presents an algorithm for the acquisition of visual object representations using a mobile robot platform. A probabilistic appearance representation of the environment, which the robot acquires before, allows for the detection of new unknown objects by surprise. High surprise values in the region of the new object capture the attention of the robot

and a rectangular image region which encompasses the new object is analyzed for SURF features. The descriptors of unknown features are added to a database if the features are static and if a sufficiently large proportion of the image patch used for the computation of the descriptor covers high surprise values in the surprise map. Experimental results show that the proposed method outperforms a state-of-the-art approach.

The main contribution of this chapter is a scheme for the acquisition of visual 3D object representations in cluttered environments while the robot's camera can freely move around the object in a space where a prior environment model is available. The process is driven by surprise so that the algorithm autonomously detects whether a new object is present in the scene and when the acquisition starts and ends. A novel feature selection strategy creates a database with a small set of characteristic features.

One issue which arises with the appearance-based segmentation of objects, as described in this chapter, is the insufficient robustness against illumination changes which can happen between the acquisition of the environment model and the acquisition of the object representations. Shading, shadows and specularities can cause regions of high surprise. Since the intensity gradients inside dark shadow or specularity regions usually are small, the number of SURF features extracted in these areas is small. However, illumination effects can distract the robot's attention to image regions which are not relevant for object learning. To this end, Chapter 6 deals with the detection of novel objects in a robot's environment under varying illumination and presents methods to suppress illumination effects.

6 Illumination-invariant Image-based Novel Object Detection

If novelty detection is based on low-level visual cues like the intensity or color captured in camera views, novelty is not only caused by changes in the environment like new or removed objects but also by illumination changes. Since in many cases illumination changes are not relevant for the tasks of a robot and distract its attention, mechanisms are required to lower the influence of shadows, shading and specularities during attentional selection. To this end, this chapter presents methods to detect changes in the environment while suppressing regions of high novelty caused by illumination changes.

A fundamental step is the acquisition of an illumination-invariant image-based model which represents the appearance of the environment at a densely spaced series of viewpoints in terms of a set of images which are free of illumination effects [MBS+09]. This is achieved by capturing multiple image sequences of a static environment under different illumination conditions using a mobile robot. In addition to the images, depth maps are computed and the pose of the camera is estimated. Illumination effects are removed in the gradient domain using an image interpolated from each image sequence at a defined viewpoint. A statistical model for the intensity and color saturation variation caused by the illumination of the scene is infered and stored for each view in the illumination-invariant image-based representation [MBM+10]. This statistical model is then used to identify image regions which exhibit intensity and/or color values which are not typical for illumination changes.

A challenging issue in this context is the detection and suppression of specularities. In this chapter a method is presented which detects specularities in the images acquired for the recovery of the illumination-invariant image-based representation and models them in a shape-based approach [MES11]. The experimental results in this chapter show that the addition or removal of objects can be detected by the proposed methods while illumination effects including pronounced specularities on metallic surfaces are ignored.

6.1 Acquisition of illumination-invariant image-based environment representations

To obtain an image-based environment representation which is free of illumination effects, M image sequences $\mathcal{I}_m,\ m = 1, \dots M$ are captured from a scene by a multi-camera system on a mobile robot platform, while the illumination conditions vary from sequence to sequence. Virtual images are rendered from each acquired image sequence at identical viewpoints around the scene. To use the view synthesis technique in Section 3.3, a depth map is computed for each captured image, as described in Section 3.2. The pose of the camera system is determined using the visual localization technique in Section 3.1.1 or the optical tracking system in Section 3.1.2. The virtual images are interpolated at a series of viewpoints whose spacing is similarly dense as the spacing of the real camera views. The obtained M

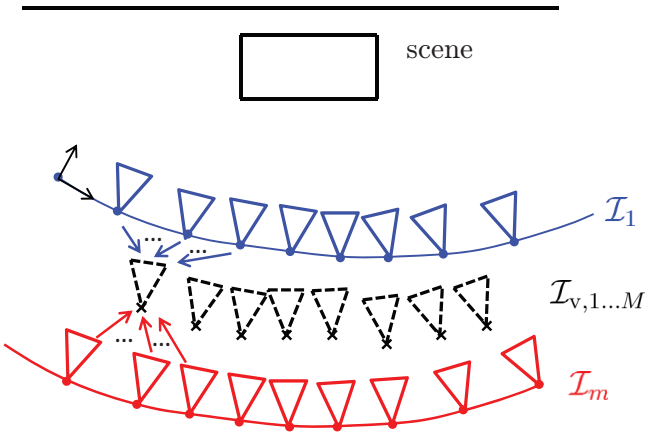


Figure 6.1: *The crosses with the dashed viewing frusta illustrate the interpolation of virtual images at a dense series of defined viewpoints. At a given viewpoint, a virtual image is rendered from each acquired image sequence, respectively.*

sequences of virtual images are denoted by $\mathcal{I}_{\mathrm{v},m},\ m = 1, \ldots, M$, as depicted in Figure 6.1.

6.1.1 Registration of multiple image sequences

If the camera pose is determined using an optical tracking system, the poses of all camera views are obtained with respect to the same world coordinate system. This is not the case if an image-based pose estimation is applied. The acquisition of the M image sequences can happen over a longer period of time and the robot might be switched off in between. As pointed out in Section 3.1.1, however, loosing the set of KLT features stored and updated by the visual localization algorithm along the image sequence makes it impossible to determine the pose of future camera views with respect to the coordinate system initialized with the acquisition of the first image. In this case, the coordinate system is re-initialized and the poses of new camera views are determined with respect to the new origin. Hence, if a virtual image is supposed to be rendered at the same absolute viewpoint in space from different image sequences, the relationship between the coordinate systems is required.

In the following, the images in a given sequence $\mathcal{I}_m$ are indexed by j. The coordinate frame of the first sequence $\mathcal{I}_1$ is used as the common coordinate frame of all sequences. To register a particular sequence $\mathcal{I}_m, \quad m = 2, ..., M$ to this coordinate frame, a sparse set of N_S support views is taken from it [MBM+10]. For illustration, three support views are represented by the camera viewing frusta in Figure 6.2.

The support views are inserted into the image sequence $\mathcal{I}_1$ between similar images, which results in an augmented image sequence $\mathcal{I}'_1$. The transformation matrices of the support views $\mathbf{S}_{m,j_n}$ $(n = 0, \ldots, N_\mathrm{S}-1)$ with respect to the common coordinate frame are determined by applying the visual localization technique in Section 3.1.1 to $\mathcal{I}'_1$. Using the transformation matrices of the support views and their poses $\mathbf{M}_{m,j_n}$ $(n = 0, \ldots, N_\mathrm{S} - 1)$ with respect to the coordinate frame of $\mathcal{I}_m$, the transformation between the two coordinate frames of the sequences is calculated, respectively, for each support view by

$$\mathbf{Q}_{m,j_n} = \mathbf{S}_{m,j_n} \cdot \mathbf{M}^{-1}_{m,j_n} \quad n = 0, \ldots, N_\mathrm{S} - 1. \tag{6.1}$$

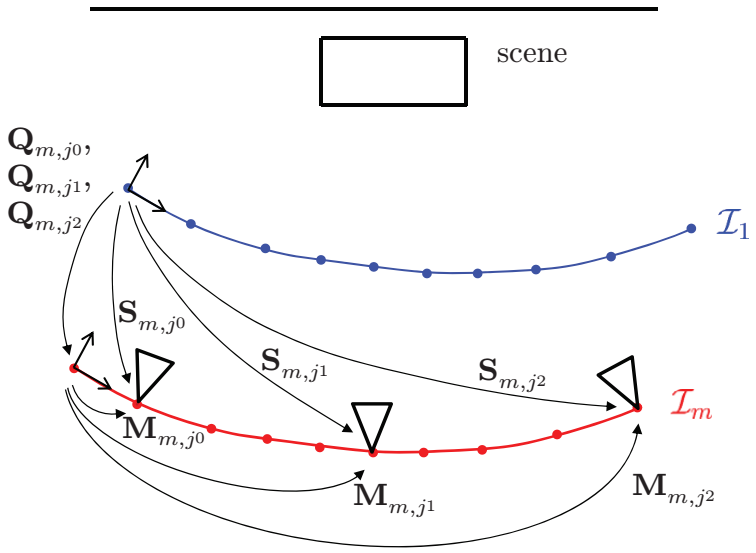


Figure 6.2: *All acquired image sequences are registered to a common coordinate frame. Using several support views in an image sequence $\mathcal{I}_m$, the transformation between the sequence and the common coordinate frame is determined.*

Although describing the same relationship, these matrices can be different from each other due to the error propagation in the visual localization over the image sequence $\mathcal{I}_m$. Using $\mathbf{Q}_{m,j_n}$, the rotation matrix and the translation vector in the transformation matrix $\mathbf{Q}_{m,j}$ between the two coordinate frames is interpolated over the whole image sequence $\mathcal{I}_m$ by second-order Lagrange polynomials. With three support views the three columns $\mathbf{r}_{t,m,j}$ ($t \in \{1,2,3\}$) of the rotation matrix in $\mathbf{Q}_{m,j}$ are interpolated by

$$\mathbf{r}_{t,m,j} = \frac{(j-j_1)\,(j-j_2)}{(j_0-j_1)\,(j_0-j_2)} \cdot \mathbf{r}_{t,m,j_0} + \frac{(j-j_0)\,(j-j_2)}{(j_1-j_0)\,(j_1-j_2)} \cdot \mathbf{r}_{t,m,j_1} + \frac{(j-j_0)\,(j-j_1)}{(j_2-j_0)\,(j_2-j_1)} \cdot \mathbf{r}_{t,m,j_2}, \tag{6.2}$$

where the vectors $\mathbf{r}_{t,m,j_n}$ are the respective columns of the rotation matrix in $\mathbf{Q}_{m,j_n}$. Likewise, the translation vector in $\mathbf{Q}_{m,j}$ is interpolated from the translation vectors in the transformation matrices $\mathbf{Q}_{m,j_0}$, $\mathbf{Q}_{m,j_1}$ and $\mathbf{Q}_{m,j_2}$. To fulfill the orthonormality constraint on general rotation matrices, the column vectors of the rotation matrix in $\mathbf{Q}_{m,j}$ are normalized to unit length. The poses $\mathbf{S}_{m,j}$ of the images in I_m with respect to the common coordinate frame are then calculated by

$$\mathbf{S}_{m,j} = \mathbf{Q}_{m,j} \cdot \mathbf{M}_{m,j}, \tag{6.3}$$

where $\mathbf{M}_{m,j}$ is the pose of a given image in $\mathcal{I}_m$ with respect to its first image.

