A Probabilistic Appearance Representation and Its Application to Surprise Detection in Cognitive Robots

Werner Maier, *Member, IEEE*, and Eckehard Steinbach, *Senior Member, IEEE*

***Abstract*—In this work, we present a novel probabilistic appearance representation and describe its application to surprise detection in the context of cognitive mobile robots. The luminance and chrominance of the environment are modeled by Gaussian distributions which are determined from the robot's observations using Bayesian inference. The parameters of the prior distributions over the mean and the precision of the Gaussian models are stored at a dense series of viewpoints along the robot's trajectory. Our probabilistic representation provides us with the expected appearance of the environment and enables the robot to reason about the uncertainty of the perceived luminance and chrominance. Hence, our representation provides a framework for the detection of surprising events, which facilitates attentional selection. In our experiments, we compare the proposed approach with surprise detection based on image differencing. We show that our surprise measure is a superior detector for novelty estimation compared to the measure provided by image differencing.**

***Index Terms*—Attention, cognitive robots, image-based representations, surprise.**

I. Introduction

COGNITIVE robots plan motion sequences and actions based on an internal representation of their environment. In the field of simultaneous localization and mapping (SLAM), various approaches have been proposed that enable mobile robots to autonomously build a topological map of their environment [1]. To this end, the robots typically fuse data from multiple laser scans or depth maps into a consistent global geometric model and acquire knowledge of the objects' distances. Geometric information about the world is useful for grasping task-relevant objects or to avoid obstacles during navigation.

However, the information about the robot's environment contained in purely geometric models is not sufficient. Without storing information about the appearance of the world, a robot would be unable to remember if the (identically shaped) cups that it just put on the table were yellow or green. Besides, it would not be possible to distinguish the blue box in the kitchen cupboard from the other equally sized and shaped boxes. Hence, color information is a crucial component of the robot's internal representation.

Texture mapping techniques from the field of computer graphics [2] have shown that realistic visualizations of structured environments can be achieved by associating real-world image data with the 3-D vertices of the geometry model. In order to capture some of the view-dependent appearance changes, a sparse set of images (view-dependent textures) can be stored along with the global geometric model.

The performance of traditional texture mapping heavily depends on the level of detail and on the material properties of the objects. In complex real-world environments, and that is where robots usually act, it is difficult to achieve high-quality results. For the visualization of glasses and other translucent objects, information about the refraction of light has to be included in the rendering process in order to achieve acceptable results. To this end, raytracing techniques [3] were developed which provide high quality results but suffer from high computational complexity. For cognitive robots, which have to make decisions in real-time and rely on rapid information retrieval from the internal model, this is not acceptable.

Image-based object or scene representations provide an interesting alternative to raytracing techniques since they allow for a realistic visualization of the environment while the computational complexity is independent of the structure of the scene. Hence, it is possible to render virtual images of scenes containing translucent and specular objects several times a second. In order to capture an image-based scene representation, a dense set of images is acquired by a rigid camera array or a mobile platform. Novel virtual views are generated by transferring and combining pixel data from nearby reference views. The amount of images that has to be acquired for realistic and artifact-free modeling depends on the accuracy of the geometry information that is used for the color transfer [4]. The more accurate the local geometry data, the smaller the number of images that have to be stored in the representation. Hence, view-dependent geometry represented by local per-pixel depth maps, which are stored along with the images, has become very popular [5].

Since cognitive robots act in dynamic environments, their internal environment representation has to be updated continuously. In order to handle the vast amount of data that is acquired by its sensors, the robot has to filter the information and focus on stimuli which are particularly relevant for task selection. Hence, a mechanism which controls the robot's attention is required. In

Manuscript received January 28, 2010; revised July 29, 2010; accepted September 13, 2010. Date of publication September 27, 2010; date of current version December 10, 2010. This work was supported in part by the DFG Excellence Initiative Research Cluster *Cognition for Technical Systems—CoTeSys*, see also www.cotesys.org.

The authors are with the Institute for Media Technology, Technische Universität München, 80333 München, Germany (e-mail: werner.maier@tum.de; eckehard.steinbach@tum.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAMD.2010.2080272

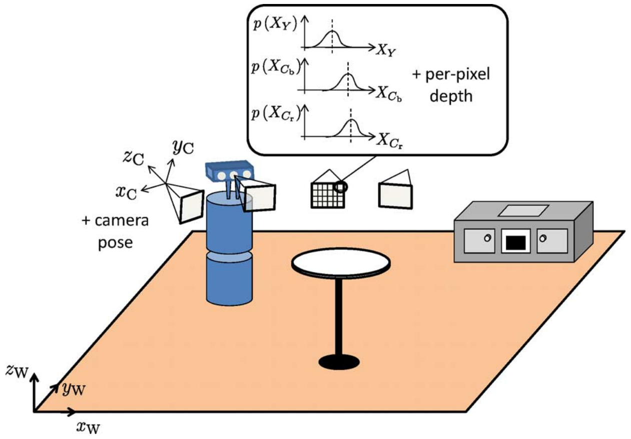


Fig. 1. Our proposed appearance representation uses Gaussian models for the luminance and the chrominance of the environment at each pixel at a viewpoint. The Gaussian distributions are infered from observations along the robot's trajectory. The representation also includes a depth map and the pose of the robot's camera head for each viewpoint.

the past, a number of methods have been proposed for the detection of novelty [6]. Furthermore, saliency models have been presented [7] in order to predict positions in an image that attract human gaze. Recently, Itti *et al.* found that, compared to other information-theoretic and saliency measures, Bayesian surprise is the strongest attractor of human attention [8].

Hence, surprise can provide an efficient means to direct the robot's attention to regions in the environment that contain objects the robot has not seen before. Once the object is segmented from the background, the robot can extract features, store them in a database and use them for recognizing it at a later time. During navigation the robot is rapidly informed about new obstacles that are not contained in the internal representation. Besides, surprise is a good detector for unexpected human motion.

In [9], we proposed a method for surprise detection that is based on image-based representations. We extend our work in [9] in two important directions. First, compared to [9] we consider dynamic scenes and second, we add a measure of uncertainty of the acquired color values to the robot's internal representation. More specifically, this work presents a novel probabilistic appearance representation that, similar to image-based models, consists of a dense series of views. However, our proposed representation does not store the raw observations of the robot but infers the parameters of probability distributions which treat luminance and chrominance values as random variables (see Fig. 1). Compared to traditional image-based models this enables a cognitive robot to identify regions with uncertain color values in the currently observed image and to distinguish between static and frequently changing objects. Furthermore, we show in this work that the proposed representation provides a framework for the computation of Bayesian surprise.

Although our algorithms have not been developed for a specific cognitive architecture, we believe that they could be used for preattentive vision in the perception modules of architectures like ACT-R [10] and ICARUS [11]. Our surprise maps indicate the position of novel objects and thus could provide input to the visual buffer which is associated with the dorsal “where“ path of ACT-R's visual system. On the other hand, the selective extraction of features from novel objects which is shown in this work can facilitate the autonomous formation of higher-level object representations. Using these object representations, the robot can recognize familiar objects in the scene and create entities associated with them in the visual buffer of ACT-R or in the perceptual buffer of the ICARUS architecture.

The paper is structured as follows. In Section II, we review related work. Section III presents the novel probabilistic appearance representation which we propose for cognitive mobile robots. The section also describes our methods for camera localization, depth estimation and view interpolation. The computation of surprise using our probabilistic representation is presented in Section IV. Section V shows experimental results obtained from a real-world scenario. In Section VI, we conclude this work.

II. Related Work

In [12], the huge body of work in the field of image-based rendering (IBR) is reviewed. IBR methods are classified with respect to the amount of geometry information which is included in the representation and used for rendering. A system which computes novel images using explicit view-dependent geometry in terms of per-pixel depth maps is presented in [5]. Since the positions of the cameras capturing the scene are static, only one initial calibration step is required, without further pose estimation during the acquisition process.

In [13], an image collection is acquired by a mobile robot during exploration and organized in a link graph. Using this representation, the robot is able to localize itself in the environment. This work investigates, in particular, the application of this representation to path planning and navigation.

Itti *et al.* found in [8] that surprise attracts human attention more than saliency and information-theoretic measures like the image entropy. They use a statistical model for the firing rates of early visual neurons and learn probability distributions over its parameters from observations using Bayesian inference. The Kullback–Leibler divergence of the posterior and the prior distribution provides a quantitative measure for surprise. Neuroscientific experiments in [14] also show that there are areas in the primary visual cortex and putamen which respond progressively more to unpredicted and progressively less to predicted visual stimuli. It was found that surprise is an important cue for associative learning [15].

In [16], 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 [8], 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 [17], [18]. Intrinsic motivation, which results from the pursuit of maximum internal reward, can be driven by learning progress [19] or by novelty and surprise [20], 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 in our work. Experiments are often performed in an artificial gridworld.

The concept of surprise is also used in [21] where a robot automatically detects new landmarks during exploration and creates a topological map which contains points of special interest like gateways.

The main contributions of this paper are the following:

- a probabilistic representation which expresses the robot's expectation and uncertainty about the appearance of its dynamically changing 3-D environment in terms of belief distributions over luminance and chrominance;
- a method for the computation of surprise maps which is based on the probabilistic concept of this representation.

In contrast to our work, the image-based rendering techniques reviewed in this section only store momentary snapshots of the scene. While [8] investigates low-level surprise in humans, we propose a method for surprise detection which does not require the simulation of visual neurons and which can be easily implemented on the graphics hardware of a robot. Furthermore, unlike [8], we keep track of the pose of the robot's camera and extract local geometry information. This allows us to match images which are captured at different viewpoints. In contrast to [13] and [21], we do not focus on objects and landmarks which are only useful for navigation but also include complex everyday objects like glasses in our representation.

