Surprise Detection and Visual Homing in Cognitive Technical Systems

Werner Maier, Elmar Mair, Darius Burschka and Eckehard Steinbach

Abstract— **One important feature of a cognitive system is to interact with its environment and to react in real-time to changes in it. A pure geometric representation of the world is therefore often insufficient. To allow a cognitive system to refer to a prior perception, it has to be able to re-localize itself with respect to its environment. Thus, a vision based agent needs to register the currently seen images within its history. In this paper, we investigate not only an image based surprise trigger, which uses Bayesian probabilistic inference techniques to detect pixel-wise changes, but also an online image based homing algorithm. This is necessary to achieve a location independent surprise trigger. Our algorithms allow a cognitive system to localize itself within its environment and to react online to changes respectively to a prior measurement using only a calibrated camera. Experiments show acceptable results in terms of a robust detection of unexpected changes in the environment.**

I. INTRODUCTION

Cognitive systems need to be aware of their environment in order to react to changes, to adapt their behavior or just to keep their environment model up to date. Geometric models only allow recognition of changes in the 3D structure in the environment. What happens if only the appearance, like the color of an object changes or a glass of water is not in its place anymore or if an object is too thin to be recognized? Application examples regarding such situations would be a human environment, where a household robot has to recognize if the towels are changed or a bottle has been removed, or also a factory, where damages on the surface of a workpiece have to be recognized from an arbitrary viewing position. In such cases it is useful to have a model of the world which contains information about its appearance. Besides, in order to detect missing objects it is often easier to make use of the visual appearance of the scene than of the geometry model, since this reduces the search for changes from a 3D space to a 2D space. Common computer graphics techniques achieve fairly realistic images of simple virtual environments by mapping textures onto the triangles of a mesh-based geometry. However, it requires sophisticated and computationally expensive methods like raytracing to model translucent or reflective objects in a realistic way. Effects like the refraction of light rays are crucial for checking if they are present or not, if they are filled or empty.

Image-based rendering techniques ([1]) turned out to be very suitable for the photorealistic modeling of such objects because computational complexity does not depend on the properties of the environment. In our approach, we use a densely acquired set of images together with approximate 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. It is not the scope of this paper to explain these underlying algorithms in detail, nevertheless, we will give brief introductions at the appropriate places to facilitate the understanding.

If mobile robots can localize themselves within an image sequence, it does not mean that they can refer to a prior perception. A relation to the history has to be found to provide the integration of the current sensor data. In other words, a cognitive system, which should be able to recognize changes in the environment over time, has to be able to register itself with respect to a specified coordinate frame. A detailed classification of visual intensity based Homing algorithms can be found in [14]. In [16] the efficiency of various methods are compared. Most of them are biologically inspired, like [15] or [17]. We will present two structure-based “snapshot” approaches, where the structure is based on images again. These methods allow a cognitive system to register within a partially seen environment.

If changes in the environment are unexpected, cognitive technical systems have to react and adapt their current action plans. In [3], it was 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 ([2]). 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. Section II proposes two online solutions to the Homing problem. In Section III, our novel method for visual surprise detection, which is based on a probabilistic approach for image-based view synthesis, is presented. Before we conclude this work,

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

E. Mair and D. Burschka are with the Department of Informatics, Technische Universität München, 85748 Garching, Germany elmar.mair@cs.tum.edu, burschka@cs.tum.edu

W. Maier and E. Steinbach are with the Department of Electrical Engineering, Technische Universität München, 80333 München, Germany werner.maier@tum.de, eckehard.steinbach@tum.de

we present in Section IV experimental results and outline the integration of our module into the demonstration scenarios envisioned in the cluster of excellence CoTeSys.

II. VISUAL HOMING

The first step in the generation of image-based models is the accurate localization of the captured images. Our real-time capable algorithm presented in [6] allows us to estimate the position and orientation of the 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. In order to localize itself in the world from an image sequence, the robot has to first recognize the motion of the world in that image stream. This is done by automatically tracking features in the left images of the captured stereo pairs with the Kanade-Lucas-Tomasi (KLT) tracker ([7] [8] [9] [10]), which fits best our requirements in terms of application, speed and robustness. The right image of the stereo system is only used to recover the exact scale of the features in 3D space using the intrinsic and extrinsic calibration data. The spatial dimensions of the features are important to estimate the exact scale of camera translation.

