Visual Homing and Surprise Detection for Cognitive Mobile Robots Using Image-Based Environment Representations

Werner Maier, Elmar Mair, Darius Burschka and Eckehard Steinbach

Abstract— **One important feature of a cognitive system is to perceive and understand its environment and to adapt its actions to changes and unforeseen situations. In this paper, we propose a scheme for visual surprise detection in cognitive mobile robots. With the robot's observation and a set of reference images previously acquired near its current viewpoint, a pixel-wise surprise trigger is computed using Bayesian probabilistic inference techniques. With appropriate mathematical approximations this algorithm can be implemented on modern graphics hardware which nearly allows for real-time surprise detection. In order to refer to prior observations, a mobile robot has to be able to re-localize itself with respect to its environment. Thus, we also present two online image-based homing algorithms which both facilitate the computation of location-independent surprise triggers. Experiments show acceptable results in terms of robust and fast detection of unexpected changes in the environment.**

I. INTRODUCTION

Cognitive technical systems need to be aware of the environment they are acting in. Common environment representations for mobile robots are based on geometric 3D maps which are acquired by the agent using laser range finders or fused from stereo vision data. These maps contain important cues for navigation and make the robot aware of potential collisions. 3D models of single objects are also indispensable for grasping. In practice, cognitive technical systems have to act like humans in dynamic real-world environments which are permanently changing over time. Detecting changes in the 3D world in order to update the internal model is very challenging since the cognitive system has to check all the points and triangles in its reference representation for alterations. It is impossible to build up an internal dynamic representation of the whole environment containing all changes over time.

The agent has to learn to evaluate its perception and to focus on events which are of interest and convey some kind of novelty. Hence, the update of the environment model has to be attention-driven. Furthermore, changes in the environment can be detected much faster with an appearance model than with a purely geometry-based model. Using the appearance for change detection, an internal virtual image is computed which serves as a reference for the current observation. In the field of computer graphics, many methods exist for the visualization of 3D models with image data. However, in case of translucent or filigree objects it is very challenging to acquire accurate geometric models with common computer vision techniques.

Image-based rendering techniques ([1]) have been shown to be very suitable for the photorealistic modeling of such objects because the computational complexity does not depend on the properties of the environment. In our approach, we use a densely acquired set of images together with view-dependent geometry information in order to predict virtual images in a given viewpoint space. A necessary cue for view synthesis is the knowledge about the positions and orientations of the capturing cameras, which have to be estimated. The localization algorithm we use is also purely based on images.

Beyond that, a cognitive system which should be able to recognize changes in the environment over time has to re-localize itself with respect to the coordinate frame of its internal representation. A detailed classification of visual intensity-based homing algorithms can be found in [8]. In [10], various methods are compared with respect to their efficiency. Most of them are biologically inspired (e.g. [9] or [11]). We present two structure-based "snapshot" approaches, where the structure is based on images again. These methods allow a cognitive system to register the captured images in a partially seen environment.

In [4], it is shown that surprise is an important cue for the direction of human attention to unexpected events. A variety of image change detection algorithms have been presented in literature ([12]). However, they all have in common that they are only applied to images taken from the same camera at a rigid position. For a mobile cognitive technical system this is not acceptable since it also needs to notice changes at positions where no previous camera image is available. Therefore, we propose in this work an algorithm for visual surprise detection in cognitive technical systems, which relies on accurate visual registration of the system's cameras and image-based environment modeling. Surprise detection is applicable from any point in the world and at any time because of the underlying homing algorithms.

The remainder of this paper is structured as follows. After a brief outline of the visual localization method used in this work in Section II, Section III proposes two online solutions to the homing problem. The image-based view synthesis method which is briefly outlined in Section IV is the basis of our novel method for visual surprise detection described in Section V. Before we conclude this work, we present in

This work is supported in part within the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see also www.cotesys.org

W. Maier and E. Steinbach are with the Media Technology Group, Technische Universität München, 80333 München, Germany werner.maier@tum.de, eckehard.steinbach@tum.de

E. Mair and D. Burschka are with the Lab for Robotics and Embedded Systems, Technische Universität München, 85748 Garching, Germany elmar.mair@mytum.de, burschka@cs.tum.edu

Section VI selected experimental results.