III. A Probabilistic Appearance Representation for Cognitive Robots

A cognitive mobile robot which is equipped with cameras is able to acquire color information and to gather evidence about the appearance of the objects in its environment. By storing images from a continuous sequence together with the corresponding camera poses, the robot can build an internal appearance representation and later remember how the environment looked like at the moment of acquisition. The ability to recall and predict the appearance of the environment from a percept history enables the 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 [22], 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 robot can already filter relevant information from early stimuli before higher cognitive layers are reached.

An image taken at a given time instant only reflects the momentary appearance of the scene but does not tell how long the environment is 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 holds a belief in the hypothesis that the scene appears in a certain color.

In the representation that we propose in this work, the luminance and chrominance values which are captured at a single pixel for a given viewpoint are modeled by Gaussian distributions

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

We use three separate probability models for the luminance Y and chrominance components C_b and C_r since there is strong evidence that in the human visual system the luminance and chrominance information is processed in decoupled pathways [23].

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. 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 [24]. Compared to the color processing in the human visual system the C_b values encode blue–yellow opponencies and the C_r values red–green opponencies.

Like in image-based representations with explicit geometry we store a per-pixel depth map at each viewpoint and the 6-degree of freedom (6 DOF) pose of the robot's camera head with respect to a defined world coordinate system. The retrieval of the pose and the computation of the depth maps are described later in this section. Fig. 1 illustrates our proposed probabilistic appearance representation.

A. Bayesian Inference of Model Parameters

In principle, there are several ways to infer the parameters of the Gaussian distribution from the acquired luminance and chrominance samples. Within the frequentist paradigm, one could estimate the mean from the average of sequentially captured samples and the precision from the squared differences of the samples from the inferred mean. The obtained maximum-likelihood estimates are then point estimates which provide 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, only a small set of luminance and chrominance samples is acquired around a given viewpoint. Hence, 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, does not infer only one single model and discards all the others, but places prior distributions over the parameters of the probability distribution. The prior distributions can be interpreted as subjective beliefs in hypotheses about the correct model. They take into account that, apart from the most likely model, there might be other models that can be valid for the observed data—of course with a lower probability.

In [9], we assume that the appearance predicted from the internal representation at a virtual viewing position does not differ much from the true appearance. Thus, we suppose that the mean value of the Gaussian distribution is known and only placed a one-dimensional gamma prior over the precision. In this work, we do not make this assumption and treat both the mean and the precision of the Gaussian distribution as unknown.

One reason why we use a Gaussian model for the luminance and chrominance values at a pixel is that the Gaussian distribution belongs to the exponential family. So there exists a conjugate prior for its parameters [25]. An important property of conjugate priors is that the posterior distribution obtained by the multiplication of the prior with the likelihood function has the same functional form as the prior distribution [26]. This makes any further analysis like the comparison of posterior and prior distribution straight forward. Furthermore, the computed posterior distribution serves as a new prior during the processing of new luminance and chrominance samples.

The conjugate prior which is used for the Bayesian inference of both the mean and the precision of the Gaussian distribution under the assumption that none of them is known is the normal-gamma distribution. It has the following form:

$$p_0\left(\mu_k, \lambda_k\right) = \\ \kappa \cdot \lambda_k^{\alpha_{0,k}-1/2} \cdot \exp\left\{-\beta_{0,k}\lambda_k\right\} \cdot \exp\left\{-\frac{\lambda_k\left(\mu_k - \tau_{0,k}\right)^2}{2\sigma_{0,k}}\right\} \quad (2)$$

with

$$\kappa = \frac{\beta_{0,k}^{\alpha_{0,k}}}{\Gamma(\alpha_{0,k})\sqrt{2\pi\sigma_{0,k}}} \quad (3)$$

as a normalization factor and $k \in \{Y, C_\mathrm{b}, C_\mathrm{r}\}$ again. $\Gamma(\alpha)$ is the gamma function

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} \cdot \exp\{-u\}\,\mathrm{d}u\,. \quad (4)$$

As we can see from (2), the normal-gamma model is a two-dimensional probability distribution which is determined by its four hyperparameters $\alpha_{0,k}$, $\beta_{0,k}$, $\tau_{0,k}$, and $\sigma_{0,k}$. Thus, in order to encode the subjective belief distribution of the robot with respect to the mean and precision of the Gaussian model, it is sufficient to store these four hyperparameters for a given pixel at a viewpoint. Note that the normal-gamma distribution is not separable with respect to the random variables μ_k and λ_k. While the first exponential term in (2) only contains the precision, the second one depends on both the mean value and the precision. An example for a normal-gamma distribution in the luminance channel is shown in Fig. 2.

When the robot makes a new observation $\mathbf{X}_\mathrm{ob} = \{X_{\mathrm{ob},k}\}_{k=Y,\mathrm{C_b},\mathrm{C_r}}$ at a viewpoint, the prior distribution in (2) is turned into a posterior distribution using Bayes' formula

$$p\left(\mu_k, \lambda_k \mid X_{\mathrm{ob},k}\right) \propto p\left(X_{\mathrm{ob},k} \mid \mu_k, \lambda_k\right) \cdot p_0\left(\mu_k, \lambda_k\right). \quad (5)$$

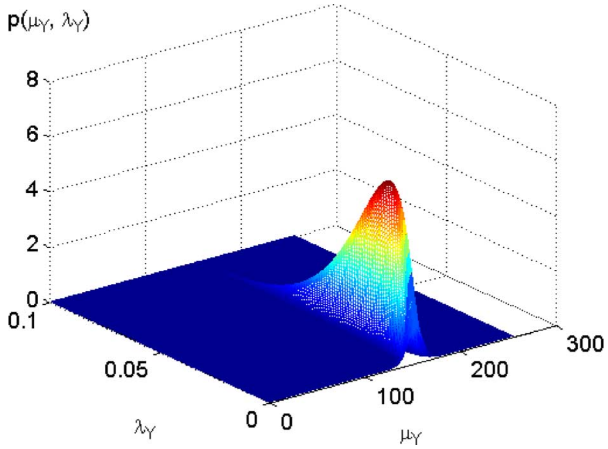


Fig. 2. An example of a normal-gamma distribution over the mean μ_Y and the precision λ_Y of the Gaussian model of the luminance channel. The expected luminance at this pixel is around 150.

Due to conjugacy, the posterior distribution in (5) is again a normal-gamma distribution with the following hyperparameters [26]

$$\alpha_k = \alpha_{0,k} + \frac{1}{2} \quad (6)$$

$$\beta_k = \beta_{0,k} + \frac{1}{2} \cdot \frac{\left(X_{\mathrm{ob},k} - \tau_{0,k}\right)^2}{\sigma_{0,k} + 1} \quad (7)$$

$$\tau_k = \frac{\sigma_{0,k} \cdot X_{\mathrm{ob},k} + \tau_{0,k}}{\sigma_{0,k} + 1} \quad (8)$$

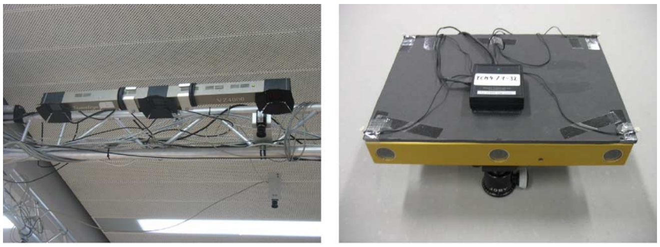
$$\sigma_k = \frac{\sigma_{0,k}}{\sigma_{0,k} + 1}. \quad (9)$$

We see from the update in (6) 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 (8) encodes the expected value of the luminance or chrominance at a pixel. Hence, the two-dimensional array of the τ_k-values is supposed to be close to a photorealistic virtual image. The parameter β_k in (7) contains a sum of squared differences between the new observed luminance or chrominance value and the corresponding expected value. The parameter σ_k in (9) determines the weight that a new luminance or chrominance value receives during the update of the expected value.

When we consider the updates in (6)–(9), 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 transferred to the internal representation which is only achieved with values for $\sigma_{0,k}$ which are significantly larger than 0.

In order to prevent the unbounded growth of α_k and β_k, a forgetting factor $f < 1$ is introduced in [27] where similar update



(a) (b)

Fig. 3. (a) One of the trackers which capture the position of the LED markers on the camera head from the laboratory's ceiling. (b) Our camera head consists of three sensors and a rectangular plate with LED markers in its corners.

equations have been derived for a gamma model as a prior over the parameter of the Poisson-distributed neural firing rates. We adapt this solution to our case. Furthermore, in order to ensure rapid adaptation to unexpected changes, the parameter $\sigma_{0,k}$ is modified before the Bayesian update as follows:

$$\sigma_{p,k} = \min\left(\max\left(\sigma_{0,k} \cdot \exp\left\{\frac{\xi}{M_{01}}\right\}, \sigma_{\min}\right), \sigma_{\max}\right) \tag{10}$$

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

$$\begin{aligned} M_{01} &= \int_{\lambda_k=-\infty}^{+\infty} \int_{\mu_k=-\infty}^{+\infty} \lambda_k \cdot p_0\left(\mu_k, \lambda_k\right) \, \mathrm{d}\mu_k \, \mathrm{d}\lambda_k \\ &= \frac{\alpha_{0,k}}{\beta_{0,k}}. \end{aligned} \tag{11}$$

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, (10) 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, which we use in our algorithm, then become

$$\alpha_k = f \cdot \alpha_{0,k} + \frac{1}{2} \tag{12}$$

$$\beta_k = f \cdot \beta_{0,k} + \frac{1}{2} \cdot \frac{\left(X_{\mathrm{ob},k} - \tau_{0,k}\right)^2}{\sigma_{p,k} + 1} \tag{13}$$

$$\tau_k = \frac{\sigma_{p,k} \cdot X_{\mathrm{ob},k} + \tau_{0,k}}{\sigma_{p,k} + 1} \tag{14}$$

$$\sigma_k = \frac{\sigma_{p,k}}{\sigma_{p,k} + 1}. \tag{15}$$

Equation (11) 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 work, we use a forgetting factor of 0.8.