6.1.2 Computation of illumination-invariant images

Using the virtual images rendered at a viewpoint depicted by a cross in Figure 6.1, an illumination-invariant image is recovered. For the computation of an illumination-invariant image from several intensity images taken at the same viewpoint under different illumina-

tion, an efficient method, which has originally been proposed in [Wei01], is applied [1]. The technique is briefly revisited in the following.

As given in (2.28), an intensity image captured by a camera can be decomposed into the product of an illumination-invariant image and an illumination image. The illumination model in (2.28) is based on the assumption that the surfaces of the scene are diffuse reflectors. However, the decomposition of a camera image into a product of an illumination-invariant image and an illumination image can also be derived from Phong's model in (2.26). Considering that a nonperfectly reflecting surface is illuminated by a light source at a given position, Phong's model can be written as

$$I_\lambda = \underbrace{k_\mathrm{d} O_{\mathrm{d}\lambda}}_{R_\lambda} \cdot \underbrace{\left[I_{\mathrm{p}\lambda} \cdot \vec{N} \cdot \vec{L} + \frac{I_{\mathrm{p}\lambda} k_\mathrm{s}}{k_\mathrm{d} O_{\mathrm{d}\lambda}} \left(\vec{R} \cdot \vec{V} \right)^n \right]}_{L_\lambda} . \tag{6.4}$$

Here, R_λ represents an illumination-invariant term, which models the diffuse spectral reflectance of the surface, and a term L_λ, which contains illumination effects and which varies for different position of the light source. The ambient light term from (2.26) is omitted and $f_\mathrm{att} = 1$. To ensure that the fraction in the second summand of L_λ in (6.4) is defined, R_λ must not be zero. Very small values in the intensity image are therefore clipped before the illumination-invariant image is recovered.

The model which decomposes a camera image into an illumination-invariant part and an illumination part can also be used for photorealistic virtual images synthesized by the graphics hardware of a computer. In the logarithmic domain, a virtual image rendered from eight real images in sequence $\mathcal{I}_m$ at the j-th viewpoint is thus given by

$$i_{\mathrm{v},m,j} = r_j + l_{m,j}, \tag{6.5}$$

where r_j is the logarithmic illumination-invariant image and $l_{m,j}$ the logarithmic illumination image.

The convolution of the virtual image with the filters $f^x = [1, -1]$ and $f^y = [1, -1]^\mathrm{T}$ provides the horizontal and vertical gradients of the image, respectively,

$$\nabla_x i_{\mathrm{v},m,j} = f^x * i_{\mathrm{v},m,j} \tag{6.6}$$

$$\nabla_y i_{\mathrm{v},m,j} = f^y * i_{\mathrm{v},m,j}. \tag{6.7}$$

Although not explicitly expressed, the computation of the gradients in (6.6) and (6.7) and all further computations are done in RGB domain, separately for each color channel.

Since illumination gradients are sparse and since the illumination-invariant image at a viewpoint remains constant over all image sequences, the gradients of the illumination-invariant image can be recovered by

$$\nabla_x \hat{r}_j = \underset{m=1,\ldots,M}{\mathrm{median}} \left(\nabla_x i_{\mathrm{v},m,j} \right) \tag{6.8}$$

$$\nabla_y \hat{r}_j = \underset{m=1,\ldots,M}{\mathrm{median}} \left(\nabla_y i_{\mathrm{v},m,j} \right). \tag{6.9}$$

[1] In [Wei01] the illumination-invariant image is also denoted by *reflectance image*. Indeed, it corresponds to the reflectance of the scene if it is assumed that all surfaces in the environment are Lambertian. Since this chapter also considers environments with specular surfaces, the recovered image is free of illumination effects but does not always represent the reflectance of the scene. Hence, the expression *reflectance image* is not used in this chapter.

This step removes all gradients which stem from time-variant illumination effects like shading gradients, shadow borders or the edges of specularities.

As the convolution in (6.6) and (6.7) is a linear operation, the following overconstrained system of equations is obtained for the recovered gradients of the illumination-invariant image

$$\nabla_x \hat{r}_j = f^x * \hat{r}_j \tag{6.10}$$
$$\nabla_y \hat{r}_j = f^y * \hat{r}_j. \tag{6.11}$$

The equations (6.10) and (6.11) are solved for the illumination-invariant image $\hat{r}_j$ using pseudoinverse filtering

$$\hat{r}_j = \left(\nabla_x \hat{r}_j * \tilde{f}^x + \nabla_y \hat{r}_j * \tilde{f}^y\right) * g. \tag{6.12}$$

The filters $\tilde{f}^x$ and $\tilde{f}^y$ are reversed versions of the respective gradient filters in (6.10) and (6.11). The filter g only depends on the gradient filters and is chosen such that it fulfills the equation

$$\left(f^x * \tilde{f}^x + f^y * \tilde{f}^y\right) * g = \delta, \tag{6.13}$$

where δ is the Dirac impulse.

The logarithmic illumination image corresponding to the virtual image rendered from $\mathcal{I}_m$ at the j-th viewpoint is obtained by

$$\hat{l}_{m,j} = i_{\mathrm{v},m,j} - \hat{r}_j. \tag{6.14}$$

6.2 Statistical models for the intensity and saturation values in illumination images

For Lambertian environments the second term of L_λ in (6.4) vanishes ($k_\mathrm{s} = 0$), which means that an illumination image describes the modulation of the light source intensity by the orientation of the surface normal and by occlusions, which lead to shadows. Hence, the illumination image is an image with gray shades if the spectrum of the light source does not change during the acquisition of the image sequences. An example for an illumination image in linear domain is shown in Figure 6.3(c). It is amplified for visualization. The correponding illumination-invariant image and the camera image are shown in Figures 6.3(b) and 6.3(a), respectively. Since the surfaces of the scene are largely Lambertian, the illumination image appears grayish and with little color saturation, as predicted by (6.4).

In image regions which exhibit strong specularities, in turn, the first term of L_λ in (6.4) can be neglected. Since the direction of light reflection $\vec{R}$ can be assumed to be very similar to the direction of the observer's viewpoint $\vec{V}$, the illumination image can be approximated in these regions by

$$L_\lambda \approx \frac{k_\mathrm{s} I_{\mathrm{p}\lambda}}{k_\mathrm{d} O_{\mathrm{d}\lambda}}. \tag{6.15}$$

Thus, (6.15) shows that the illumination image describes the ratio between the specularly reflected light and the diffuse reflectance of the surface. In case of an illuminant emitting

(a) (b) (c)

Figure 6.3: *(a) Camera image. (b) Illumination-invariant image. (c) Illumination image.*

light of a bright color (e.g. white or yellow), the intensity of the specular reflection in the numerator is in general larger than the intensity of the diffuse reflection in the denominator. Therefore, the luminance of specularity regions is relatively high in L_λ. Furthermore, if the color of the illuminant differs from the intrinsic color of the surface, the color saturation in specularity regions is increased in L_λ.

Figure 6.4(a) shows a camera image and Figure 6.4(b) its corresponding illumination image, which exhibits these effects on the surface of the green object and on the surface of the tin can. The specularities on the green object lead to blueish color shades in the illumination image. This is also observed for the specularity on the metal cap of the mug.

To examine the luminance and the saturation components of multiple illumination images, nine camera images of the scene in Figure 6.4(a) are taken under varying lighting conditions. In HSV color space, the histograms of the luminance and saturation components show that the majority of the values lie around 1 (see Figure 6.4(c) and Fig. 6.4(d)). The histograms drop quickly towards lower and higher luminance and saturation values and exhibit a skew towards high luminance and saturation values. Hence, statistical models which describe the effects of lighting changes on the illumination images can be infered from the variation of their intensity and their color saturation. A probability distribution which approximates well the histograms of the luminance and the saturation of illumination images in linear domain is the gamma distribution. Using the N intensity values x_Y and the N saturation values x_S of the illumination images computed from all M interpolated virtual images at a given viewpoint of the dashed image sequence in Figure 6.1, the parameters of the gamma distributions

$$p(x_k) = \frac{b_k^{a_k}}{\Gamma(a_k)} \cdot x_k^{a_k-1} \cdot \exp\left(-b_k x_k\right) \tag{6.16}$$

can be infered by Maximum Likelihood estimation. The index $k = \{Y, S\}$ represents either the luminance or saturation component and $\Gamma(\,\cdot\,)$ is the gamma function in (2.10). The inference of the parameters a_k and b_k is done as shown in Section 2.1.1.

Changes in the environment which are not caused by lighting but e.g. by moved, new or removed objects can be detected by checking for regions in the illumination image of a new camera image whose luminance and color saturation values form outliers with respect to the infered statistical models in (6.16) (see Section 6.3). However, it is also high luminance and saturation values from specularities as well as low luminance values from dark shadows