II. VISUAL LOCALIZATION

The first step in the generation of image-based models is the accurate localization of the captured images. Our real-time capable algorithm allows us to estimate the position and orientation of the calibrated camera during the acquisition of the image sequence. Since our method does not require any external references like, for example, artificial markers in the scene or the dimensions of a known object in the world, it makes our algorithm very flexible and suitable for a cognitive system navigating in real-world environments. Features tracked by the Kanade-Lucas-Tomasi (KLT) tracker [6] are used to estimate the camera position with a robustified variant of the visual GPS (RVGPS) ([18]) algorithm. A second camera is used to initialize the 3D structure of the features, which is necessary for RVGPS. Thus, apart from the highly accurate initialization the scale of the translation is known, too.

III. VISUAL HOMING

Our localization method provides acceptable results with respect to position and orientation as long as the tracker finds enough matches between the images. However, it fails as soon as there are too few tracked landmarks for pose estimation. This may happen if the torsional moment of the camera is too high, so that all references leave its field of view or just the tracker's search range. After a simple reboot, the relation to the prior run can also be lost. Even if all detected features were saved on a hard drive, the cameras could not be registered within the prior world coordinate frame, as soon as the robot moves outside the known trajectory. We need to register the new sequence with respect to the reference coordinate system in order to establish a relationship to the previously acquired data. Since we do not use external markers as reference, which could be used to determine the origin of the reference frame, we need to initially specify an arbitrary origin. All the information which is necessary to refer to this origin, whenever required, has to be stored – a so called "snapshot" has to be taken. In this chapter, we present two different approaches to solve this problem.

Let us assume that we have done a first run, where we acquired an image sequence. Now we want to perform a second run and estimate the camera poses with respect to the coordinate frame of the first run. W.l.o.g., we call the first sequence S_1 and the second sequence S_2. The reference image is assumed to be the first of S_1 with the initialized KLT feature set. It defines the origin of the coordinate frame and is denoted as $I_{1.1}$. Now, S_2 should be registered with respect to the coordinate frame of S_1, the so called reference frame. The first image in S_2 is called $I_{2.1}$ and our aim is now to localize it with respect to $I_{1.1}$.

First of all we have to find a relation between the two viewpoints. To find feature correspondences between two images, which are related to each other by an affine transformation, we can not use KLT any longer. However, in the last decades various detector-descriptor combinations were investigated, which are also able to deal with such transformations. The most well-known and widely used one is probably SIFT ([14]). A well-known drawback is its speed. SIFT uses complex detection functions and large descriptor vectors which make it independent of any affine transformation but slow down the whole algorithm. A newer, alternative approach is SURF ([13]), which is supposed to be at least as accurate and robust, but faster than SIFT. Hence, we use SURF to find correspondences between images which show the same scene but from different viewpoints.