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

B. Camera Localization

In order to continuously determine the current pose of the robot's camera during motion, we use multiple active-optical real-time 3-D trackers which are mounted on the ceiling of the laboratory. One of the trackers is shown in Fig. 3(a). The tracker sensor captures the position of four LED markers which are attached to the camera head of the robot and transmits the marker data to a host computer. As depicted in Fig. 3(b), the camera head consists of a three-sensor camera system, a ballhead, and a rectangular plate which contains the four LED markers in the corners. The plate is strongly attached to the camera system in a way that it is perpendicular to a camera's image plane. The markers emit infrared light in a predefined pattern and are controlled via radiocommunication.

We define a local coordinate system $x_{\mathrm{C}}y_{\mathrm{C}}z_{\mathrm{C}}$ on the plate and determine the positions of the LEDs in that coordinate frame. When the robot moves and acquires images, we permanently capture the current world coordinates of the LEDs. In order to get the current rotation of the local coordinate frame $x_{\mathrm{C}}y_{\mathrm{C}}z_{\mathrm{C}}$ with respect to the world coordinate frame $x_{\mathrm{W}}y_{\mathrm{W}}z_{\mathrm{W}}$, we use the least-squares approach in [28]. The translation $\mathbf{t}$ of the camera head with respect to the origin of the world coordinate system is provided by the tracked 3-D position of the front-left LED in Fig. 3(b).

C. Depth Estimation

At each viewpoint in Fig. 1, a dense depth map is calculated which contains the distance of each scene point to the projection center of the camera. The camera system, depicted in Fig. 3(b), captures three images simultaneously. The sensor in the middle provides the image which updates the prior distribution in (2). The image from the left camera and the image from the right camera are used for multiview stereo matching, respectively. Using two images for depth recovery allows for handling occlusions.

The computation of depth maps is done in two steps. First, a plane-sweep approach [29] tests several depth hypotheses for each pixel of the middle image and computes a matching cost by comparing its luminance value to the corresponding luminance values in the left and right image, respectively. Second, an efficient implementation of the max-product algorithm [30] then infers the most probable plane label for each pixel and provides as a result a smooth and dense depth map. The depth map is stored at each viewpoint together with the pose information and the parameters of the posterior distribution in (12)–(15).

D. 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, in order to retrieve the prior distributions for the pixels of a currently captured image, it has to interpolate the parameters of the probability model in (2) from nearby views from the internal representation. Section III-A describes 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*.

The internal representation might comprise hundreds or thousands of reference parameter images, depending on the complexity of the environment. Hence, the robot cannot keep the whole representation in working memory all the time. Furthermore, it would not make sense to use reference parameter images for view interpolation covering a part of the environment which lies completely outside the current field of view of the robot's camera. Thus, each time, the robot interpolates a new virtual parameter image, it determines a subset of seven reference views which are the closest ones in terms of distance and viewing direction.

The view interpolation is done in several passes. First, the local geometry of the environment is reconstructed in terms of triangle meshes using the depth maps of the selected reference views. The corresponding reference parameter images are mapped on the 3-D vertices. The meshes are then projected on the image plane of the virtual camera. Finally, the values $\alpha_{0,k}$, $\beta_{0,k}$, $\tau_{0,k}$, and $\sigma_{0,k}$ in the virtual parameter image are computed by averaging the warped reference parameters at each pixel. The obtained values are the parameter values of the normal-gamma prior in (2).

IV. Computation of Surprise Maps

The probabilistic appearance representation presented in Section III provides a framework for the detection of surprising events and for attentional selection. The posterior distribution obtained by (12)–(15) 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 to describe Bayesian surprise in terms of how much a new observation changes the robot's prior belief is provided by the Kullback–Leibler divergence [8]. As shown in Appendix B, the Kullback–Leibler divergence of two normal-gamma distributions can be written in a closed form. 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 $S_k,\ k \in \{Y, C_\mathrm{b}, C_\mathrm{r}\}$ at a given pixel are then computed by

$$S_k = \mathcal{KL}\left(p\left(\mu_k, \lambda_k\right); p_0\left(\mu_k, \lambda_k\right)\right) = T_1 + T_2 + T_3 + T_4 + T_5 \tag{16}$$

where

$$T_1 = \log\left(\frac{\beta_k^{\alpha_k} \cdot \Gamma(\alpha_{k,0}) \cdot \sqrt{\sigma_{k,0}}}{\beta_{k,0}^{\alpha_{k,0}} \cdot \Gamma(\alpha_k) \cdot \sqrt{\sigma_k}}\right) \tag{17}$$

$$T_2 = (\beta_{k,0} - \beta_k) \cdot \frac{\alpha_k}{\beta_k} \tag{18}$$

$$T_3 = (\alpha_k - \alpha_{k,0})\left[\psi(\alpha_k) - \log(\beta_k)\right] \tag{19}$$

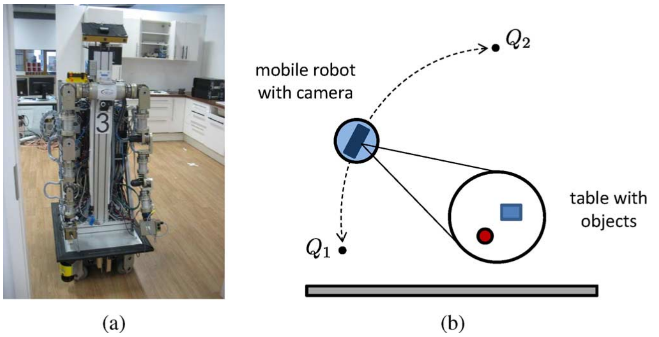


Fig. 4. (a) The mobile platform used for image acquisition [31]. (b) During the acquisition of the image sequence $\mathcal{I}_1$, the robot moves multiple times from point Q_1 to point Q_2 along an approximately circular trajectory. The trajectories between the two points are similar but never identical.

$$T_4 = \frac{1}{2\sigma_{k,0}}\left[\left(\tau_{k,0}^2 - 2\tau_{k,0}\tau_k + \tau_k^2\right) \cdot \frac{\alpha_k}{\beta_k} + \sigma_k\right] \tag{20}$$

$$T_5 = -\frac{1}{2}. \tag{21}$$

In (17) and (19), $\log(\cdot)$ denotes the natural logarithm and $\psi(\alpha_k)$ is the digamma function

$$\psi(\alpha_k) = \frac{\frac{\mathrm{d}}{\mathrm{d}x}\Gamma(x)\mid_{x=\alpha_k}}{\Gamma(\alpha_k)}\ . \tag{22}$$

The surprise values which are computed in the luminance and chrominance channels are finally combined to a total surprise score

$$S = S_Y + S_{\mathrm{C_b}} + S_{\mathrm{C_r}}\ . \tag{23}$$

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.

V. Experimental Results

In order to evaluate the computation of surprise based on our proposed probabilistic appearance representation we captured a long image sequence $\mathcal{I}_1$ with 1283 frames using the mobile platform shown in Fig. 4(a). The robot's camera acquired images at a resolution of 320 × 240 pixels and at a frame rate of 7 frames per second (fps). While the estimation of the camera head's poses was performed in real-time at 25 Hz, the computation of the depth maps required several seconds per view. Hence, for further processing, the images were saved on a hard disk and the depth maps were computed in an off-line step. The interpolation of the prior parameters, the inference of the posterior distributions according to the (12)–(15) and the computation of the surprise maps in (16) are performed by a graphics processing unit (GPU). The execution time of these steps is on average 120 ms using an NVIDIA GeForce GTX 275. Hence, our module for surprise detection can be used in real-time applications with frame rates up to 7 fps.

The robot started at point Q_1 in Fig. 4(b) and was controlled to go along a circular trajectory with its camera head looking towards the center of the circle. When it reached Q_2 it stopped and immediately went back on the circle to Q_1. Arrived at Q_1 again, it repeated the motion multiple times. At the turning points the

TABLE I
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.

robot continued the data acquisition so that the image sequence $\mathcal{I}_1$ was not interrupted. Each time it reached Q_1 the scene was changed by a human who added and removed 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 I.

Our 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 was stored. The parameters of the normal-gamma prior distribution for the inference of the first reference parameter image were 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 were infered during the first run by using the robot's observation and a prior which was 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 III-D) was 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 were replaced by the depth map and the pose at the robot's current viewpoint as well.

A. Evaluation of Bayesian Surprise Based on Our Probabilistic Appearance Representation

Figs. 5–7 illustrate three surprise maps which we computed in the phases B, C, and D, respectively. In Fig. 5(a), the image captured by the robot at frame 465 is shown. Fig. 5(b) illustrates the values of $\tau_{k,0}$ which are interpolated at the robot's viewpoint from the internal representation. We transformed $\tau_{k,0}$ to the RGB domain in order to facilitate a better comparison to the captured image. The parameters $\tau_{k,0}$ encode at each pixel the luminance and chrominance of the scene which the robot expects at its viewpoint. As depicted in Fig. 5(b), our representation enables the robot to predict a virtual image with a high realism.