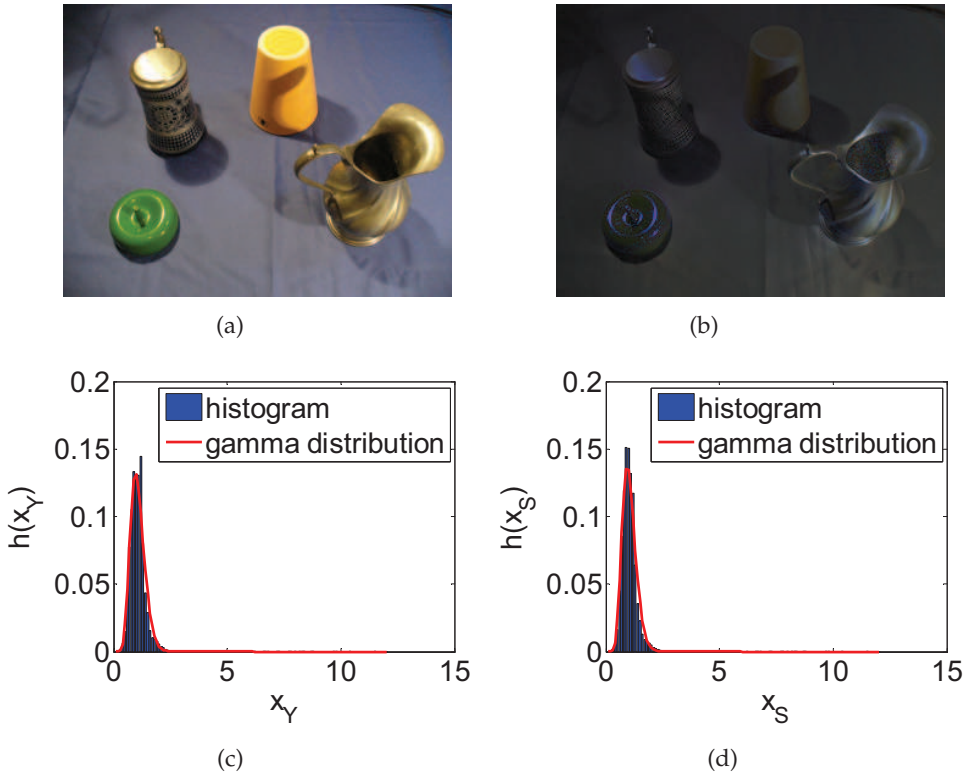


Figure 6.4: *(a) An image taken of a scene under certain lighting conditions. (b) The illumination of the scene. (c) Histogram of the luminance components of 9 illumination images. (d) Histogram of the saturation components of 9 illumination images.*

which have a low probability density in (6.16). Specularities and shadows depend on the geometric properties of the object surfaces and have similar shape under similar positions of the light source. Hence, their shapes can be recognized in the illumination image of a new camera image and the novelty values in the corresponding regions can be suppressed. Thus, after learning the parameters of the gamma distributions in (6.16) from the luminance and saturation values of all pixels of the illumination images, the binary shapes of the specularity and shadow regions are extracted from the illumination images used for learning the gamma models.

Since specularities on gray surfaces under white light do not lead to higher color saturation in the illumination image but only to higher luminance values, only the luminance component of the illumination image is used to detect specularities and strong shadows. To this end, the information of each luminance value in the illumination images is measured. The information of the luminance at a pixel (u, v) in the M illumination images is given by

$$i_Y(u, v) = -\log P\left(x_Y(u, v)\right), \tag{6.17}$$

where $P\left(x_Y(u, v)\right)$ is the probability of the luminance value $x_Y(u, v)$. The luminance values of the digital illumination images are quantized and thus discrete. The probability of a luminance value is approximated by the product of the probability density of the luminance value in (6.16) and the width of the quantization interval.

Next, the algorithm searches for regions which encompass information values which lie above the threshold

$$T_Y = \frac{\overline{i_Y}}{1-\kappa}, \tag{6.18}$$

where $\overline{i_Y}$ is the average information over all illumination images. The information maps are binarized by setting all values above this threshold to 1 and all values below to 0. Images containing the binary shapes are stored at each viewpoint in the virtual image sequence in Figure 6.1 together with the parameters a_Y, b_Y, a_S and b_S. This data forms part of the environment representation.

6.3 Detection of novel objects under varying illumination in environments with specular surfaces

When a new camera image I_λ^{new} is taken, an illumination image L_λ^{new} can be computed using an illumination-invariant image R_λ interpolated from the environment representation at the current position of the robot's camera

$$L_\lambda^{\text{new}} = \frac{I_\lambda^{\text{new}}}{R_\lambda} = \frac{k_{\text{d}}^{\text{new}} O_{\text{d}\lambda}^{\text{new}} \cdot \left[I_{\text{p}\lambda} \cdot \vec{N} \cdot \vec{L} + \frac{I_{\text{p}\lambda} k_{\text{s}}}{k_{\text{d}}^{\text{new}} O_{\text{d}\lambda}^{\text{new}}} \left(\vec{R} \cdot \vec{V} \right)^n \right]}{k_{\text{d}} O_{\text{d}\lambda}}. \tag{6.19}$$

In regions where no specularities are exhibited ($k_{\text{s}} \to 0$ or $\vec{R} \cdot \vec{V} \to 0$) the illumination image in (6.19) results in

$$L_\lambda^{\text{new}} = \frac{k_{\text{d}}^{\text{new}} O_{\text{d}\lambda}^{\text{new}}}{k_{\text{d}} O_{\text{d}\lambda}} \cdot I_{\text{p}\lambda} \cdot \vec{N} \cdot \vec{L}. \tag{6.20}$$

Thus, the illumination image does not only exhibit a modulation pattern of the light source intensity but also indicates changes in the diffuse spectral reflectance of the surfaces in the environment. Such a change can happen, if an object has been moved, added or removed whose intrinsic color differs from the intrinsic color of the background. Then, the corresponding region in the illumination image contains, as in case of specularities and shadows, luminance and color saturation values which are outliers in the illumination models in (6.16) and provide high information values $i_Y(u,v)$ and $i_S(u,v)$. For a robust detection of the objects, however, illumination effects have to be identified and their information values have to be attenuated.

In this section, a technique is proposed which is based on matching binary shapes to identify illumination effects in a new camera image. As already mentioned in Section 6.2, both the shapes of specularities and the shapes of shadows depend on the geometry of the object surfaces in the environment and thus have similar shapes under similar lighting conditions. Furthermore, specularities can only be present on objects with specular surfaces, not in the rest of the environment. Hence, if a binary shape is extracted in a region of the new illumination image, the algorithm looks for a similar shape extracted in the corresponding region of one of the illumination images during training.

For shape matching descriptors based on Zernike moments, as proposed in [KH90], are used. The Zernike moment of order n with repetition m for an image region $f(u,v)$ inside

the unit circle is computed as

$$A_{nm} = \frac{n+1}{\pi} \sum_u \sum_v f(u,v) V^*_{nm}(u,v), \quad u^2 + v^2 \leq 1 \tag{6.21}$$

where $V_{nm}(u,v)$ is the complex Zernike polynomial given by

$$V_{nm}(u,v) = R_{nm}(u,v) \cdot \exp\{jm\theta\}. \tag{6.22}$$

θ is the angle between the vector $(u,v)^{\mathrm{T}}$ and the u-axis. The radial polynomial $R_{nm}(u,v)$ is computed by

$$R_{nm}(u,v) = \sum_{s=0}^{n-\frac{|m|}{2}} (-1)^s \cdot \frac{(n-s)!}{s! \cdot \left(\frac{n+|m|}{2} - s\right)! \cdot \left(\frac{n-|m|}{2} - s\right)!} \cdot \left(u^2+v^2\right)^{\frac{n}{2}-s}. \tag{6.23}$$

The asterisk in (6.21) denotes the complex conjugate of the Zernike polynomial in (6.22). The Zernike polynomials have to be computed only once at the beginning of the algorithm since they are independent of the image data.

The magnitudes of the Zernike moments up to order $n = 120$ are summarized in a vector and build the descriptor $\mathbf{d}$.

The steps of the algorithm proposed in this chapter for the detection of novel objects in presence of specular surfaces in the environment are:

1. Compute the illumination image from a new camera image and an illumination-invariant image interpolated from the illumination-invariant image-based environment representation described in Section 6.1.2.
2. Identify regions in the shading image with information values $i_Y(u,v) > T_Y$. For this, the parameters of the gamma distributions are also interpolated from nearby reference viewpoints.
3. For each region $\mathcal{R}'$:
   a) Compute the intersection BB_{I} of the bounding box of $\mathcal{R}'$ and the bounding box of a shape $\mathcal{R}_{\mathrm{M}}$ stored in the representation. The binary images containing the shapes $\mathcal{R}_{\mathrm{M}}$ are interpolated from the representation at the viewpoint where the new image is taken.
   b) If $BB_{\mathrm{I}} \neq \emptyset$, compute the Zernike descriptors $\mathbf{d}'_{\mathrm{I}}$ and $\mathbf{d}_{\mathrm{M,I}}$ of the partial regions $\mathcal{R}'_{\mathrm{I}}$ and $\mathcal{R}_{\mathrm{M,I}}$ which lie inside BB_{I}.
   c) Compute the distance of the two shapes using the Euclidean distance $e = \left\|\mathbf{d}'_{\mathrm{I}} - \mathbf{d}_{\mathrm{M,I}}\right\|$.
   d) Store with the pixel region covered by $\mathcal{R}'_{\mathrm{I}}$ the minimum distance e obtained from the region $\mathcal{R}_{\mathrm{M,I}}$ in the environment representation which matches the region $\mathcal{R}'_{\mathrm{I}}$ best.
4. Compute for each pixel (u,v) in the new camera image the novelty measure

$$\Delta(u,v) = \frac{e(u,v)}{\hat{e}} \cdot \sqrt{i_Y^2(u,v) + i_S^2(u,v)}. \tag{6.24}$$

   $\hat{e}$ denotes the maximum shape distance across all pixels.

Hence, the better the shape of a region with high information matches a shape from the representation the more the novelty measure is suppressed.

6.4 Experimental results

In the following, experimental results obtained from image data sets acquired by two mobile robot platforms are shown. The experiments cover the detection of novel changes in the environment due to new or removed objects both in Lambertian environments and in environments with specular surfaces. The variable κ in (6.18) is set to $\kappa = 0.5$ in Section 6.4.1 and to $\kappa = 0.2$ in Section 6.4.2.