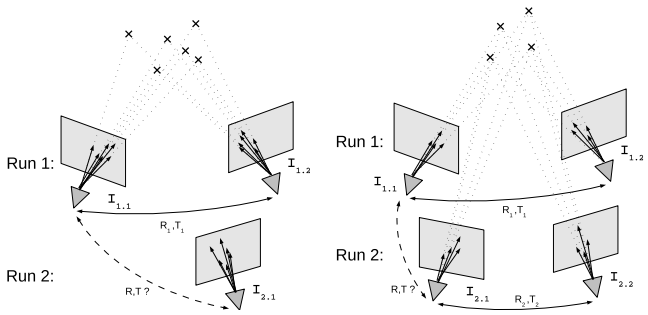
A. Homing based on three images

In our first homing approach (further on Homing1) we use RVGPS for extrinsic parameter estimation (see Section II). RVGPS is now used to estimate the rotation and translation between the current and the reference frame. We only need $I_{1.1}$ and its initialized points of interest (POIs), provided by SURF. This 3D structure forms the so called snapshot of the origin. SURF and KLT use different detectors, hence the stereo-registration method used by the visual localization scheme in Section II cannot be used between $I_{1.1}$ and $I_{2.1}$. Instead the POIs are initialized by a so called *structure from motion* approach: The distance of the camera between $I_{1.1}$ and $I_{1.2}$, which is estimated by our visual localization routine, is used as baseline for stereo triangulation. Once the SURF-features are initialized, we only need at least 3 SURF correspondences between $I_{1.1}$ and $I_{2.1}$ to apply RVGPS in order to estimate the six degrees of freedom (DOFs). Of course, the robustness and accuracy rapidly increases if more matching features are available. Thus, big parts of the same scene should be seen by these three images to ensure that enough matches are found. Otherwise one can also use more than one image in S_1 to initialize more POIs in $I_{1.1}$. The more points are available, the higher is the probability that correspondences in $I_{2.1}$ are found and the higher is also the accuracy of the motion estimation. Fig. 1(a) illustrates the principle of the Homing1 algorithm, which needs only 3 images.

B. Homing based on four images

The Homing1 variant has shown that its results strongly depend on the accuracy of the POIs' structure. Our second approach has been developed with the aim of avoiding that inconvenience by not using RVGPS for localizing $I_{2.1}$ with respect to $I_{1.1}$. Instead we are looking for an optimal matching of two 3D structures in the different coordinate frames of S_1 and S_2 in order to estimate the six DOFs. To calculate two corresponding structures for our second homing algorithm (Homing2) we need for each sequence S_1 and S_2 two images, their extrinsic parameters and the SURF correspondences in all 4 images. The extrinsic parameters for structure initialization in $I_{1.1}$ resp. $I_{2.1}$ are estimated in the same way as in the Homing1 method - by the visual localization routine and subsequent stereo triangulation. Using Arun's algorithm ([15]) we can calculate the transformation matrix between the two frames of S_1 and S_2. The result of

this method is obviously more robust, because we do not estimate the transformation matrix and the structure of the point set at the same time, like in Homing1. On the other hand we need to find SURF matches in 4 images, which is more problematic than with 3 images due to the smaller common feature intersection. Fig. 1(b) depicts the principle of the Homing2 algorithm based on 4 images.



(a) Homing using 3 images. (b) Homing using 4 images.

Fig. 1. *Fig. 1(a)*: The transformation between sequence 1 and 2 is estimated using the RVGPS algorithm. Thus, only one image of run 2 is necessary. *Fig. 1(b)*: The point structure in run 2 is initialized independently of the reference sequence (run 1), so that a higher accuracy is provided at the cost of the robustness (usually fewer common features are found).

Which algorithm to use therefore strongly depends on the application and the scene. Since the errors do not vary much (compare the subfigures in Fig. 7), the more robust but less accurate Homing1 algorithm is preferable in most cases.

IV. IMAGE-BASED VIEW PREDICTION

One type of image-based scene representation that recently has become very popular uses view-dependent geometry and texture. Instead of computing a global geometry model which is valid for any viewpoint and viewing direction, the geometry of the scene is locally estimated and only holds for a small region in the viewpoint space. It has been shown that this approach is suitable especially when the scene contains specular and translucent objects. To extract local geometry information, per-pixel depth maps are calculated for each reference image, i.e., the left image of each captured stereo pair. Loopy belief propagation [2] minimizes a matching cost volume and yields the most probable depth value for each pixel, assuming that the scene is smooth between depth discontinuities. A triangulated mesh is reconstructed from each depth map and simplified using the algorithm in [3]. While these steps are done off-line, the view selection and view synthesis, as explained in the following, are performed on-line.

The view selection only chooses a small subset of all captured images which contribute to view interpolation, each time a new frame is rendered. In real-world environments most surfaces are non-Lambertian, which means that the reflected intensity depends on the position of the viewer. Hence, in our approach, the reference cameras are ranked in terms of their orthogonal distance with respect to predefined rays within the viewing frustum of the virtual camera. Before a new frame is rendered, the seven closest ones are selected.

Novel virtual views are synthesized in a two-pass procedure. In the first rendering pass, the color data of the selected reference images which is associated with the triangle vertices of the view-dependent meshes is warped into the virtual view. Pixels that lie inside the projected triangles are interpolated from the color values at the corners. The second rendering pass then determines the final color of each pixel in the virtual view.