The surprise map in Fig. 5(c) clearly indicates the glass as a novel object which was added to the scene at the beginning of phase B and is not contained in the internal representation. Fig. 5(d) depicts the sum of the parameters $\beta_0^\Sigma = \sum_{k\in\{Y,\mathrm{C_b},\mathrm{C_r}\}} \beta_{k,0}$ which express the uncertainty of the appearance at each pixel [see (11)]. 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

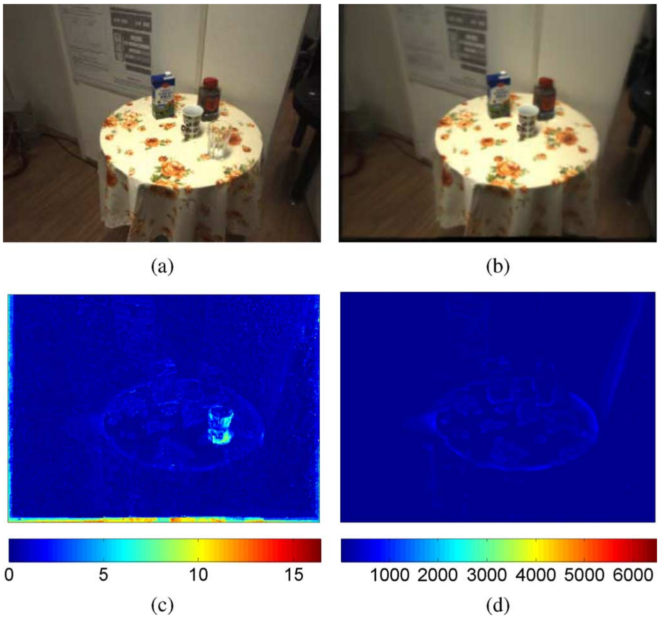


Fig. 5. (a) Frame 465 of the image sequence $\mathcal{I}_1$. (b) The parameters $\tau_{0,k}$ correspond to the robot's expected appearance. For illustration, the values of $\tau_{0,k}$ were transformed to RGB domain. (c) The surprise map indicates the glass as a novel object. (d) The parameters $\beta_{0,k}^\Sigma$ show that the robot's uncertainty about the appearance is low across the image.

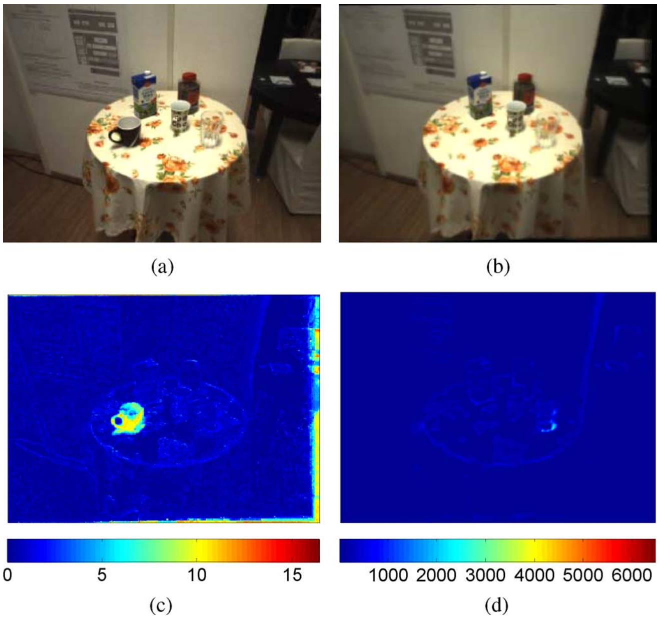
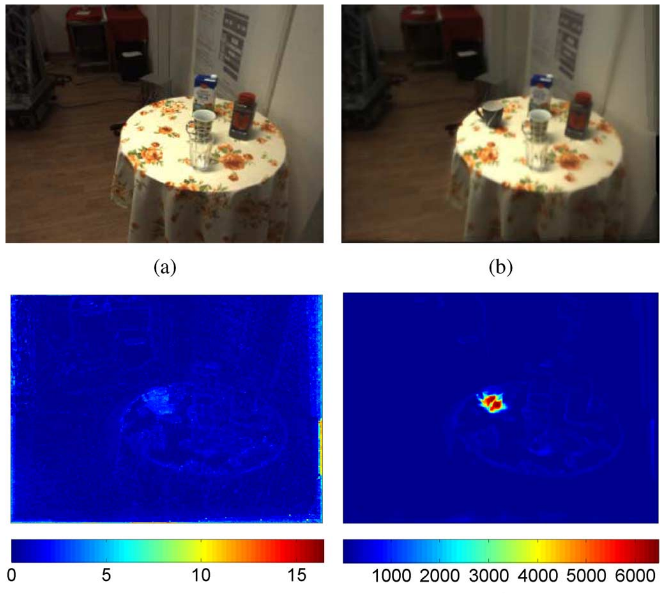
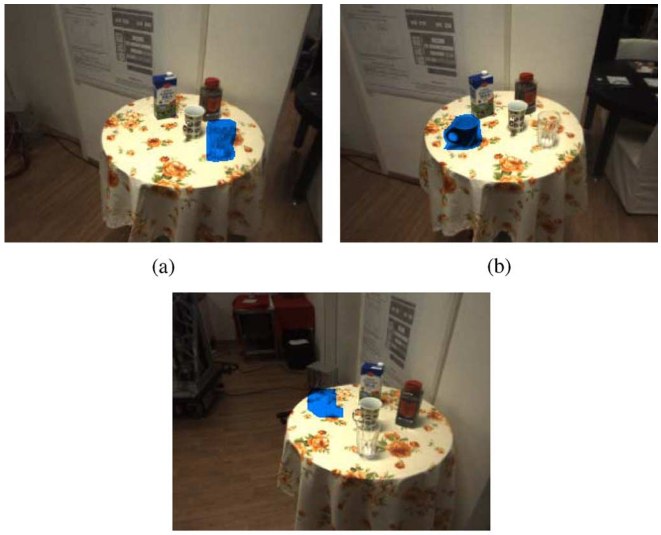


Fig. 6. (a) Frame 800 of the image sequence $\mathcal{I}_1$. (b) The parameters $\tau_{0,k}$ represent the robot's expected appearance. For illustration, the values of $\tau_{0,k}$ were transformed to RGB domain. (c) The surprise map indicates the cup as a novel object. (d) The parameters $\beta_{0,k}^\Sigma$ show that the robot's uncertainty about the appearance is low across the image.



(a) (b) (c) (d)

Fig. 7. (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 $\tau_{0,k}$ were transformed to RGB domain. (c) The surprise map shows only slightly elevated values in the region of the missing cup. (d) The parameters $\beta_{0,k}^{\Sigma}$ show a region of high uncertainty where the cup was removed.



(a) (b) (c)

Fig. 8. We manually labeled the image regions showing the glass in (a) and the cup in (b) in order to create a mask for the evaluation of the robot's surprise about these objects. In (c), we labeled the region where the cup has been in order to measure the robot's surprise about the missing cup. (a) Case II. (b) Case II. (c) Case II.

to small shifts of the object edges in the observation and the predicted image during phase A.

Fig. 6(a) shows an observation of the robot in phase C and Fig. 6(b) the appearance of the scene which the robot expects from the internal representation. The glass which was a novel object in phase B is already shown in this virtual image. The surprise map in Fig. 6(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 Fig. 7(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 Fig. 7(b), the robot is only little surprised that it has disappeared. Since the sudden appearance of the cup at the beginning of phase B 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_0^{Σ} in Fig. 7(d)]. The robot expects low luminance values but the infered Gaussian model has a small precision. That is why 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_0^{Σ} in Fig. 7(d) in the region of the glass indicates that our proposed representation is able to store the appearance of complex transparent objects with a relatively low uncertainty.

We made a quantitative evaluation of the surprise maps over the whole image sequence $\mathcal{I}_1$. For this, we manually drew polygons into the images around the regions of the glass and the cup, starting with the frame in which they appear on the table for the first time. Fig. 8 shows the blue-tinged masks. The mask in Fig. 8(c) indicates the region on the table that contained the cup before it was removed. It is used in order to measure the robot's surprise about the missing cup. In order to remove noise from the surprise maps we averaged the values over 4×4-blocks. Fig. 9(a) shows the maximum values which are measured in a block within the masks that indicate the glass and the cup, respectively. We start our measurements 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.

We start the measurement of the maximum surprise values in a block within the region of the cup at frame 675 as soon as it is on the table (orange curve). Fig. 9(a) shows that the cup evokes larger surprise values than the glass did at the beginning of phase B. This is because the stimulus difference between the bright table cloth and the dark cup is also 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 glass 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 Fig. 9(b)]. The surprise values averaged over the whole region of the glass are of course much lower than the maximum surprise values because the stimulus difference is very low at sites where the glass hardly refracts the light.

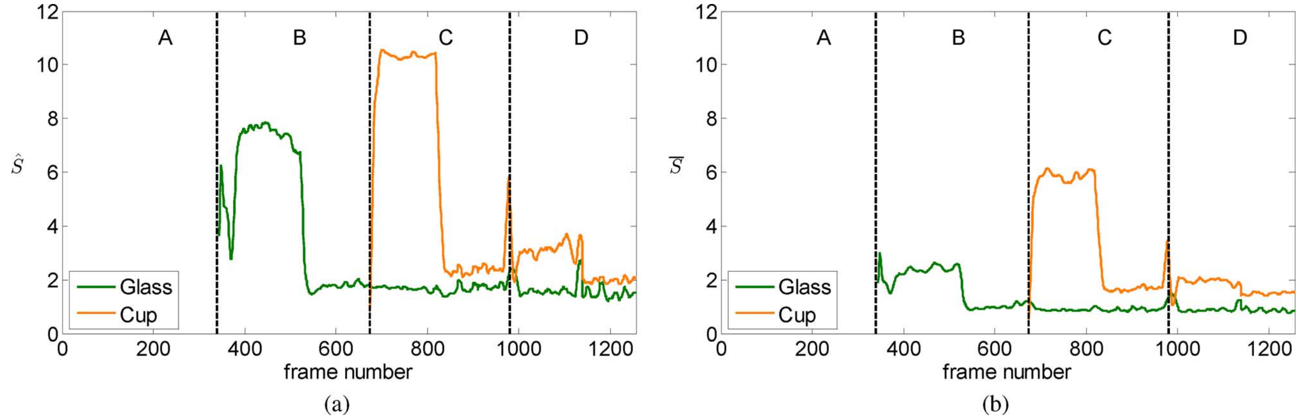


Fig. 9. (a) The maximum surprise value of a 4 × 4-block within the regions of the glass (green) and the cup (orange). (b) The average surprise value over all 4 × 4-blocks within the regions of the glass (green) and the cup (orange). In both cases, we see high values during the robot's first run from Q_1 to Q_2 after the new object has been put on the table.