6.4.1 Detection of novel changes in Lambertian environments

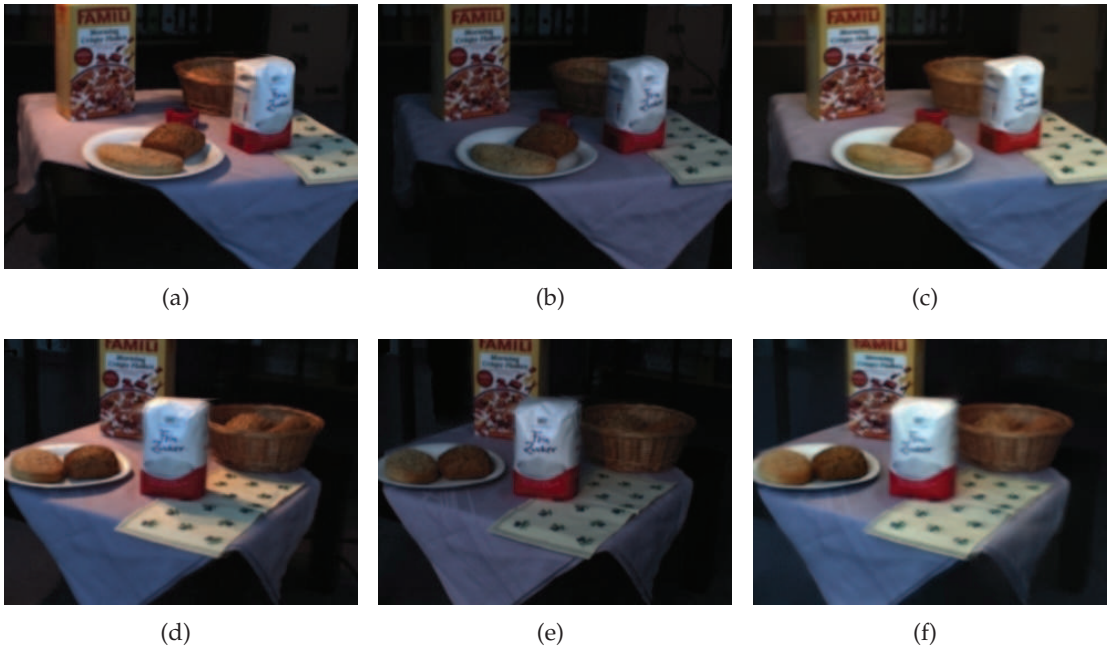
In a first experiment, nine image sequences $\mathcal{I}_m$ ($m = 1, \ldots, 9$) are acquired of a small scene, using the Pioneer 3-DX robot [mob] in Figure 3.5. The scene consists of several objects like a sugar box, a cereals box, both made of paper, a basket with rolls, a paper plate with rolls and a small red tape dispenser. The objects are on a table covered with diffuse reflecting table cloth (see Figure 6.6(c)). Thus, the surfaces in the scene are largely Lambertian. Each sequence consists of 100 stereo image pairs. The robot is controlled to follow a quarter-circle with the stereo camera looking towards the center of the circle, in a similar way as described in the experiments in Section 3.4. Due to the inaccurate internal odometry of the robot, however, the trajectories are not perfect quarter-circles. Furthermore, the robot is manually steered to the starting point of the next trajectory before the acquisition of a new image sequence. Consequently, the trajectories are similar but never completely identical.

The first image sequence $\mathcal{I}_1$ is captured under outdoor daylight which falls through a window of the laboratory behind the robot. Next, two sequences $\mathcal{I}_2$ and $\mathcal{I}_3$ are acquired under indoor illumination. Here, the lamps mounted on the ceiling of the laboratory, which provide white light, are used. For the remaining six image sequences $\mathcal{I}_4$ to $\mathcal{I}_9$ the scene is illuminated by daylight and by the light of two lamps which are mounted on a tripod and which are placed at various positions around the scene. The lamps provide yellow-white light so that not only the positions but also the spectra of the light sources vary between the different runs.

The reference images for the interpolation of virtual images are provided by the robot's left camera. Hence, for each left image in the nine sequences a depth map is computed, using the method described in Section 3.2. The pose of the robot's left camera is estimated during the acquisition of the images, using the image-based approach in Section 3.1.1. Therefore, the image sequences have to be registered to a common coordinate system, as described in Section 6.1.1.

Figures 6.5(a), 6.5(b), 6.5(d) and 6.5(e) show four virtual images rendered at different viewpoints from the image sequences $\mathcal{I}_4$ and $\mathcal{I}_5$. These virtual images are photorealistic and hardly exhibit artifacts due to erroneous poses of the reference images or depth maps. The illumination-invariant images recovered from all nine virtual images at the two viewpoints are shown in Figures 6.5(c) and 6.5(f). Obviously, the shadows in the virtual images cast by the objects on the table are largely gone in the illumination-invariant images.

To evaluate the method for illumination-invariant novelty detection in Section 6.3 another image sequence $\mathcal{I}_{\mathrm{ob}}$ is acquired. The two lamps are positioned here in a way that the illumination was different from all the runs before. Furthermore, one roll is removed from the plate. One image, which is depicted in Figure 6.6(a), is chosen as the robot's observation for



(a) (b) (c)

(d) (e) (f)

Figure 6.5: *(a),(d): Virtual images rendered at different viewpoints from images acquired under dominant artificial illumination. (b),(e): Virtual images rendered at the same viewpoints from images taken under dominant daylight illumination. (c),(f): Reflectance images recovered from all virtual images at these viewpoints. The illumination effects which are visible in (a),(b),(d) and (e) are largely removed.*

(a) (b) (c)

Figure 6.6: *Table scene: (a) Observation which is taken after one roll has been removed from the plate. (b) Illumination-invariant image of the scene. (c) Reference image for NGC and SCT.*

the experiments. The corresponding virtual illumination-invariant image rendered at the observation's viewpoint is shown in Figure 6.6(b).

The proposed information-theoretic novelty measure in Section 6.3, is shown in Figure 6.7(a). It clearly indicates a region of high novelty around the missing roll (values between 60 and 120) while the shadow regions have very low values (up to 20). Elevated values along the edges of the objects and in the background are due to a slight inaccuracy in the

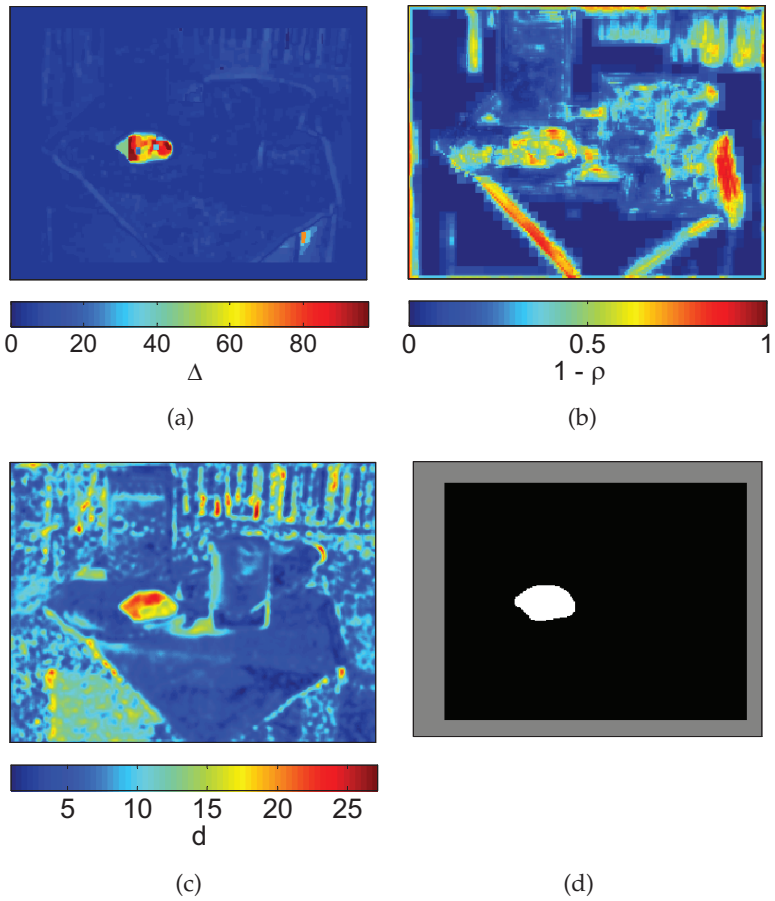


Figure 6.7: *(a) Novelty measure, as proposed in Section 6.3, (b) Normalized Gradient Correlation ($1 - \rho$), (c) Spherical Coordinate Transform, (d) Ground truth.*

estimation of the observation's pose.

Two change detection methods based on Normalised Gradient Correlation (NGC) ([OH07]) and the Spherical Coordinate Transform (SCT) ([SS07], [MMHM01]) are compared to the method in Section 6.3. Both reference methods are outlined in Section 2.4. The results are shown in Figures 6.7(b) and 6.7(c). Both for NGC and SCT a virtual image rendered from the image sequence I_4 is chosen as a reference image (see Figure 6.6(c)). High correlation coefficients ρ obtained by NGC indicate blocks with no changes while a change is likely if the coefficient is low (see Section 2.4). For coherent visualization the measure $1 - \rho$ is chosen. As proposed in [OH07], the gradient correlation coefficients are computed on three resolution layers and combined for the change measure. Compared to the method in Section 6.3, NGC is much more sensitive to pose inaccuracies and clearly indicates object edges as regions of high novelty since there the gradient structure between blocks from the observation and the corresponding blocks from the reference image is different. The method based on the SCT reliably detects the missing roll but also exhibits comparably high chromaticity differences d in the rest of the change map.

In a second experiment the method for novelty detection in Section 6.3 is evaluated for an-

(a) (b) (c)

Figure 6.8: *Washing machine scene: (a) Observation which shows the hair dryer as a new added object. (b) Illumination-invariant image of the scene. (c) Reference image for NGC and SCT.*

other scene which is shown in Figure 6.8. In contrast to the previous table scene all images of the washing machine scene are taken at a static viewpoint while the illumination changes. The method in Section 6.3 is again compared to the two reference approaches NGC and SCT. Figure 6.9(a) shows the novelty map obtained from the proposed method. Again, it provides high novelty values in the region of the observation image that exhibits the hair dryer that is added to the scene as a new object. Remarkably, the algorithm is able to distinguish between a black object and a shadow, which is difficult for the other two methods. The change from the white color of the washing machine to the deep black of the hair dryer leads to an attenuation of the intensity that is not typical for a shadow. That is why it is classified as a reflectance change in the proposed scheme. The result of the NGC shows that the edges of the shadows produce high values in the map in Figure 6.9(b). Due to the faint chromaticity change the values in the change map in Figure 6.9(c) in the region of the hair dryer are not higher than at many other pixels.

For a quantitive comparison of the three methods, the receiver operating characteristic curves are computed. For the table scene the false positive rate is evaluated and plotted vs. the true positive rate, using the ground truth in Figure 6.7(d), which is determined manually. The gray regions in the ground truth map are excluded from the evaluation since there the virtual illumination-invariant image does not provide any information. In Figures 6.10(a) and 6.10(b) the ROC curves are obtained by varying a threshold between the minimum and the maximum value of the maps in Figures 6.7(a), 6.7(b) and 6.7(c). Pixels which contain novelty values above the threshold and lie within the white region in Figure 6.7(d) provide true positives whereas pixel whose novelty values lie below that threshold in that area provide false negatives. Likewise, pixels in the novelty maps which contain values above the threshold and lie within the black region in Figure 6.7(d) provide false positives whereas pixel whose novelty values lie below that threshold in the black area provide true negatives. The TPRs are computed as given in (5.8) and the FPRs are determined as given in (5.9). The ROC curve for the method in Section 6.3 shows higher TPRs at low FPRs smaller than 0.2 than the ROC curves for the other methods. Analyzing the ROC curves in Figure 6.10(b) that are obtained for the washing machine scene it is shown that the one for the method in Section 6.3 almost shows ideal behavior while the perfomance of the other methods is clearly worse.