Our approach for view synthesis infers probabilistic models for the single pixel colors in the virtual image. It is assumed that the warped color values of the reference images are data samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_7]$ which are independently drawn from a Gaussian distribution whose mean value μ is identical to the true color value at the respective pixel in the virtual view (see Fig. 2 for illustration).

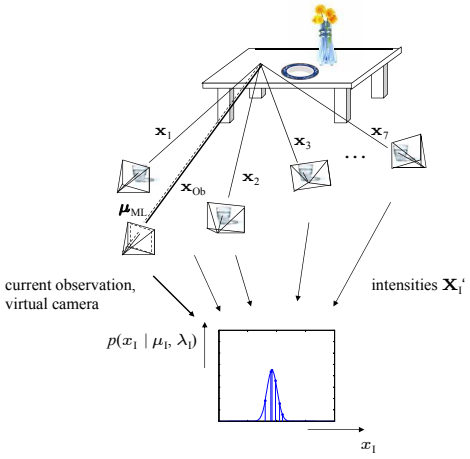


Fig. 2. View synthesis: The pixels of the virtual image are predicted from a set of reference images. The color values from the reference images are assumed to be samples from a Gaussian distribution whose mean is the true color value. Surprise detection: If the current observation yields a sample value which is, due to changes in the scene, largely different from the reference samples, a surprise trigger is generated in a given pixel region.

The task of the second rendering pass is to find an acceptable estimate for the true color value μ. A common method to determine the parameters of a probability distribution from sample data is maximum-likelihood (ML) estimation which yields

$$\mu_{\mathrm{ML}} = \frac{1}{7} \sum_{k=1}^{7} \mathbf{x}_k \tag{1}$$

where $\mathbf{x}_k = [x_{\mathrm{R},k}, x_{\mathrm{G},k}, x_{\mathrm{B},k}]^{\mathrm{T}}, k = 1, \ldots, 7$ are RGB-tripel. Thus, the estimated color value at a given pixel position in the virtual image is the arithmetic mean of the pixels colors from the seven reference images.

V. SURPRISE DETECTION

The ML estimates for the mean and the covariance of the Gaussian distribution are point estimates which give one model which describes the statistical properties of the sample

data. However, the estimates still deviate from their true values and there are other less probable parameterizations for the Gaussian distribution. Unlike ML estimation, Bayesian inference takes into account all possible models and puts priors over the parameters of the probability distribution of the sample data. In [4], a Bayesian framework was presented for modeling and quantifying human surprise in a mathematical way. Inspired by that, we propose in the following a scheme for Bayesian visual surprise detection based on the probabilistic concept for view synthesis.

For surprise detection the set of samples consists of seven RGB-tripels from reference images captured in the past and an additional color value from the current observation. As depicted in Fig. 2, the virtual camera and the real camera capturing the current image have identical position and orientation. Hence, accurate localization of the cognitive system's camera is crucial for robust surprise detection. Similar to the processing of color information in the human visual system ([16]), we compute from each RGB reference image a luminance signal and two color opponency signals (red-green and blue-yellow), respectively. Thus, surprise detection does not have to be performed jointly in RGB-space but can be done independently in three decoupled pathways. For the luminance of a pixel in the virtual image the following likelihood function for a univariate Gaussian model results:

$$p(\mathbf{X}_\mathrm{I} \mid \mu_\mathrm{I}, \lambda_\mathrm{I}) = \prod_{k=1}^{7} \left(\frac{\lambda_\mathrm{I}}{2\pi}\right)^{\frac{1}{2}} \exp\left\{-\tfrac{\lambda_\mathrm{I}}{2}\left(x_{\mathrm{I},k} - \mu_\mathrm{I}\right)^2\right\}. \quad (2)$$

$\mathbf{X}_\mathrm{I} = [x_{\mathrm{I},1}, \ldots, x_{\mathrm{I},7}]$ is a vector containing the luminance samples from the reference images. μ_I denotes the true luminance value at the pixel in the virtual image which is also the mean of the Gaussian distribution. For the choice of the prior distributions it is more convenient to use the precision λ_I, which is defined by the reciprocal of the variance ($\lambda_\mathrm{I} \equiv \frac{1}{\sigma_\mathrm{I}^2}$). Assuming that the mean is given by its ML estimate $\mu_{\mathrm{I,ML}} = \sum_{k=1}^{7} x_{\mathrm{I},k}$, we put a prior over the precision which has the form of a gamma distribution

$$p(\lambda_\mathrm{I}) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda_\mathrm{I}^{a_0-1} \exp\left\{-b_0 \lambda_\mathrm{I}\right\}. \quad (3)$$