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 which can be seen on http://www.lmt.ei.tum.de/videos/surprise.php.

B. 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. In order to evaluate this, we introduce a measure which we call attentional selectivity (AS) and which we calculate as follows:

$$\mathrm{AS} = 10\log_{10}\left(\frac{\overline{S}}{\hat{S}_{\mathrm{out}}}\right)\ \mathrm{dB} \tag{24}$$

where $\overline{S}$ is the surprise value averaged over all 4×4-blocks within the region of the object of interest. $\hat{S}_{\mathrm{out}}$ denotes the maximum value of all blocks outside the region. In our case, we exclude blocks near the borders of the surprise map since there the virtual images which predict the appearance of the scene often do not contain any information [cf. right border of Fig. 6(b)]. This automatically leads to high surprise values. 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, we introduce the peak attentional selectivity (PAS) which is computed as

$$\mathrm{PAS} = 10\log_{10}\left(\frac{\hat{S}}{\hat{S}_{\mathrm{out}}}\right)\ \mathrm{dB} \tag{25}$$

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

In the following analysis, we evaluate the AS and the PAS with respect to the regions of the glass and the cup over the image sequence $\mathcal{I}_1$. For comparison, we also evaluate the AS and PAS obtained by the method *image differencing*. Here, the appearance of the environment is stored in terms of a (nonprobabilistic) image-based representation and a virtual image is interpolated at the current viewpoint of the robot from nearby reference images. The image-based representation is continuously updated by replacing the latest selected reference image by the current observation. The absolute difference between the luminance and chrominance values in a new observation and in the predicted image is used in order to detect changes between the two images. The AS and PAS are computed using (24) and (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 our proposed representation is that the robot only stores deterministic snapshots of the environment and does not hold any information about the uncertainty of the appearance.

Fig. 10(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. When looking at the corresponding AS values in Fig. 10(b), we notice 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 Fig. 11. Here, we see that the edges around the objects in Fig. 11(d) show up elevated absolute difference values which are due to pose inaccuracies. This type of noise decreases the AS. In contrast, the surprise values around the object edges in Fig. 11(c) are relatively low, since our 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

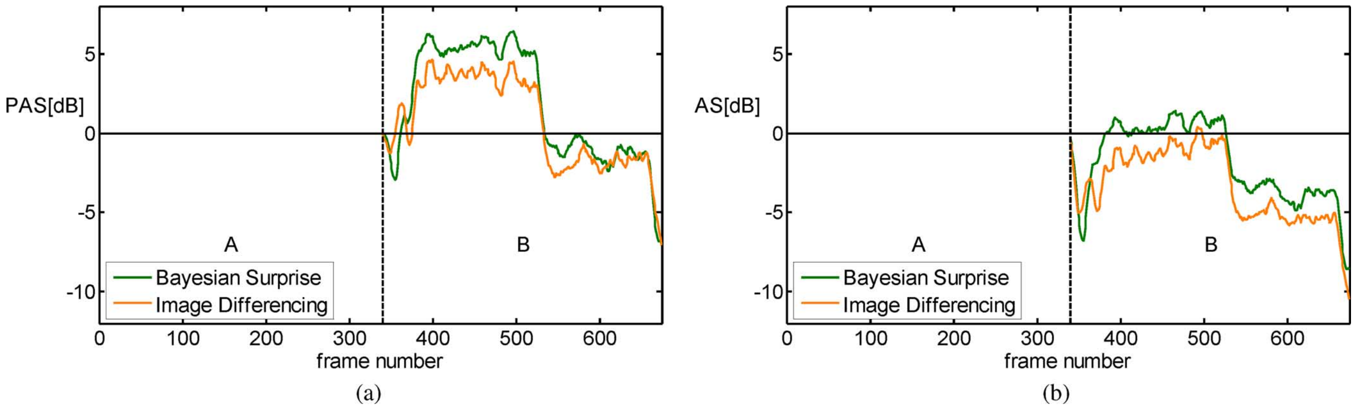


Fig. 10. The region of the glass is evaluated. Both the PAS values in (a) and the AS values in (b) are higher for Bayesian surprise than for image differencing. In (b), the AS values below 0 dB obtained by image differencing during the robot's run from Q_1 to Q_2 show that the glass does not convey more novelty than the rest of the scene. In contrast, Bayesian surprise detects the glass as a novel object.

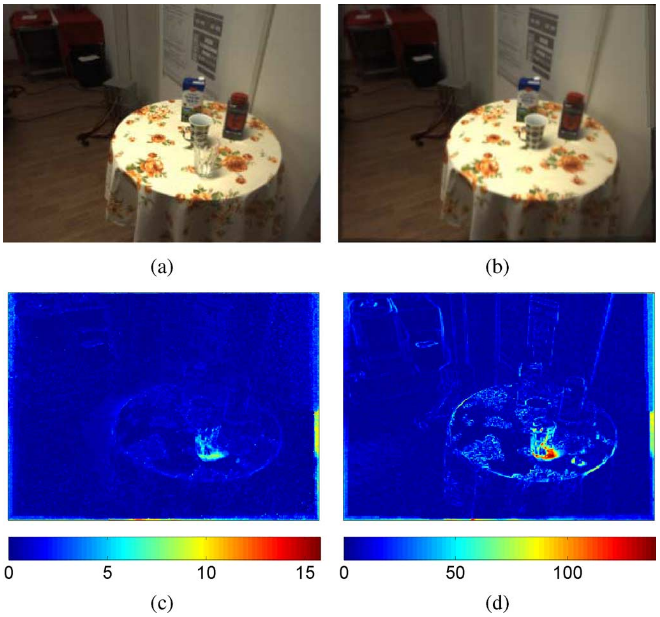


Fig. 11. (a) Frame 400 of the image sequence $\mathcal{I}_1$. (b) The virtual image which is interpolated from the robot's image-based representation and hence predicts the appearance of the scene. (c) The surprise values computed by our method are higher in the region of the glass than in the rest of the map. (d) The map obtained by image differencing is sensitive to pose inaccuracies and shows false positives near the edges of the objects.

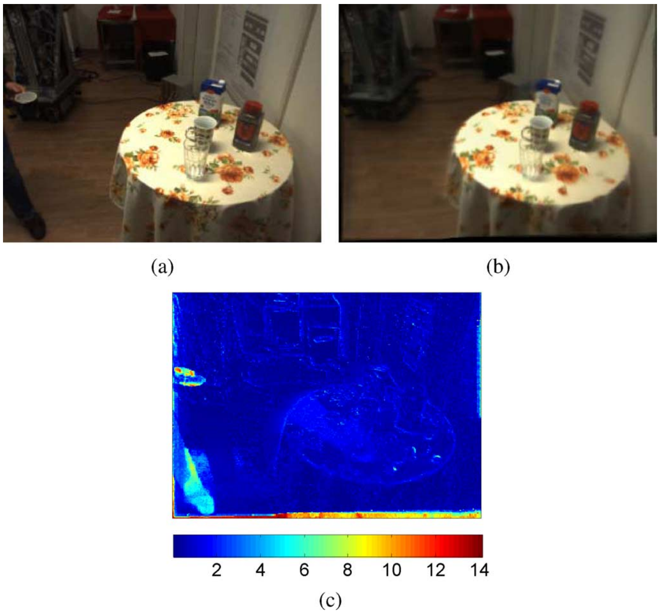
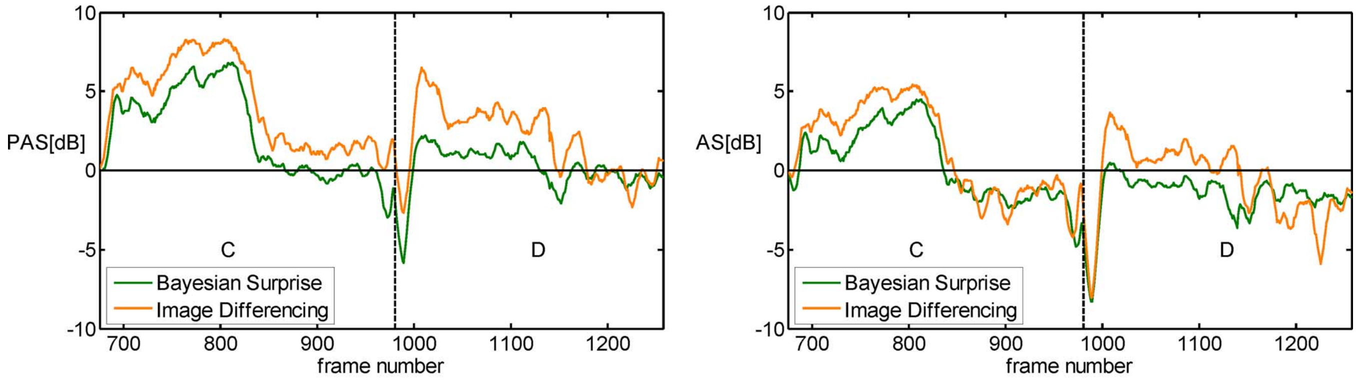


Fig. 12. The human who is about to put a cup on the table in frame 662 (a) is not expected by the robot (b) and thus, causes high surprise values near the left border (c).

the left (see Fig. 12). Both methods detect the human as a novelty and provide surprise values which are higher than the ones within the region of the glass.