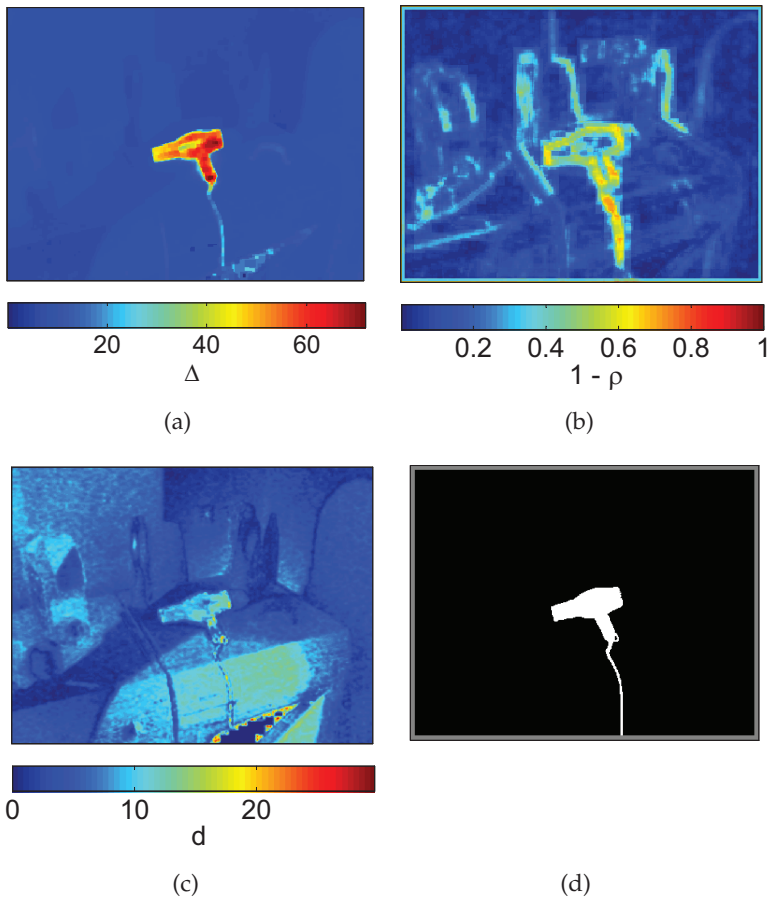


Figure 6.9: *(a) Novelty measure, as proposed in Section 6.3, (b) Normalized Gradient Correlation $(1 - \rho)$, (c) Spherical Coordinate Transform, (d) Ground truth.*

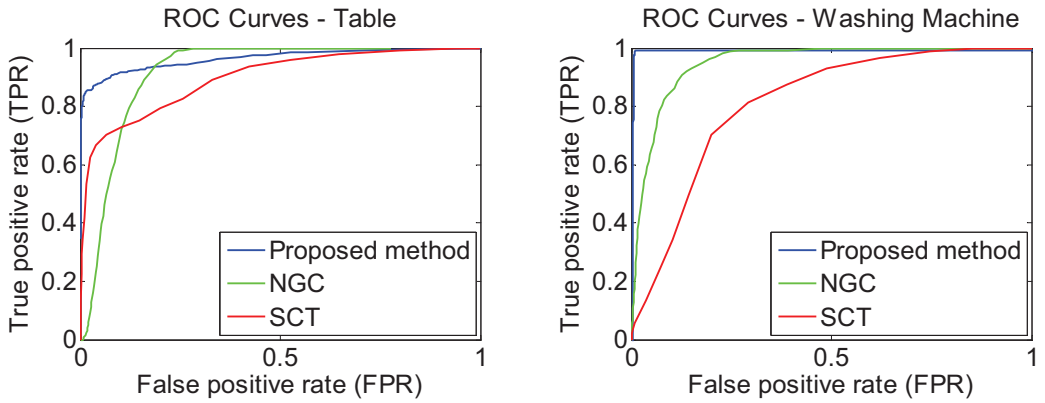
6.4.2 Detection of novel changes in environments with specular surfaces

For further experimental validation of the approach in Section 6.3 two experiments are conducted in environments which contain specular surfaces.

In the first experiment eight image sequences $\mathcal{I}_m$ $(m = 1, ..., 8)$ are acquired of a scene using the mobile robot platform employed in the experiments in Section 4.3. The robot is controlled to move along a trajectory, which is similar to the quarter circle between the points Q_1 and Q_2 depicted in Figure 4.3(b). In each run the scene is illuminated by two lamps mounted on a tripod whose position is changed around the scene from run to run. The setup is shown in Figure 6.11 [2]. The scene consists of several objects on a table covered with a textured table cloth. Among the objects there are a tin plate with a highly specular surface, an object made of translucent glass and several colored objects.

For each image in the eight image sequences which is captured by the center camera of

[2]The robot platform in Figure 6.11 is part of the CoTeSys Central Robotics Laboratory (CCRL), which is supported within the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see also `www.cotesys.org`.



(a) (b)

Figure 6.10: *ROC curves for the table scene (a) and the washing machine data set (b). They show the performance of the three approaches in terms of the true positive rate vs. the false positive rate.*

(a) (b) (c)

Figure 6.11: *The acquisition of multiple image sequences by a mobile robot platform while the position of the lamps is changed.*

the Bumblebee® XB3 a depth map is computed as described in Section 3.2. The 6D pose of the Bumblebee® XB3 is estimated for each captured image with respect to a common world coordinate system, using the optical tracking system in Section 3.1.2. Hence, a registration of the image sequences as described in Section 6.1.1 and as required in the experiments in Section 6.4.1 is not necessary.

Before the acquisition of a new sequence of 70 images a white cup is added to the scene. Figure 6.12(a) shows frame 20 of the image sequence and Figure 6.12(b) shows the corresponding illumination-invariant image interpolated from the environment representation.

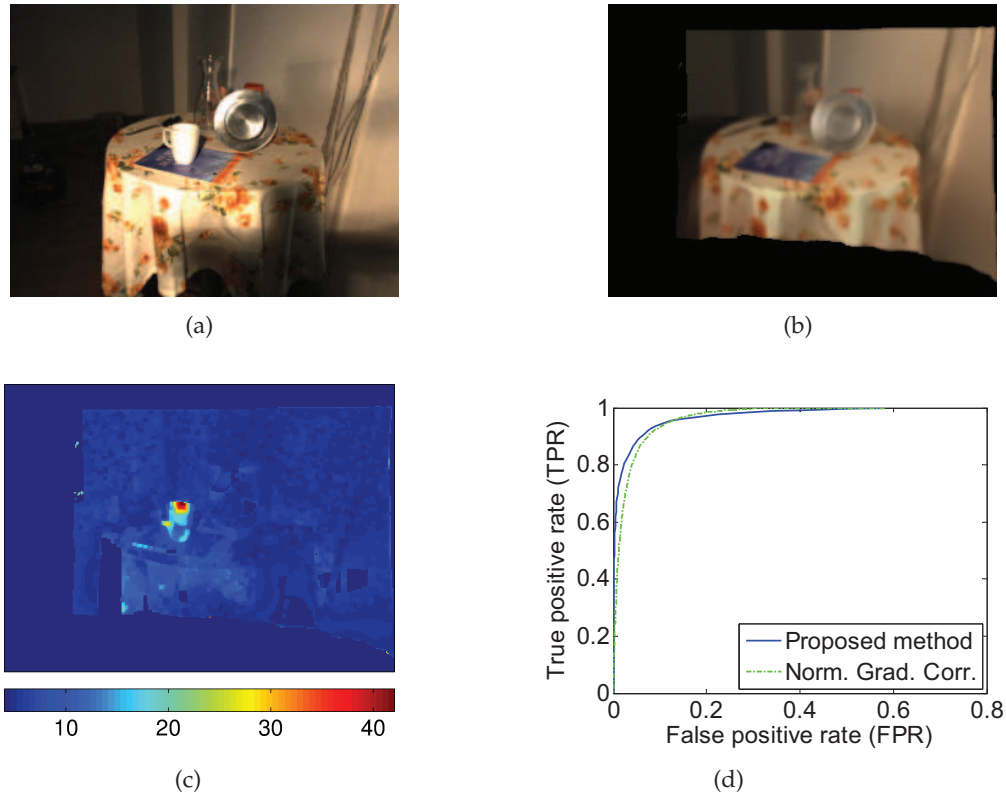


Figure 6.12: *(a) Image taken of the scene with the new object. (b) Interpolated illumination-invariant image. (c) Novelty map. (d) ROC curves.*

Despite the strong reflections on the tin plate and the shadow which the robot casts on the scene near the lower right image corner, the values in the novelty map in Figure 6.12(c) are on average higher for the region of the new cup than for the rest of the image. For a quantitative analysis the region covered by the new white cup is labeled in all 70 frames. The ROC curves in Figure 6.12(d) are obtained by sweeping a threshold through the value range of the novelty maps of all 70 frames. For high FPRs the performance of the approach in Section 6.3 is similar to the performance of the normalised gradient correlation [OH07]. At FPRs smaller than 10%, however, the approach in Section 6.3 also provides higher TPRs than the NGC method.