Here $\Gamma(a_0) = \int_0^\infty t^{a_0-1} \exp\left\{-t\right\} \mathrm{d}t$ denotes the gamma function which serves as a normalization constant. The shape of the distribution thus depends on the two hyperparameters a_0 and b_0.

With Bayes' formula the posterior distribution of the precision given the sample data is calculated from the likelihood function and the prior up to a scaling factor by

$$p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I}) \propto p(\mathbf{X}_\mathrm{I} \mid \mu_{\mathrm{I,ML}}, \lambda_\mathrm{I}) \cdot p(\lambda_\mathrm{I}) \quad (4)$$

Note that the posterior is again a gamma distribution with the hyperparameters $a = a_0 + \frac{7}{2}$ and $b = b_0 + \frac{1}{2}\sum_{k=1}^{7}\left(x_{\mathrm{I},k} - \mu_{\mathrm{I,ML}}\right)^2$ which depend on the sample data. The kind of prior whose posterior has the same functional form is called a conjugate prior. The advantage of conjugate priors is that their posteriors can again be used as priors for further analysis.

Now we augment our set of luminance samples by the luminance value which the current observation of the cognitive technical system provides ($\mathbf{X}'_\mathrm{I} = [x_{\mathrm{I},1}, \ldots, x_{\mathrm{I},7}, x_{\mathrm{I,ob}}]$). The posterior distribution over λ_I is then calculated by

$$p(\lambda_\mathrm{I} \mid \mathbf{X}'_\mathrm{I}) \propto p(x_{\mathrm{I,ob}} \mid \mu_{\mathrm{I,ML}}, \lambda_\mathrm{I}) \cdot p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I}) \quad (5)$$

which results in a gamma distribution with the hyperparamters $a' = a + \frac{1}{2}$ and $b' = b + \frac{1}{2}\left(x_{\mathrm{I,ob}} - \mu_{\mathrm{I,ML}}\right)^2$.

In [17], the Kullback-Leibler divergence (KLD) as the difference between the posterior distribution over the model parameters given a new observation and the prior distribution is proposed as a quantitative measure for surprise

$$\begin{aligned}&\mathrm{KLD}\left(p(\lambda_\mathrm{I} \mid \mathbf{X}'_\mathrm{I});\ p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I})\right) = \\ &= \int_{\lambda_\mathrm{I}} p(\lambda_\mathrm{I} \mid \mathbf{X}'_\mathrm{I}) \log\left(\frac{p(\lambda_\mathrm{I} \mid \mathbf{X}'_\mathrm{I})}{p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I})}\right) \mathrm{d}\lambda_\mathrm{I}. \end{aligned} \quad (6)$$

It can be shown that the KLD between two gamma distributions is a function of their hyperparameters

$$\begin{aligned}&\mathrm{KLD}\left(p(\lambda_\mathrm{I} \mid \mathbf{X}'_\mathrm{I});\ p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I})\right) = \\ &= a \cdot \log\left(\frac{b'}{b}\right) + \log\left(\frac{\Gamma(a)}{\Gamma(a')}\right) + b \cdot \frac{a'}{b'} \\ &\quad + (a' - a) \cdot \psi(a') \end{aligned} \quad (7)$$

where $\psi(a') = \frac{\frac{\mathrm{d}}{\mathrm{d}x}\Gamma(x)\big|_{x=a'}}{\Gamma(a')}$ is the digamma function. We evaluate (7) for each pixel in the virtual image and as a result get a pixel-wise surprise trigger.