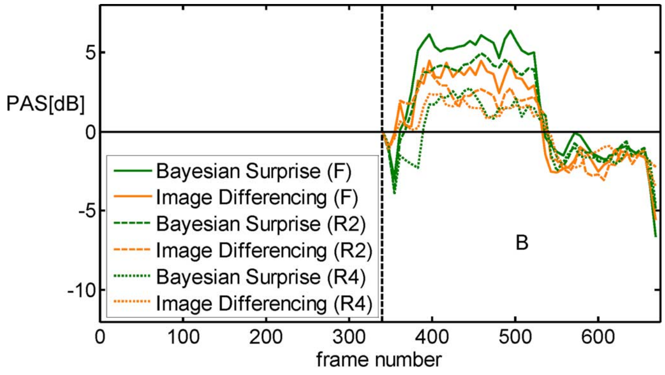
Fig. 13(a) compares the PAS values obtained by Bayesian surprise to the PAS values obtained by image differencing with respect to the region of the cup. Due to the large stimulus difference, which cause large absolute difference values especially in the luminance channel, the PAS obtained by image differencing is higher than the PAS obtained by Bayesian surprise. However, both the PAS and the AS values in Fig. 13(b) lie above 0 dB so that the cup is clearly detected as a novel object. In phase D, we see 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 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. 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, we investigate the effect of a reduction of the number of reference parameter images in the probabilistic appearance representation on the peak attentional selectivity. Fig. 14 compares the PAS values in phase B obtained by Bayesian surprise to the PAS values obtained by image differencing for different densitites of viewpoints. 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



(a) (b)

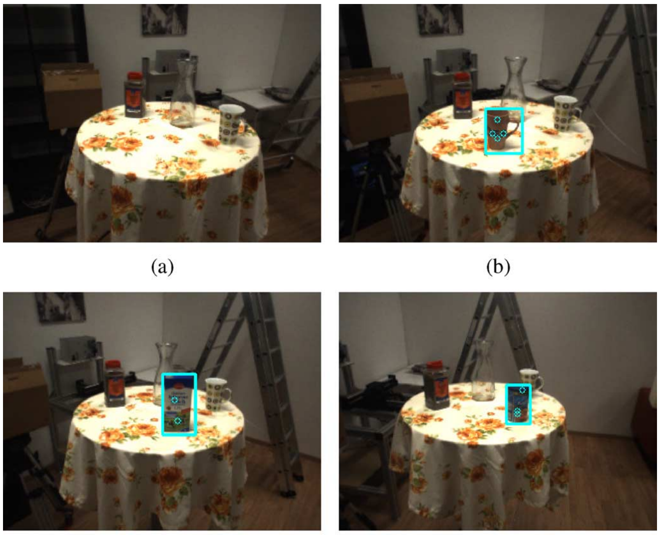
Fig. 13. The region of the cup is evaluated. Both the PAS values in (a) and the AS values in (b) are higher for image differencing than for Bayesian surprise. However, during the robot's first run from Q_1 to Q_2 the PAS and AS values obtained by Bayesian surprise are clearly above 0 dB. The high values of the PAS and AS obtained by image differencing at the beginning of phase D lead to a strong attentional control to 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. This is reflected by the lower AS values obtained by Bayesian surprise.



(a)

Fig. 14. 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.

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)." We see from the curves in Fig. 14 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 environment representation. In general, however, the optimal number of reference views in the representation always depends on the complexity of the scene [4].



(a) (b) (c) (d)

Fig. 15. (a) Frame 250 of the image sequence $\mathcal{I}_2$, which was used to infer the probabilistic environment representation. (b) Frame 185 of the image sequence $\mathcal{I}_3$, which shows the bounding rectangle around a new cup and four keypoints detected inside. (c) Frame 523 of the image sequence $\mathcal{I}_3$, which shows the bounding rectangle around a new milk carton and two keypoints detected inside. (d) Frame 824 of the image sequence $\mathcal{I}_3$, which shows the bounding rectangle around a new coffee can and three keypoints detected inside.

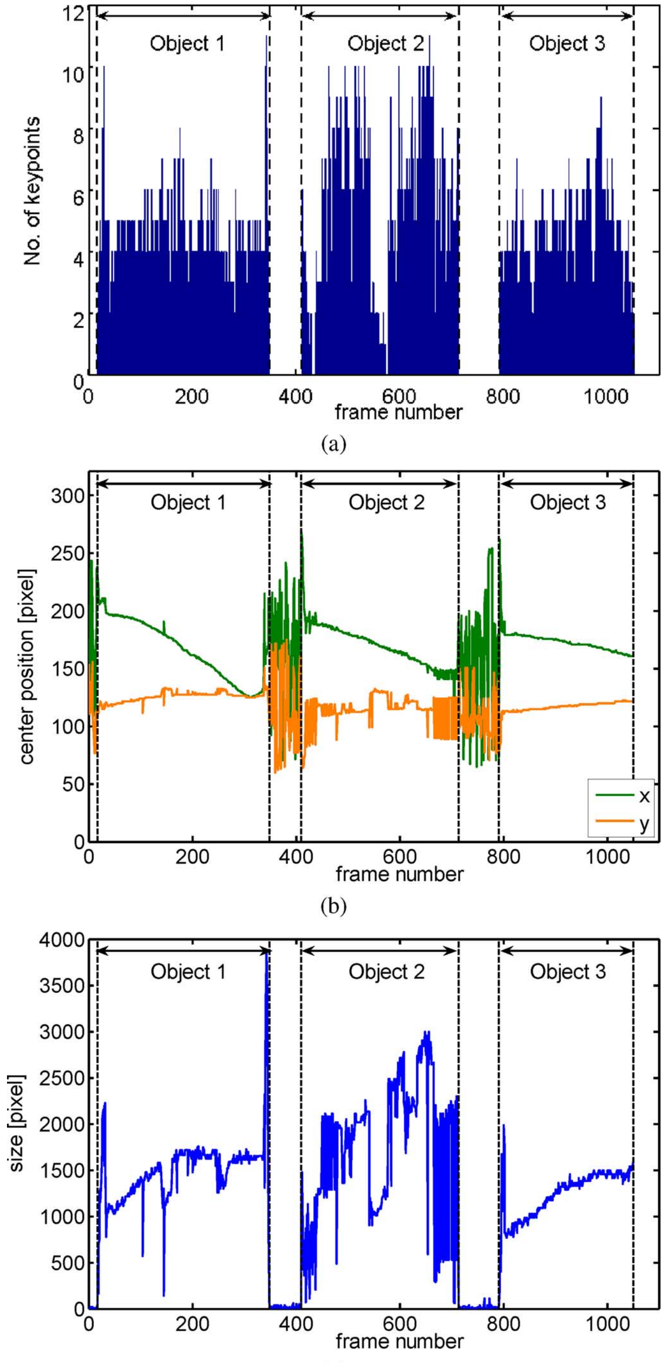
C. Selective Extraction of Visual Features Using Surprise Maps

Our surprise maps can be used to guide the attention of the robot and to select regions from the captured images which contain novelty. Hence, in this work, we consider the application of surprise detection to the selective extraction of descriptive features from new objects which are presented to the robot in a familiar environment. Due to the short computational time, we use speeded-up robust features (SURF) [32], [33], which are invariant to scale, rotation and, to a certain degree, to illumination. The association of the selected SURF features with a new created object class in a database facilitates the formation of higher-level object representations which can be used later for recognition.

For our experiments, we acquired two other image sequences $\mathcal{I}_2$, and $\mathcal{I}_3$ in a different part of our laboratory using the platform in Fig. 4(a). The robot again captured images of a table scene while moving on a circular trajectory between two turning points. The size of the images is 320 × 240 pixels. During the first image sequence $\mathcal{I}_2$ a probabilistic environment model containing 316 reference parameter images is infered. In the second image sequence $\mathcal{I}_3$, three unknown objects were presented to the robot – a cup (frames 40–315), a milk carton (frames 435–690), and a coffee can (frames 810–1050). Using a flood filling algorithm [34], whose seed point is the pixel location of the maximum surprise value in a surprise map, we identify the part of the image which is most surprising to the robot. We compute the bounding box of this region and select all features inside. Fig. 15 shows one image of the image sequence $\mathcal{I}_2$ [Fig. 15(a)] and three images of the sequence $\mathcal{I}_3$ [Fig. 15(b)–(d)], which indicate the computed bounding box around the new objects. The circles inside the bounding boxes visualize the keypoints of the extracted SURF features. We see in Fig. 16(a) that the number of features extracted inside the bounding box, in general, increases when a new object is presented to the robot while no features are extracted when there is no new object in the scene (e.g., between frames 316 and 434 or between frames 691 and 809). When there is no new object, the bounding boxes are usually very small [see Fig. 16(c)] and located at random positions in the surprise maps [see Fig. 16(b)] since the position of the maximum surprise value varies a lot with noise. In contrast, when the robot detects a new object, the focus of attention follows the object which can be seen from the smooth curves of the horizontal (x) and vertical (y) pixel position of the bounding boxes' center in Fig. 16(b). Hence, surprise strongly guides the robot's attention.

D. Limitations

One of the current limitations of our 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. However, a purely vision-based localization of the camera head is challenging due to the highly dynamic environment of the robot and odometry data is usually too inaccurate for our application.