In a second experiment the approach in Section 6.3 is validated with a series of images which are acquired using a camera mounted on a tripod at a single viewpoint. 16 images are taken while the position of a light source is varied half around the scene. A white object is placed into the scene and six new images are taken with the position of the light source changing. One of these images is shown in Figure 6.13(a). The novelty map in Figure 6.13(c) clearly indicates the novel object in the scene while illumination effects are not exhibited. In the region of the strong specularity on the plate the novelty map shows very low values. This is because its shape matches very well a shape which has been extracted from the 16 illumination images before. Considering the ROC curves in Figure 6.13(d), which are obtained by evaluating all six new images, the NGC algorithm shows the highest TPRs if high FPRs are

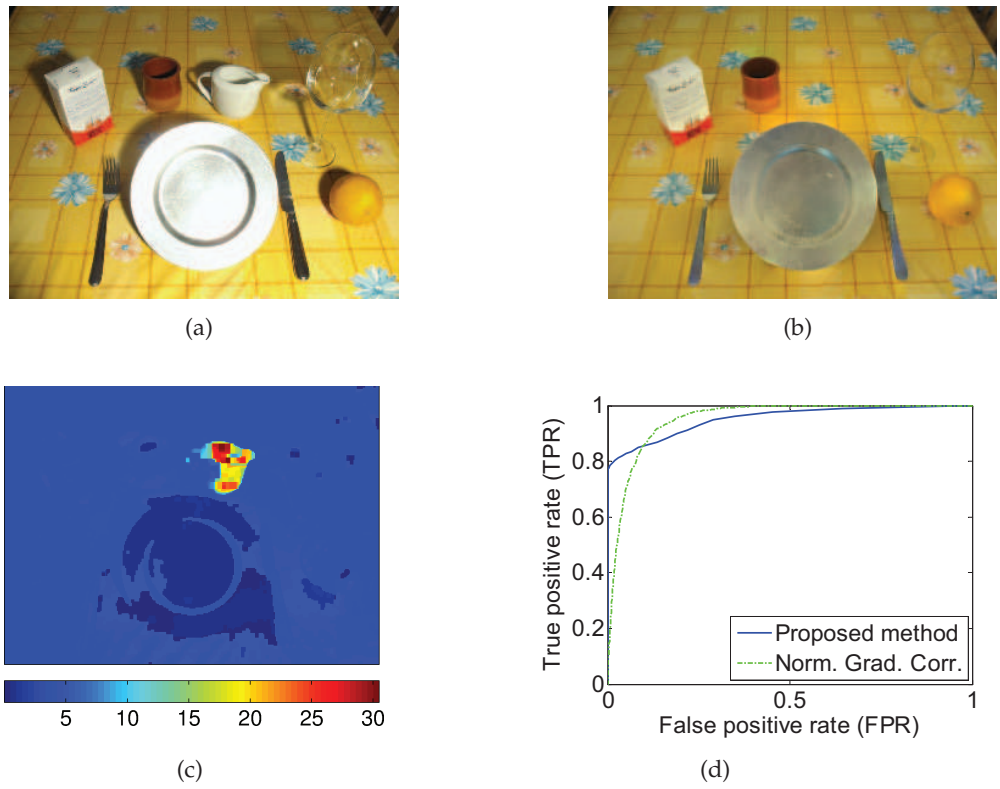


Figure 6.13: *(a) Image taken of a scene under certain lighting conditions. (b) Illumination-invariant image. (c) Novelty map. (c) ROC curves.*

permitted. This is because the unicolored object strongly changes the gradient structure of the table surface with the line and flower pattern. However, below a FPR of 7% the method in Section 6.3 again outperforms the NGC algorithm.

6.5 Discussion

This section points out the limitations of the approach presented in this chapter for the image-based detection of novel objects in a robot's environment under varying illumination conditions. Furthermore, the insights gained from the experimental results are discussed.

Limitations

Illumination effects which are static and therefore visible in all image sequences acquired during the training phase will also be visible in the illumination-invariant image. Regions of the environment which are never properly illuminated by the light source and therefore always exhibit shadows with undefined edges in the images are a typical example for these illumination effects, as well as specularities on surfaces with a small radius, which are constant for different positions of the light source. Similarly, if the spectrum of the light source

is narrowbanded and the light source emits light of a given color during the training phase, the illumination-invariant image will be tinted in that color. However, illumination effects visible in the illumination-invariant image will not impair the detection of novel objects if these effects will also be visible in images the robot captures after the training phase. Only if the light source is positioned in a way that a shadow region of the illumination-invariant image is illuminated in a new image or if the spectrum of the light source changes, novelty will be measured. If the spectrum of the light source changes during the training phase, the variance of the gamma distribution for the color saturation in (6.16) is increased, which means that the novelty measure might be lower when colored objects are detected.

During the training the scene has to be static. This condition can be relaxed to a certain degree, as long as scene changes are visible only in a small minority of the image sequences. Since the median is in general robust against outliers, the gradients of moving, new or disappeared objects in the scene are ignored in (6.8) and (6.9), if they are exhibited only in one or two image sequences and if the total number of image sequences is large.

New changes in the environment which cause similar intensity and color saturation variations in illumination images as shadows do cannot be detected by the method proposed in this chapter. This can happen in rare cases, e.g. if a gray object is placed on a white surface.

Furthermore, the detection of novel objects and the suppression of specularities depends on the robustness of the shape matching technique in Section 6.3 and the choice of the threshold in (6.18). If a novel object is erroneously classified as a specularity (in case of very simple object shapes), its novelty values are attenuated. Besides, if a specularity is not recognized in a new illumination image, the FPR in novelty detection is increased.

Comparative evaluation

Change detection methods based on NGC and SCT are very fast and do not require any training. Compared to the approach proposed in this chapter, the acquisition of multiple images and the inference of statistical models describing the intensity and saturation variation caused by the illumination in the environment is not necessary.

However, if NGC is applied, small errors in the registration of images to a common coordinate frame increase the number of false positives, since the edges of static objects are not aligned in the images and thus the gradient structures of corresponding blocks in the images are different. Registration errors also affect the method for novelty detection proposed in this chapter. However, elevated novelty values along a thin line tracing an object edge can be eliminated by morphological filtering operations and are not as disturbing as whole blocks of false positives as they might be provided by NGC. Hence, NGC performs well if the camera is static as in surveillance applications. It is less suitable, however, for mobile robot applications since slight pose inaccuracies are inevitable. Furthermore, NGC produces false positives along shadow borders, since they provide intensity gradients which are not exhibited in the reference image. In contrast, the performance of the approach proposed in this chapter is not affected by intensity discontinuities caused by illumination effects.

Using SCT, changes can only be detected, if the hue and/or color saturation values between corresponding regions in two images are different. Mere intensity changes, as caused e.g. by a black object in front of a white background, are not indicated if SCT is used. In contrast, the method proposed in this chapter does indicate intensity changes caused by new, moved

or removed objects if they are atypical for shadows.

Besides, neither change detection methods based on NGC nor methods based on SCT propose mechanisms to handle and suppress specularities. In the experiments of this chapter it is shown that the attenuation of the novelty values from specularities is effective, using the proposed approach.

6.6 Summary

This chapter presents an approach for the detection of novel changes in the environment of a mobile robot. The approach is image-based and robust against illumination changes. Novelty detection is based on an illumination-invariant image-based environment representation, which the robot computes from multiple image sequences which are taken of the scene under different illumination conditions. Using the illumination-invariant image at a given viewpoint, illumination effects are extracted from the captured camera images and represented in illumination images. The addition or removal of an object from the environment can be detected in the illumination image computed from a new camera image and an illumination-invariant image interpolated at the current viewpoint from the environment representation. In general, the illumination image exhibits colors in that region which are not typical.

One of the main contributions of this chapter is a statistical model which describes the variation of the intensity and the color saturation in illumination images. Since illumination images largely resemble gray-scale images and may contain sparse regions of pronounced color, the statistics of their values are well described by gamma models for the intensity and saturation component in HSV color space. Image regions which exhibit the addition or removal of an object usually provide values which are outliers in terms of the infered gamma distributions. Another main contribution of this chapter is the suppression of specularities and strong shadows during novelty detection, which also form outliers. An approach is proposed which finds matches between the shapes of outlier regions extracted from a new illumination image and shapes of specularities and shadows which have been stored in memory during the training phase.

The experiments in this chapter analyze the performance of the proposed approach using both data sets acquired by a static camera and data sets acquired by a camera on a moving robot platform. The results show a robust detection of scene changes while illumination effects are largely suppressed.

A research item in future work is to develop a method to keep the illumination-invariant image-based representation up to date if objects are added, moved or removed from the environment.

7 Conclusion

In this thesis, two novel types of appearance-based environment representations for cognitive mobile robots are investigated. These environment representations facilitate the detection of novel events in the robot's surroundings, while illumination changes can be identified and suppressed.

Probabilistic appearance representation

The probabilistic appearance representation proposed in this thesis represents the luminance and chrominance of a 3D environment in terms of prior distributions at densely spaced viewpoints. The parameters of the priors are learned in a Bayesian approach from images which the robot takes along its way through the environment. Depth information and camera pose data are stored at each viewpoint and facilitate the interpolation of the priors at intermediate viewpoints. The probabilistic description of luminance and chrominance enables the robot to reason about the uncertainty of the appearance which results from contradictory luminance and chrominance values acquired for a scene point in the past. If the uncertainty is low, the robot is able to make predictions of the environment's appearance which resemble a photorealistic virtual image.

The statistical appearance representation also provides a method for the assessment of the level of surprise of new captured luminance and chrominance values. Surprise quantifies how much a new luminance or chrominance sample changes the corresponding prior distribution stored in the environment model. The level of surprise is usually low in parts of the environment which often change. In turn, surprise is high if a luminance or chrominance change is perceived after a series of similar and consistent luminance or chrominance samples.

Experimental results show that the surprise maps computed by the proposed method clearly indicate novel scene changes like objects which are added or removed. It is shown that the Bayesian approach for surprise detection is superior to a more simpler method which detects scene changes by calculating the difference between a camera image and an image rendered from an image-based representation of the environment. Further experiments show that the Bayesian surprise measure still detects new objects if the density of the reference views in the representation is reduced.

Surprise detection is applied in this thesis to the acquisition of feature-based object representations in cluttered environments. When a new object is present in the scene, a region of high surprise values is exhibited in the surprise map. Inside this region local image features are extracted and added to a database, while the robot moves around the object. Experimental results show that the learned objects can be reliably recognized in front of a different cluttered background.

Illumination-invariant image-based representation

In parallel to the probabilistic appearance representation an illumination-invariant image-based representation of the environment is acquired. While the probabilistic appearance representation stores the momentary statistics of the luminance and chrominance of the 3D environment captured in the past, the illumination-invariant representation holds an appearance-based model of the static part of the environment which is free of lighting effects. The illumination-invariant environment model is computed from multiple image sequences taken of the scene under different illumination conditions over time.