For fast and parallel calculation of pixel-wise surprise triggers, modern graphics hardware can be used. Since common graphics APIs like Direct3D and OpenGL do not allow for the direct calculation of gamma and digamma functions, (7) has to be modified. In our pixel shader implementation, we approximate the gamma function using the Stirling series

$$\Gamma(z) \approx \sqrt{\frac{2\pi}{z}} \cdot \left(\frac{z}{e}\right)^z \cdot \exp\left(\frac{1}{12z} - \frac{1}{360z^3} + \frac{1}{1260z^5}\right) \quad (8)$$

where $e = 2.71828\ldots$ is the Euler's number.

The digamma function is approximated by

$$\psi(z) \approx -\frac{1}{z} - \gamma + \sum_{n=1}^{5}\left(\frac{1}{n} - \frac{1}{z+n}\right) \quad (9)$$

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

VI. EXPERIMENTAL RESULTS

In this section, we show some test results of our visual navigation algorithm and the visual output obtained from our image-based modeling technique applied to a household scene. We further tested our methods for visual homing and surprise detection. Fig. 3 shows the acquisition of an image sequence S_1 with a stereo camera head (640x480 pixels) mounted on a Pioneer 3-DX robot. The robot went along an approximately circular trajectory around a table set with household objects like glasses, plates etc. The stereo camera was looking towards the objects and captured 213 pairs of images.

The visual localization of image sequence S_1 in a world coordinate frame is illustrated in Fig. 4.

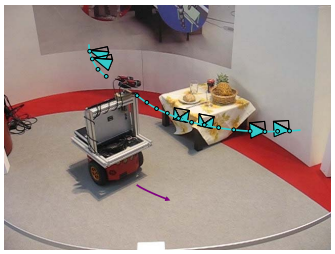


Fig. 3. Acquisition of a set of images with a stereo camera head mounted on a Pioneer 3-DX.

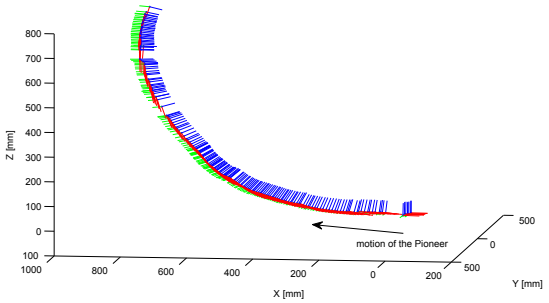
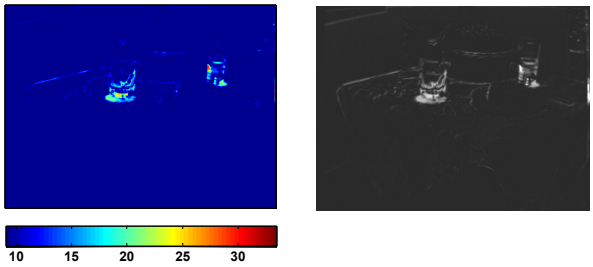


Fig. 4. 3D-plot of the localization results, which were used for the model generation in Section IV. The blue lines show the viewing direction of the capturing camera. The pioneer was programmed to go along a quarter circle around the scene.

(a) (b)

Fig. 5. (a) Observation of the cognitive system. (b) Virtual image rendered from a set of reference images from S_1 at the current position of the observing camera.

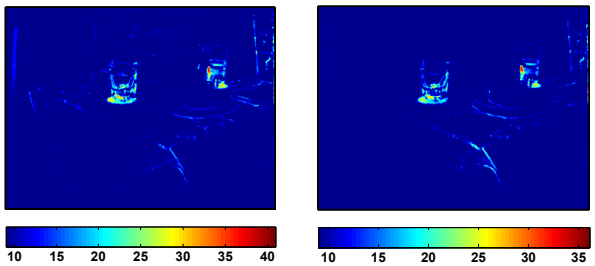


(a) (b)

Fig. 6. (a) Surprise trigger obtained from the pixel-wise calculation of the KLD between prior and posterior distribution over the precision of the color samples. (b) Approximated surprise trigger computed on the graphics hardware.