(a)

(b)

(c)

Fig. 16. (a) The number of keypoints which are detected inside the bounding box of the surprise region in the frames of the image sequence $\mathcal{I}_3$. At each keypoint, a feature descriptor is computed. (b) The horizontal (x) and vertical (y) pixel position of the center of the bounding box along the image sequence $\mathcal{I}_3$. (c) The size of the bounding box along the image sequence $\mathcal{I}_3$.

VI. Conclusion

While image-based representations only provide snapshots of the environment at the moment of acquisition, the probabilistic appearance representation presented in this work enables the robot to reason about the uncertainty of the currently captured luminance and chrominance values. The parameters of prior distributions are inferred by the robot's observations and stored at a dense series of viewpoints. Using the pose information and per-pixel depth maps which are stored together with the parameter images, the robot is able to interpolate the parameters of the prior distribution at viewpoints where it has not been before. We show in our experimental results that our representation provides a realistic expectation of the environment's appearance even if complex transparent objects like glasses are present in the scene.

Our representation provides a framework for the computation of Bayesian surprise. A comparison to the image differencing method, which computes the difference between the currently acquired image and a virtual image interpolated from an image-based representation, shows that the surprise maps obtained by our method are more robust to noise due to pose inaccuracies. Furthermore, experimental results show that our surprise measure is a better detector for novelty than the measure provided by image differencing. Our surprise measure can be used to guide the robot's attention to novel image parts and facilitates selective feature extraction.

The approach which we present in this work, provides two important cues which can be used by a robot for the selection of action sequences. First, the detection of image regions which exhibit high surprise values allow for the selective extraction of object features. By tracking these features over several frames the robot can identify static objects in the environment and plan trajectories around them (under the consideration of obstacles) in order to get full representations of these objects. Second, the robot can develop policies to minimize the uncertainty about the environment's appearance in its internal representation by taking images preferably in regions where the uncertainty is highest. As pointed out in Section III-A, the update of the robot's expected appearance is driven by uncertainty.

Appendix A
Moments of the Normal-Gamma Distribution

The moment M_{01} of the normal-gamma distribution is calculated by

$$M_{01} = \int_{\mu=-\infty}^{+\infty} \int_{\lambda=-\infty}^{+\infty} \lambda \cdot \frac{\beta^{\alpha}}{\Gamma(\alpha)\sqrt{2\pi\sigma}} \cdot \lambda^{\alpha-1/2} \cdot \exp\{-\beta\lambda\} \cdot \exp\left\{-\frac{\lambda(\mu-\tau)^2}{2\sigma}\right\} \mathrm{d}\lambda\, \mathrm{d}\mu. \quad (26)$$

With the substitution

$$\epsilon := \alpha + 1 \quad (27)$$

we obtain

$$\begin{aligned} M_{01} &= \int_{\mu=-\infty}^{+\infty} \int_{\lambda=-\infty}^{+\infty} \frac{\beta^{\epsilon-1}}{\Gamma(\epsilon-1)\sqrt{2\pi\sigma}} \cdot \lambda^{\epsilon-1/2} \cdot \exp\{-\beta\lambda\} \cdot \exp\left\{-\frac{\lambda(\mu-\tau)^2}{2\sigma}\right\} \mathrm{d}\lambda\, \mathrm{d}\mu \\ &= \frac{\epsilon-1}{\beta} \int_{\mu=-\infty}^{+\infty} \int_{\lambda=-\infty}^{+\infty} \frac{\beta^{\epsilon}}{\Gamma(\epsilon)\sqrt{2\pi\sigma}} \cdot \lambda^{\epsilon-1/2} \cdot \exp\{-\beta\lambda\} \cdot \exp\left\{-\frac{\lambda(\mu-\tau)^2}{2\sigma}\right\} \mathrm{d}\lambda\, \mathrm{d}\mu = \frac{\epsilon-1}{\beta} \\ &= \frac{\alpha}{\beta}. \end{aligned} \quad (28)$$

Here, we used the relationship

$$\Gamma(\epsilon) = (\epsilon-1)\Gamma(\epsilon-1). \quad (29)$$

Using (28), the moment M_{11} is calculated by

$$\begin{aligned} M_{11} &= \int_{\mu=-\infty}^{+\infty} \int_{\lambda=-\infty}^{+\infty} \mu\lambda \cdot \frac{\beta^{\alpha}}{\Gamma(\alpha)\sqrt{2\pi\sigma}} \cdot \lambda^{\alpha-1/2} \cdot \exp\{-\beta\lambda\} \cdot \exp\left\{-\frac{\lambda(\mu-\tau)^2}{2\sigma}\right\} \mathrm{d}\lambda\, \mathrm{d}\mu \\ &= \int_{\lambda=-\infty}^{+\infty} \lambda \cdot \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \lambda^{\alpha-1} \cdot \exp\{-\beta\lambda\} \cdot \left[\int_{\mu=-\infty}^{+\infty} \mu \cdot \frac{1}{\sqrt{2\pi\frac{\sigma}{\lambda}}} \exp\left\{-\frac{(\mu-\tau)^2}{2\frac{\sigma}{\lambda}}\right\} \mathrm{d}\mu\right] \mathrm{d}\lambda \\ &= \int_{\lambda=-\infty}^{+\infty} \lambda \cdot \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \lambda^{\alpha-1} \cdot \exp\{-\beta\lambda\} \cdot \tau\, \mathrm{d}\lambda \\ &= \tau \cdot \frac{\alpha}{\beta}. \end{aligned} \quad (30)$$

Similarly, the moment M_{21} is computed by

$$\begin{aligned} M_{21} &= \int_{\mu=-\infty}^{+\infty} \int_{\lambda=-\infty}^{+\infty} \mu^2\lambda \cdot \frac{\beta^{\alpha}}{\Gamma(\alpha)\sqrt{2\pi\sigma}} \cdot \lambda^{\alpha-1/2} \cdot \exp\{-\beta\lambda\} \cdot \exp\left\{-\frac{\lambda(\mu-\tau)^2}{2\sigma}\right\} \mathrm{d}\lambda\, \mathrm{d}\mu \\ &= \int_{\lambda=-\infty}^{+\infty} \lambda \cdot \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \lambda^{\alpha-1} \cdot \exp\{-\beta\lambda\} \cdot \left[\int_{\mu=-\infty}^{+\infty} \mu^2 \cdot \frac{1}{\sqrt{2\pi\frac{\sigma}{\lambda}}} \exp\left\{-\frac{(\mu-\tau)^2}{2\frac{\sigma}{\lambda}}\right\} \mathrm{d}\mu\right] \mathrm{d}\lambda \\ &= \int_{\lambda=-\infty}^{+\infty} \lambda \cdot \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \lambda^{\alpha-1} \cdot \exp\{-\beta\lambda\} \cdot \left[\tau^2 + \frac{\sigma}{\lambda}\right] \mathrm{d}\lambda \\ &= \tau^2 \cdot \frac{\alpha}{\beta} + \sigma. \end{aligned} \quad (31)$$

Appendix B
The Kullback–Leibler Divergence of Two Normal-Gamma Distributions

Be $p_0(\mu,\lambda)$ and $p(\mu,\lambda)$ two normal-gamma distributions

$$p_0(\mu,\lambda) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)\sqrt{2\pi\sigma_0}} \cdot \lambda^{\alpha_0-1/2} \cdot \exp\{-\beta_0\lambda\} \cdot \exp\left\{-\frac{\lambda(\mu-\tau_0)^2}{2\sigma_0}\right\} \quad (32)$$

$$p(\mu,\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)\sqrt{2\pi\sigma}} \cdot \lambda^{\alpha-1/2} \cdot \exp\{-\beta\lambda\} \cdot \exp\left\{-\frac{\lambda(\mu-\tau)^2}{2\sigma}\right\}. \quad (33)$$

The Kullback–Leibler divergence of the two normal-gamma distributions is computed by

$$\mathcal{KL}(p(\mu,\lambda); p_0(\mu,\lambda)) = \int_{\mu=-\infty}^{+\infty} \int_{\lambda=-\infty}^{+\infty} p(\mu,\lambda) \cdot \log\left(\frac{p(\mu,\lambda)}{p_0(\mu,\lambda)}\right) \mathrm{d}\lambda\, \mathrm{d}\mu. \quad (34)$$

Here, it is convenient to use the natural logarithm since the normal-gamma distribution contains two natural exponential terms.