Using the illumination-invariant images, the illumination effects are extracted from the acquired camera images and statistical models for the variation of luminance and color saturation in the illumination images are infered. The addition or removal of objects in the scene can be detected in the illumination image extracted from a new camera image, by searching for luminance and saturation values which form outliers with respect to the infered statistical models.

A shape matching technique allows for the identification of strong specularities and shadows in new camera images. Since the shape of a specularity is subject to the shape and curvature of the object surface as well as the position of the light source, a similar shape is expected to be found on a static object of the background in a new captured image if the lighting is similar to one the training cases. Specularity and shadow regions are suppressed in a novelty map so that the attention of the robot is not distracted.

In experiments the performance of the algorithms is analyzed using both images acquired by a static camera and image data acquired by a camera on a mobile robot platform. It is shown that the proposed algorithms detect the addition and removal of objects while illumination effects are largely ignored. The technique is able to distinguish the addition of a black object from a shadow and the addition of a white object from a specularity. In comparison to state-of-the-art techniques, the proposed method does not require that the environment be colored and textured and can cope with shadow borders and specularities.

Outlook

For further research work it would be interesting to extend the approach for surprise detection to more abstract knowledge representations of the environment. The probabilistic appearance representation presented in this thesis stores knowledge of the luminance and chrominance of the scene and provides a very low-level description of the environment based on early visual cues. This representation can support the acquisition of higher-level knowledge representations. One step towards this direction is shown in this thesis in terms of the acquisition of feature-based object representations. However, many planning algorithms require symbolic representations of the world which are much more abstract. Thus, the goal would be to develop a surprise framework which works on a multi-level knowledge representation of the world.

In terms of the surprise-driven acquisition of object representations the algorithm currently only considers the addition of new unknown objects to the familiar environment. The case that objects of the background might be removed, which also causes surprise and triggers a feature learning process, is not considered. However, the algorithm can be easily extended to cope with this situation. By comparing the mean depth of the region which provokes high

surprise values to the corresponding mean depth predicted from the environment model, the algorithm can decide if a new object has been added to the scene or if an object has been removed.

Another open issue related to the illumination-invariant image-based representation proposed in this thesis is how the representation can be updated when a new object has been detected or when an object has disappeared. The challenge here is to transfer the new appearance of the scene from one captured image to the illumination-invariant representation without illumination effects like the shading of object surfaces.

Bibliography

Publications by the author

[BMS08] I. Bauermann, W. Maier, and E. Steinbach. Progressive rendering from RDTC optimized streams. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pages 1169–1172, 2008.

[MBM+10] W. Maier, F. Bao, E. Mair, E. Steinbach, and D. Burschka. Illumination-invariant image-based novelty detection in a cognitive mobile robot's environment. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 5029–5034, 2010.

[MBS+09] W. Maier, F. Bao, E. Steinbach, E. Mair, and D. Burschka. Illumination-invariant image-based environment representations for cognitive mobile robots using intrinsic images. In *Proc. Vision, Modeling, and Visualization Workshop (VMV)*, pages 379–380, 2009.

[MES11] W. Maier, M. Eschey, and E. Steinbach. Image-based object detection under varying illumination in environments with specular surfaces. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 1417–1420, 2011.

[MMBS09] W. Maier, E. Mair, D. Burschka, and E. Steinbach. Visual homing and surprise detection for cognitive mobile robots using image-based environment representations. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 807–812, 2009.

[MS10] W. Maier and E. Steinbach. A probabilistic appearance representation and its application to surprise detection in cognitive robots. *IEEE Transactions on Autonomous Mental Development (TAMD)*, 2(4):267–281, 2010.

[MS11] W. Maier and E. Steinbach. Surprise-driven acquisition of visual object representations for cognitive mobile robots. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 1621–1626, 2011.

General publications

[AB91] E. H. Adelson and J. R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, pages 3–20, 1991.

[ADMT01] T. Aach, L. Dümbgen, R. Mester, and D. Toth. Bayesian illumination-invariant motion detection. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 640–643, 2001.

[AHB87] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 9(5):698–700, 1987.

[AKB08] M. Agrawal, K. Konolige, and M. R. Blas. CenSurE: Center surround extremas for realtime feature detection and matching. In *Proc. European Conference on Computer Vision (ECCV)*, pages 102–115, 2008.

[AS97] S. Avidan and A. Shashua. Novel view synthesis in tensor space. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1034–1040, 1997.

[BBM+01] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured lumigraph rendering. In *Proc. 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 425–432, 2001.

[BETG08] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2008.

[BH04] D. Burschka and G. D. Hager. V-GPS(SLAM): Vision-based inertial system for mobile robots. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 409–415, 2004.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Spring Street, NY, 2006.

[BK11] M. Begum and F. Karray. Visual attention for robotic cognition: A survey. *IEEE Transactions on Autonomous Mental Development*, 3(1):92–105, 2011.

[BNJ03] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3(7):993–1022, 2003.

[Bra00] G. Bradski. The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 25(11):120, 122–125, 2000.

[BS00] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley and Sons, Third Avenue, NY, 2000.

[BS08a] I. Bauermann and E. Steinbach. RDTC optimized compression of image-based scene representations (part I): Modeling and theoretical analysis. *IEEE Transactions on Image Processing*, 17(5):709–723, 2008.

[BS08b] I. Bauermann and E. Steinbach. RDTC optimized compression of image-based scene representations (part II): Practical coding. *IEEE Transactions on Image Processing*, 17(5):724–736, 2008.

[BT78] H. G. Barrow and J. M. Tenenbaum. Recovering intrinsic scene characteristics from images. In *Computer Vision Systems*, pages 3–26. Academic Press, New York, 1978.

[CGZ08] H. Y. Chong, S. J. Gortler, and T. Zickler. A perception-based color space for illumination-invariant image processing. *ACM Transactions on Graphics (TOG)*, 27(3), 2008.

[CH67] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[CK99] P. Chang and J. Krumm. Object recognition with color cooccurrence histograms. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 498–504, 1999.

[Col96] R. T. Collins. A space-sweep approach to true multi-image matching. In *Proc.*

IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pages 358–363, 1996.

[CTCS00] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum. Plenoptic sampling. In *Proc. 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 307–318, 2000.

[CW69] S. C. Choi and R. Wette. Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics*, 11(4):683–690, 1969.

[CW93] S. Chen and L. Williams. View interpolation for image synthesis. In *Proc. 20th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 279–288, 1993.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, Third Avenue, NY, 2001.

[dOFD+09] H. E. M. den Ouden, K. J. Friston, N. D. Daw, A. R. McIntosh, and K. E. Stephan. A dual role for prediction error in associative learning. *Cerebral Cortex*, 19(5):1175–1185, 2009.

[DTM96] P. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry-based and image-based approach. In *Proc. 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 11–20, 1996.

[DW11] D. G. Dansereau and S. B. Williams. Seabed modeling and distractor extraction for mobile AUVs using light field filtering. In *Proc. International Conference on Robotics and Automation (ICRA)*, pages 1634–1639, 2011.

[DYB98] P. Debevec, Y. Yu, and G. Borshukov. Efficient view-dependent image-based rendering with projective texture-mapping. In *Proc. Eurographics Workshop on Rendering*, pages 105–116, 1998.

[ESK03] J.-F. Evers-Senne and R. Koch. Image based interactive rendering with view dependent geometry. In *Proc. Computer Graphics Forum*, pages 573–582, 2003.

[ESNK06] J.-F. Evers-Senne, A. Niemann, and R. Koch. Visual reconstruction using geometry guided photo consistency. In *Proc. Vision, Modeling and Visualization (VMV)*, pages 57–64, 2006.

[Eva09] C. Evans. Notes on the OpenSURF library. Tech. Rep. CSTR-09-001, University of Bristol, 2009.

[EZW97] S. Engel, X. Zhang, and B. Wandell. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, 388(6637):68–71, 1997.

[FAS+01] P. C. Fletcher, J. M. Anderson, D. R. Shanks, R. Honey, T. A. Carpenter, T. Donovan, N. Papadakis, and E. T. Bullmore. Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature Neuroscience*, 4(10):1043–1048, 2001.

[Fau93] O. Faugeras. *Three-dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, Cambridge, MA, 1993.

[FBBC97] B. Funt, K. Bamard, M. Brockington, and V. Cardei. Luminance-based multi-

scale retinex. In *Proc. Congress of the International Colour Association (AIC Colour)*, 1997.

[FDL04] G. D. Finlayson, M. S. Drew, and C. Lu. Intrinsic images by entropy minimization. In *Proc. European Conference on Computer Vision (ECCV)*, pages 582–595, 2004.

[FH06] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision (IJCV)*, 70(1):41–54, 2006.

[FHLD06] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28(1):59–68, 2006.

[FV97] W. T. Freeman and P. A. Viola. Bayesian model of surface perception. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 787–793, 1997.

[FvDFH96] J. D. Foley, A. v. Dam, S. K. Feiner, and J. F. Hughes. *Computer Graphics: Principles and Practice, Second Edition in C*. Addison-Wesley, Boston, MA, 1996.

[GBG+94] H. Greenspan, S. Belongie, R. Goodman, P. Perona, S. Rakshit, and C. H. Anderson. Overcomplete steerable pyramid filters and rotation invariance. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 222–228, 1994.

[GGSC96] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. In *Proc. 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 43–54, 1996.

[HKP+99] B. Heigl, R. Koch, M. Pollefeys, J. Denzler, and L. van Gool. Plenoptic modeling and rendering from image sequences taken by a hand-held camera. In *Proc. 21. Symposium f. Mustererkennung (DAGM)*, pages 596–603, 1999.

[Hor02] G. Horstmann. Evidence for attentional capture by a surprising color singleton in visual search. *Psychological Science*, 13(6):499–505, 2002.

[HW02] X. Huang and J. Weng. Novelty and reinforcement learning in the value system of developmental robots. In *Proc. 2nd Int. Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pages 47–55, 2002.

[HWP10] A. Hendel, D. Weinshall, and S. Peleg. Identifying surprising events in video using Bayesian topic models. In *Proc. Asian Conference on Computer Vision (ACCV)*, pages 448–459, 2010.

[HZ03] R. Hartley and A. Zisserman. *Multiple View Geometry*. Cambridge University Press, The Edinburgh Building, Cambridge, UK, 2003.

[IB05] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 631–637, 2005.

[IB09] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009.

[IK01] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for

rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–1259, 1998.

[Jen01] H. W. Jensen. *Realistic Image Synthesis Using Photon Mapping*. AK Peters, South Avenue, Natick, MA, 2001.