In order to test our algorithm for surprise detection we captured another image sequence S_2 on a trajectory which was close to the first one but not identical. We changed the scene before by removing the two glasses. The task of the cognitive system is to detect these changes. This is usually quite challenging for an artificial cognitive system due to the difficulties involved with building up an internal representation of the glasses. One image from S_2, which is the current observation of the cognitive system, was localized with respect to the world coordinate system of S_1. We "manually" looked for a similar image from S_1 with known pose and estimated the observation's pose with respect to it using the method from Section II. The observation is depicted in Fig. 5(a) together with a photorealistic virtual image rendered from reference images which were selected only from S_1 (Fig. 5(b)). The virtual image was rendered with our method described in Section IV. Note that there is no real camera image from S_1 which was acquired exactly at the position of the observation. Applying our algorithm from Section V on the luminance signals of the two images, we obtained the surprise trigger shown in Fig. 6(a). The surprise trigger was calculated in MATLAB using (7). The figure clearly shows a region of high KLD values around the missing glasses.

With the approximations in (8) and (9), we obtained the surprise trigger in Fig. 6(b) which was calculated on the graphics hardware by a pixel shader implemented in Direct3D. For better visualisation the surprise trigger was amplified by a factor of 10. For a static observation, we measured an average frame rate of 14 frames per second (at a resolution of 320×240 pixels). Fig. 7 shows the results for surprise detection when comparing the two strategies for the homing problem. Since the pose is not that accurate in case of automatic localization, the surprise trigger is higher in regions where indeed no changes occured compared to Fig. 6(a). However, there is still a pronounced region around the missing glasses with high surprise trigger compared to the rest of the surprise map.



(a) (b)

Fig. 7. Bayesian surprise trigger: (a) The cognitive system automatically localizes itself using the Homing1 algorithm described in Section III-A. (b) Automatic localization with the Homing2 algorithm presented in Section III-B. Even if the results for the Homing2 algorithm seem to be more accurate, the Homing1 method is preferable due to its robustness.

Furthermore, we evaluated our algorithm for surprise detection using an image-based representation of a metallic workpiece which we acquired in a factory environment. A

dense set of images was taken of the workpiece with a stereo camera mounted on an industrial robot. The robot's arm was controlled in a way that the stereo camera moved along a zigzag path across a part of a spherical surface. In Fig. 8, a comparison is shown between simple differencing, a method which is widely used in image change detection due to its simplicity, and our Bayesian approach for surprise detection. The image in Fig. 8(d) was obtained by calculating the pixel-wise difference between the observation (Fig. 8(a)) and the virtual image (Fig. 8(b)). Although the virtual image appears visually correct, the predicted colors are different from the color values captured by a real camera. The reason for this are errors in the depth maps which occur during depth estimation due to the non-Lambertian metallic surface of the workpiece. Obviously, the simple differencing method shows large differences between observation and rendered image at sites where the workpiece actually has not changed. Our Bayesian approach for surprise detection (Fig. 8(c)), in contrast, only indicates the missing part of the workpiece.

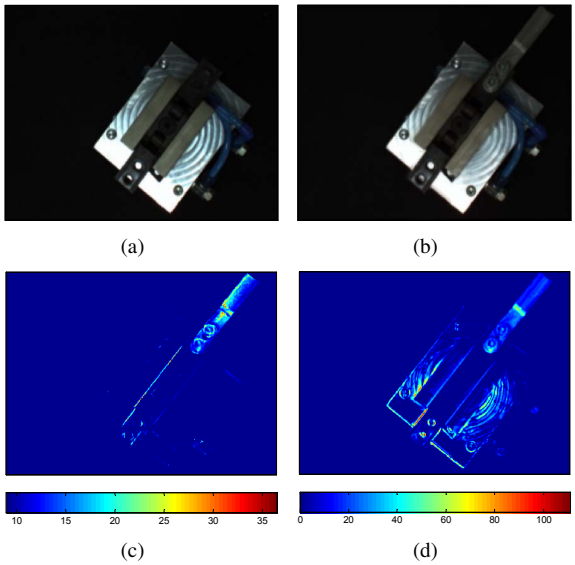


Fig. 8. (a) Observation made of an incomplete workpiece. (b) Virtual image predicted from reference images of the error-free workpiece. The surprise trigger computed with our proposed method (c) detects the actual changes reliably while the simple differencing method (d) indicates changes along the surface of the workpiece which are due to intensity variations between different viewpoints.