In order to keep the notation uncluttered, we write the Kullback–Leibler divergence as a summation of five terms

$$\mathcal{KL}\left(p\left(\mu,\lambda\right);p_0\left(\mu,\lambda\right)\right) = T_1 + T_2 + T_3 + T_4 + T_5. \quad (35)$$

For T_1, we obtain

$$\begin{aligned} T_1 &= \int_{\mu=-\infty}^{+\infty}\int_{\lambda=-\infty}^{+\infty} p\left(\mu,\lambda\right)\cdot\log\left(\frac{\frac{\beta^{\alpha}}{\Gamma(\alpha)\sqrt{2\pi\sigma}}}{\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)\sqrt{2\pi\sigma_0}}}\right)\,\mathrm{d}\lambda\,\mathrm{d}\mu \\ &= \log\left(\frac{\beta^{\alpha}\cdot\Gamma(\alpha_0)\cdot\sqrt{\sigma_0}}{\beta_0^{\alpha_0}\cdot\Gamma(\alpha)\cdot\sqrt{\sigma}}\right). \end{aligned} \quad (36)$$

Using (28), T_2 is computed by

$$\begin{aligned} T_2 &= \left(\beta_0-\beta\right)\cdot\int_{\mu=-\infty}^{+\infty}\int_{\lambda=-\infty}^{+\infty}\lambda\cdot p\left(\mu,\lambda\right)\,\mathrm{d}\lambda\,\mathrm{d}\mu \\ &= \left(\beta_0-\beta\right)\cdot\frac{\alpha}{\beta}. \end{aligned} \quad (37)$$

T_3 is computed by

$$\begin{aligned} T_3 = &\left(\int_{\lambda=-\infty}^{+\infty}\log\left(\lambda\right)\cdot\frac{\beta^{\alpha}}{\Gamma(\alpha)}\cdot\lambda^{\alpha-1/2}\cdot\exp\left\{-\beta\lambda\right\}\right. \\ &\left.\cdot\left[\int_{\mu=-\infty}^{+\infty}\frac{\frac{1}{\sqrt{\lambda}}}{\sqrt{2\pi\frac{\sigma}{\lambda}}}\cdot\exp\left\{-\frac{\left(\mu-\tau\right)^2}{2\frac{\sigma}{\lambda}}\right\}\mathrm{d}\mu\right]\mathrm{d}\lambda\right) \\ &\cdot\left(\alpha-\alpha_0\right) \\ = &\left(\int_{\lambda=-\infty}^{+\infty}\log\left(\lambda\right)\cdot\frac{\beta^{\alpha}}{\Gamma(\alpha)}\cdot\lambda^{\alpha-1}\cdot\exp\left\{-\beta\lambda\right\}\mathrm{d}\lambda\right) \\ &\cdot\left(\alpha-\alpha_0\right) = \left(\psi\left(\alpha\right)-\log\left(\beta\right)\right)\cdot\left(\alpha-\alpha_0\right). \end{aligned} \quad (38)$$

Using (28), (30), and (31), we obtain T_4 by

$$\begin{aligned} T_4 &= \frac{1}{2\sigma_0}\int_{\lambda=-\infty}^{+\infty}\int_{\mu=-\infty}^{+\infty}\left[\lambda\mu^2-2\lambda\mu\tau_0+\lambda\tau_0^2\right]\cdot p\left(\mu,\lambda\right)\mathrm{d}\mu\mathrm{d}\lambda \\ &= \frac{1}{2\sigma_0}\left[\tau^2\cdot\frac{\alpha}{\beta}+\sigma-2\tau_0\tau\cdot\frac{\alpha}{\beta}+\tau_0^2\cdot\frac{\alpha}{\beta}\right]. \end{aligned} \quad (39)$$

Similarly, we obtain T_5 by

$$\begin{aligned} T_5 &= -\frac{1}{2\sigma}\int_{\lambda=-\infty}^{+\infty}\int_{\mu=-\infty}^{+\infty}\left[\lambda\mu^2-2\lambda\mu\tau+\lambda\tau^2\right]\cdot p\left(\mu,\lambda\right)\mathrm{d}\mu\mathrm{d}\lambda \\ &= -\frac{1}{2\sigma}\left[\tau^2\cdot\frac{\alpha}{\beta}+\sigma-2\tau^2\cdot\frac{\alpha}{\beta}+\tau^2\cdot\frac{\alpha}{\beta}\right] = -\frac{1}{2}. \end{aligned} \quad (40)$$

ACKNOWLEDGMENT

The authors would like to thank D. Althoff, O. Kourakos, M. Lawitzky, and M. Rambow from the Institute of Automatic Control Engineering, as well as N. Alt from the Institute of Media Technology at the Technische Universität München for their support during the acquisition of the image sequence.

REFERENCES

[1] S. Thrun, "Robotic mapping: A survey," in *Exploring Artificial Intelligence in the New Millenium*, G. Lakemeyer and B. Nebel, Eds. San Mateo, CA: Morgan Kaufmann, 2002, pp. 1–35.

[2] P. E. Debevec, C. J. Taylor, and J. Malik, "Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach," in *Proc. Int. Conf. Comput. Graph. Interact. Techniq. (ACM SIGGRAPH)*, New Orleans, LA, 1996, pp. 11–20.

[3] T. J. Purcell, I. Buck, W. R. Mark, and P. Hanrahan, "Ray tracing on programmable graphics hardware," *ACM Trans. Graphics*, vol. 21, no. 3, pp. 703–712, 2002.

[4] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic sampling," in *Proc. Int. Conf. Comput. Graph. Interact. Techniq. (ACM SIGGRAPH)*, New Orleans, LA, 2000, pp. 307–318.

[5] C. L. Zitnick *et al.*, "High-quality video view interpolation using a layered representation," in *Proc. Int. Conf. Comput. Graph. Interact. Techniq. (ACM SIGGRAPH)*, Los Angeles, CA, 2004, pp. 600–608.

[6] M. Markou and S. Singh, "Novelty detection: A review—Part 1: Statistical approaches," *Signal Process.*, vol. 83, no. 12, pp. 2481–2497, 2003.

[7] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Rev.: Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.

[8] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vis. Res.*, vol. 49, no. 10, pp. 1295–1306, 2009.

[9] 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 Int. Conf. Robot. Autom. (ICRA)*, Kobe, Japan, 2009, pp. 807–812.

[10] J. R. Anderson *et al.*, "An integrated theory of the mind," *Psychol. Rev.*, vol. 111, no. 4, pp. 1036–1060, 2004.

[11] P. Langley and D. Choi, "A unified cognitive architecture for physical agents," in *Proc. AAAI Conf. Artif. Intell.*, Boston, MA, 2006, pp. 1469–1474.

[12] H.-Y. Shum, S.-C. Chan, and S. B. Kang*, Image-Based Rendering*. Berlin, Germany: Springer-Verlag, 2007.

[13] F. Fraundorfer, C. Engels, and D. Nister, "Topological mapping, localization and navigation using image collections," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, San Diego, CA, 2007, pp. 3872–3877.

[14] H. E. M. den Ouden *et al.*, "A dual role for prediction error in associative learning," *Cereb. Cortex*, vol. 19, no. 5, pp. 1175–1185, 2009.

[15] P. C. Fletcher *et al.*, "Responses of human frontal cortex to surprising events are predicted by formal associative learning theory," *Nature Neurosci.*, vol. 4, no. 10, pp. 1043–1048, 2001.

[16] J. Schmidhuber, J. Storck, and J. Hochreiter, "Reinforcement driven information acquisition in non-deterministic environments," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, Paris, France, 1995, pp. 159–164.

[17] J. Schmidhuber, "Self-motivated development through rewards for predictor errors/improvements," in *Proc. AAAI Spring Symp. Develop. Robot.*, Stanford, CA, 2005.

[18] A. Stout, G. D. Kondaris, and A. G. Barto, "Intrinsically motivated reinforcement learning: A promising framework for developmental robot learning," in *Proc. AAAI Spring Symp. Develop. Robot.*, Stanford, CA, 2005.

[19] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Trans. Evol. Comput.*, vol. 11, no. 2, pp. 265–286, Mar. 2007.

[20] X. Huang and J. Weng, "Novelty and reinforcement learning in the value system of developmental robots," in *Proc. 2nd Int. Workshop Epig. Robot.: Model. Cogn. Develop. Robot. Syst.*, Edinburgh, Scotland, 2002, pp. 47–55.

[21] A. Ranganathan and F. Dellaert, "Bayesian surprise and landmark detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Rome, Italy, 2009, pp. 2007–2023.

[22] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[23] S. Engel, X. Zhang, and B. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," *Nature*, vol. 388, no. 6637, pp. 68–71, 1997.

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

[25] C. M. Bishop*, Pattern Recognition and Machine Learning*. Berlin, Germany: Springer-Verlag, 2006.

[26] J. M. Bernardo and A. F. M. Smith*, Bayesian Theory*. New York: Wiley, 2000.

[27] L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn. (CVPR)*, San Diego, CA, 2005, pp. 631–637.

[28] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 698–700, Sep. 1987.

[29] R. T. Collins, “A space-sweep approach to true multi-image matching,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn. (CVPR)*, San Francisco, CA, 1996, pp. 358–363.

[30] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient belief propagation for early vision,” *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 41–54, 2006.

[31] D. Brščić *et al.*, Multi Joint Action in CoTeSys – Setup and Challenges CoTeSys Cluster of Excelence: Technische Universität München & Ludwig-Maximilians-Universität München, Munich, Germany, 2010, Tech. Rep. CoTeSys-TR-10-01.

[32] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, “Speeded-up robust features (SURF),” *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[33] C. Evans, Notes on the OpenSURF Library University of Bristol, Bristol, U.K., 2009, Tech. Rep. CSTR-09-001.

[34] G. Bradski, “The OpenCV Library,” *Dr. Dobb’s Journal of Software Tools*, pp. 120–126, 2000.

Werner Maier (S’08–M’10) studied electrical engineering at the Technische Universität München, München, Germany, and at the Universidad de Navarra, Spain. He received the B.Sc. degree in 2004 and the degree "Dipl.-Ing. (Univ)" in 2006.

In October 2006, he joined the Media Technology Group at the Technische Universität München where he has been working as a member of the research and teaching staff. His current research interests are in the field of environment modeling for cognitive technical systems.

Eckehard Steinbach (S’96–A’99–M’04–SM’08) studied Electrical Engineering at the University of Karlsruhe, Karlsruhe, Germany, the University of Essex, Essex, U.K., and the ESIEE in Paris. From 1994 to 2000, he was a member of the research staff of the Image Communication Group at the University of Erlangen-Nuremberg, Germany, where he received the Engineering Doctorate in 1999.

From February 2000 to December 2001, he was a Postdoctoral Fellow with the Information Systems Laboratory of Stanford University, Stanford, CA. In February 2002, he joined the Department of Electrical Engineering and Information Technology of the Munich University of Technology, Munich, Germany, where he is currently a Full Professor for Media Technology. His current research interests are in the area of audio–visual–haptic information processing and communication, as well as networked and interactive multimedia systems.