[KH90] A. Khotanzad and Y. H. Hong. Invariant image recognition by Zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 12(5):489–497, 1990.

[KS04] S. B. Kang and R. Szeliski. Extracting view-dependent depth maps from a collection of images. *International Journal of Computer Vision (IJCV)*, 58(2):139–163, 2004.

[Kul59] S. Kullback. *Information Theory and Statistics*. John Wiley and Sons, Third Avenue, NY, 1959.

[LB05] M. D. Levine and J. Bhattacharyya. Detecting and removing specularities in facial images. *Computer Vision and Image Understanding (CVIU)*, 100(3):330–356, 2005.

[Len98] J. Lengyel. The convergence of graphics and vision. *IEEE Computer*, 31(7):46–53, 1998.

[LF94] S. Laveau and O. D. Faugeras. 3-D scene representation as a collection of images. In *Proc. International Conference on Pattern Recognition (ICPR)*, pages 689–691, 1994.

[LGW+04] W. Lai, X.-D. Gu, R.-H. Wang, W.-Y. Ma, and H.-J. Zhang. A content-based bit allocation model for video streaming. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pages 1315–1318, 2004.

[LH96] M. Levoy and P. Hanrahan. Light field rendering. In *Proc. 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 31–42, 1996.

[Li99] Z. Li. Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proc. of the National Academy of Sciences of the United States of America (PNAS)*, 96(18):10530–10535, 1999.

[LK81] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, 1981.

[LM71] E. H. Land and J. J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61(1):1–11, 1971.

[Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LZW04] A. Levin, A. Zomet, and Y. Weiss. Separating reflections from a single image using local features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 306–313, 2004.

[Mah36] P. C. Mahalanobis. On the generalised distance in statistics. *Proc. of the National Institute of Science, India*, 2(1):49–55, 1936.

[MB95] L. McMillan and G. Bishop. Plenoptic modeling: An image-based rendering

system. In *Proc. 22nd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 39–46, 1995.

[MGZK09] H. J. Müller, T. Geyer, M. Zehetleitner, and J. Krummenacher. Attentional capture by salient color singleton distractors is modulated by top-down dimensional set. *Journal of Experimental Psychology: Human Perception and Performance*, 35(1):1–16, 2009.

[MMHM01] B. W. Minten, R. R. Murphy, J. Hyams, and M. Micire. Low-order-complexity vision-based docking. *IEEE Transactions on Robotics and Automation*, 17(6):922–930, 2001.

[MN99] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley and Sons, Third Avenue, NY, 1999.

[MS03] M. Markou and S. Singh. Novelty detection: A review – part 1: Statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003.

[MSSB09] E. Mair, K. H. Strobl, M. Suppa, and D. Burschka. Efficient camera-based pose estimation for real-time applications. In *Proc. IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 2696–2703, 2009.

[NC06] G. Neo and F. K. Chua. Capturing focused attention. *Perception and Psychophysics*, 68(8):1286–1296, 2006.

[OF96] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[OH07] R. O'Callaghan and T. Haga. Robust change-detection by normalised gradient-correlation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[OKH07] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Trans. on Evolutionary Computation*, 11(2):265–286, 2007.

[ORP00] N. M. Oliver, B. Rosario, and A. P. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000.

[PCD+97] K. Pulli, M. Cohen, T. Duchamp, H. Hoppe, L. Shapiro, and W. Stuetzle. View-based rendering: Visualizing real objects from scanned range and color data. In *Proc. Eurographics Workshop on Rendering*, pages 23–34, 1997.

[Pho75] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975.

[Pla00] M. Planck. Zur Theorie des Gesetzes der Energieverteilung im Normalspektrum. *Verhandlungen der Deutschen Physikalischen Gesellschaft*, 2(17):245, 1900.

[PvGV+04] M. Pollefeys, L. van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision (IJCV)*, 59(3):207–232, 2004.

[RAAKR05] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005.

[RD09] A. Ranganathan and F. Dellaert. Bayesian surprise and landmark detection. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 2017–2023, 2009.

[RRK+10] M. Rambow, F. Rohrmüller, O. Kourakos, D. Brščić, D. Wollherr, S. Hirche, and M. Buss. A framework for information distribution, task execution and decision making in multi-robot systems. *IEICE Transactions on Information and Systems*, E93-D(6):1352–1360, 2010.

[RTF+04] R. Raskar, K.-H. Tan, R. Feris, J. Yu, and M. Turk. Non-photorealistic camera: Depth edge detection and stylized rendering using multi-flash imaging. *ACM Transactions on Graphics (TOG)*, 23(3):679–688, 2004.

[Sch05] J. Schmidhuber. Self-motivated development through rewards for predictor errors / improvements. In *Proc. AAAI Spring Symposium on Developmental Robotics*, 2005.

[Sch10] J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

[SCK07] H.-Y. Shum, S.-C. Chan, and S. B. Kang. *Image-Based Rendering*. Springer, Spring Street, NY, 2007.

[Ser82] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, New York, 1982.

[SGW+07] J. J. Steil, M. Götting, H. Wersing, E. Körner, and H. Ritter. Adaptive scene dependent filters for segmentation and online learning of visual objects. *Neurocomputing*, 70(7-9):1235–1246, 2007.

[SH99] H.-Y. Shum and L.-W. He. Rendering with concentric mosaics. In *Proc. 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 299–306, 1999.

[SH03] P. Sajda and F. Han. Perceptual salience as novelty detection in cortical pinwheel space. In *Proc. First International IEEE EMBS Conference on Neural Engineering*, pages 43–46, 2003.

[Sim97] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *Proc. IEEE Asilomar Conference on Signals, Systems and Computers*, pages 673–678, 1997.

[SJ00] G. Schaufler and H. W. Jensen. Ray tracing point sampled geometry. In *Proc. Rendering Techniques 2000: 11th Eurographics Workshop on Rendering*, pages 319–328, 2000.

[SKC03] H.-Y. Shum, S. B. Kang, and S.-C. Chan. Survey of image-based representations and compression techniques. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(11):1020–1037, 2003.

[SKS+02] R. Swaminathan, S. B. Kang, R. Szeliski, A. Criminisi, and S. K. Nayar. On the motion and appearance of specularities in image sequences. In *Proc. European Conference on Computer Vision (ECCV)*, pages 508–523, 2002.

[SLBS10] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82, 2010.

[SLL+07] O. Stasse, D. Larlus, B. Lagarde, A. Escande, F. Saidi, A. Kheddar, K. Yokoi, and F. Jurie. Towards autonomous object reconstruction for visual search by the humanoid robot HRP-2. In *Proc. IEEE-RAS International Conference on Humanoid Robots*, pages 151–158, 2007.

[SMD+08] A. Smolic, K. Müller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand. Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 2448–2451, 2008.

[SRE+05] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 370–377, 2005.

[SS03] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I–195–I–202, 2003.

[SS07] M. Sridharan and P. Stone. Structure-based color learning on a mobile robot under changing illumination. *Autonomous Robots*, 23(3):161–182, 2007.

[SS09] M. Sridharan and P. Stone. Color learning and illumination invariance on mobile robots: A survey. *Robotics and Autonomous Systems*, 57(6-7):629–644, 2009.

[SSH95] J. Schmidhuber, J. Storck, and J. Hochreiter. Reinforcement driven information acquisition in non-deterministic environments. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN)*, pages 159–164, 1995.

[ST94] J. Shi and C. Tomasi. Good features to track. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[STT+11] N. Shroff, Y. Taguchi, O. Tuzel, A. Veeraraghavan, S. Ramalingam, and H. Okuda. Finding a needle in a specular haystack. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 5963–5970, 2011.

[TFA05] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(9):1459–1472, 2005.

[TG80] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[The92] J. Theeuwes. Perceptual selectivity for color and form. *Perception and Psychophysics*, 51(6):599–606, 1992.

[TKS03] Y. Tsin, S. B. Kang, and R. Szeliski. Stereo matching with reflections and translucency. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I–702 – I–709, 2003.

[TM02] D. S. Taubman and M. W. Marcellin. *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Kluwer, Norwell, MA, 2002.

[TN06] K. Takahashi and T. Naemura. Layered light-field rendering with focus measurement. *Signal Processing: Image Communication*, 21(6):519–530, 2006.

[TTN08] Y. Taguchi, K. Takahashi, and T. Naemura. Real-time all-in-focus video-based rendering using a network camera array. In *Proc. IEEE 3DTV-Conference*, pages 241–244, 2008.

[Wei01] Y. Weiss. Deriving intrinsic images from image sequences. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 68–75, 2001.

[Whi80] T. Whitted. An improved illumination model for shaded display. *Communications of the ACM*, 23(6):343–349, 1980.

[WIS+10] K. Welke, J. Issac, D. Schiebener, T. Asfour, and R. Dillmann. Autonomous acquisition of visual multi-view object representations for object recognition on a humanoid robot. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 2012–2019, 2010.

[XPKB09] T. Xu, T. Pototschnig, K. Kühnlenz, and M. Buss. A high-speed multi-GPU implementation of bottom-up attention using CUDA. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 41–47, 2009.

[XRB04] B. Xie, V. Ramesh, and T. Boult. Sudden illumination change detection using order consistency. *Image and Vision Computing*, 22(2):117–125, 2004.

[ZKU+04] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics (TOG)*, 23(3):600–608, 2004.

Cited websites

[kin] *Kinect for Xbox 360*. `http://www.xbox.com/kinect`

[mba] *Monterey Bay Aquarium Research Institute (MBARI)*. `http://www.mbari.org/aved/`

[mob] *Adept MobileRobots*. `http://www.mobilerobots.com/researchrobots/researchrobots/pioneerp3dx.aspx`

[pgb] *Point Grey Bumblebee XB3*. `http://www.ptgrey.com/products/bbxb3/bumblebeeXB3_stereo_camera.asp`

[pti] *PhoeniX Technologies Incorporated (PTI)*. `http://www.ptiphoenix.com/VZmodels.php`

[SSF+] STROBL, K. H. ; SEPP, W. ; FUCHS, S. ; PAREDES, C. ; ARBTER, K.: *DLR CalDe and DLR CalLab*. `http://www.robotic.dlr.de/callab/`

[sur] *Video: Surprise Detection in Cognitive Mobile Robots*. `http://www.lmt.ei.tum.de/videos/surprise.php`