VII. CONCLUSION AND FUTURE WORK

In this work, we presented an approach for visual surprise detection from image-based representations of a cognitive system's environment. Bayesian probabilistic inference allows computing pixel-wise surprise triggers. Experimental results show that our proposed method outperforms existing methods like simple differencing with respect to the reliability of the indicated changes. The approximated surprise trigger calculated by the graphics hardware still sufficiently indicates unexpected changes. Accurate self-localization of mobile cognitive systems in their environment tackles the well-known homing problem and is crucial for robust surprise detection. We proposed two solutions for the homing problem which show acceptable results.

Our future research work will focus on the segmentation of environments into static and dynamic objects. Our algorithm for surprise detection should contribute to the generation of ontologies for an understanding of the environment and execution of tasks on higher cognitive levels.

VIII. ACKNOWLEDGEMENT

We thank the German Aerospace Center (DLR) for making available a Pioneer 3-DX for our experiments. Furthermore, we would like to thank Dr. Frank Wallhoff, Alexander Bannat and Jürgen Gast for their assistance during the acquisition of the image sequence with the JAHIR industrial robot.

REFERENCES

[1] H.-Y. Shum, S.B. Kang and S.-C. Chan, "Survey of Image-Based Representations and Compression Techniques," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 11, 2003, pp. 1020-1037.

[2] P. Felsenszwalb and D. Huttenlocher, "Efficient Belief Propagation for Early Vision," *in Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2004, pp. I261-I268.

[3] M. Garland and P.S. Heckbert, "Simplifying Surfaces with Color and Texture Using Quadric Error Metrics," *in Proc. IEEE Conf. on Visualization*, Durham, USA, 1998, pp. 263-269.

[4] L. Itti and P. Baldi, "Bayesian Surprise Attracts Human Attention," *in Adv. in Neural Information Processing Systems*, vol. 19, 2006, pp. 547-554.

[5] 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*, 1981, pp. 674-679.

[6] C. Tomasi and T. Kanade, "Detection and Tracking of Point Features," *in Carnegie Mellon University Technical Report CMU-CS-91-132*, April 1991.

[7] J. Shi and C. Tomasi, "Good Features to Track," *in Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 1994, pp. 593-600.

[8] R. Möller and A. Vardy, "Local Visual Homing by Matched-Filter Descent in Image Distances," *in Biological Cybernetics*, vol. 95, no. 5, 2006, pp. 413-430.

[9] W. Stürzl and H.A. Mallot, "Efficient visual homing based on Fourier transformed panoramic images," *in Robotics and Autonomous Systems*, vol. 54, no. 4, 2006, pp. 300-313.

[10] A. Vardy and R. Möller, "Biologically plausible visual homing methods based on optical flow techniques," *Connection Science*, vol. 17, 2005, pp. 47-89.

[11] A. Vardy, "Low-level visual homing," *in Advances in Artificial Life - Proc. of the 7th European Conf. on Artificial Life (ECAL)*, 2003, pp. 875-884.

[12] R.J. Radke, S. Andra, O. Al-Kohafi and B. Roysam, "Image Change Detection Algorithms: A Systematic Survey," *IEEE Trans. on Image Processing*, vol. 14, no. 3, 2005, pp. 294-307.

[13] H. Bay, T. Tuytelaars and L. Van Gool, "SURF: Speeded Up Robust Features," *in Proc. of the Ninth European Conf. on Computer Vision*, May 2006, pp. 404-417.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *in Int. Journal of Computer Vision*, vol. 60, no. 2, 2004, pp. 91-110.

[15] K. S. Arun, T. S. Huang and S. D. Blostein, "Least-squares fitting of two 3-D point sets," *in IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 9, no. 5, 1987, pp. 698-700.

[16] S. Engel and X. Zhang, "Colour Tuning in Human Visual Cortex Measured with Functional Magnetic Resonance Imaging," *Nature*, vol. 388, no. 6637, 1997, pp. 68-71.

[17] L. Itti and P. Baldi, "A Principled Approach to Detecting Surprising Events in Video," *in Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, San Diego, USA, 2005, pp. 631-637.

[18] D. Burschka and G. D. Hager, "V-GPS - Image-Based Control for 3D Guidance Systems," *in Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, October 2003, pp. 1789-1795.