While our localization method provides acceptable results with respect to position and orientation of the capturing camera for image-based model generation, it fails as soon as the robot looses the tracked set of features. This may happen if the acceleration of the camera is too high, so that the tracker looses all its references, or the robot is simply switched off. Even if all features were saved on hard drive over the whole time, the cameras could not be registered within the prior world coordinate frame, as soon as the robot has been moved outside the known trajectory. We need to register the new sequence with respect to the old origin. 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 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 how this problem can be solved.

Let us assume that we have done a first run, where we retrieved an image sequence and now we want to make a second run, but with the camera poses respectively 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 image 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 S_1, which means that the localization in S_2 is expressed in the coordinate frame of S_1, the so called reference frame. The first image in S_2 is called $I_{2.1}$ and the goal is to register 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 affected by an arbitrary 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 known and used is probably SIFT ([12]). Its largest disadvantage is the speed. SIFT uses complex detection functions and large descriptor vectors which make it independent of any affine transformation, but slows down the whole algorithm. An alternative, newer approach is SURF ([11]), which is supposed to be even more accurate, more robust and faster than SIFT. Therefore, we use SURF to find correspondences between images, which show the same scene, but from an arbitrary view point. SURF and KLT use different detectors, hence the stereo-registration method used for the visual localization can not be used between $I_{1.1}$ and $I_{2.1}$.

A. Three image based Homing

In our first Homing algorithm (further on Homing1) we use the same algorithm for extrinsic parameter estimation as we use within the visual localization module. This technique is called RVGPS and is an iterative method to estimate the transformation matrix between two sets of vectors. 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, as snapshot. To initialize the POIs, a second image in S_1, $I_{1.2}$, and the transformation matrix between this image and $I_{1.1}$ are necessary. The extrinsic parameters and the SURF correspondences between these two images are used to determine the 3D-structure for the SURF POIs extracted from $I_{1.1}$ by simple triangulation. Once these features are initialized, we only need at least 3 SURF correspondences between $I_{1.1}$ and $I_{2.1}$ to apply RVGPS for motion estimation. Of course, the robustness and accuracy increases rapidly if you have more matching features. Thus, big parts of the same scene should be seen by these three images to ensure that enough matches are found. Otherwise you can also use more than one image in S_1 to determine the three dimensions for more POIs in $I_{1.1}$. The more points are initialized, the higher is the probability that you find correspondences within $I_{2.1}$ and the higher is also the accuracy of the motion estimation. Figure 1 illustrates the principle of the Homing1 algorithm.

Figure 2 shows the SURF matches needed for the Homing1 algorithm. Figure 2(a) contains the correspondences between $I_{1.1}$ and $I_{1.2}$, which are required to initialize the point structure in the reference frame $I_{1.1}$. Figure 2(b) represents the matches between $I_{1.1}$ and $I_{2.1}$. These correspondences are used to estimate the transformation of $I_{2.1}$ to $I_{1.1}$ using RVGPS.

B. Four image based Homing

The Homing1 variant has shown that its results depend strongly on the accuracy of the POIs’ structure. Our second approach has been developed with the aim to avoid that lack by not using RVGPS, but an optimal matching of the two 3D structures in the different coordinate frames. To calculate two

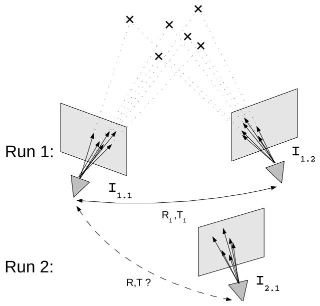
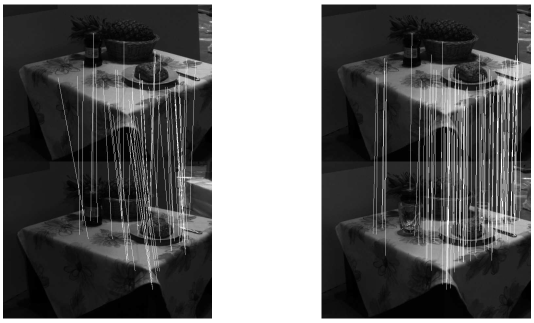


Fig. 1. Visual Homing out of 3 images. The transformation between sequence 1 and 2 is estimated using the RVGPS algorithm. Thus, only one image of run 2 is necessary.



(a) $I_{1.1}$ and $I_{1.2}$

(b) $I_{1.1}$ and $I_{2.1}$

Fig. 2. These figures show the SURF matches for the Homing1 algorithm. The result is used for the surprise trigger in the left image of figure 10. 33 common correspondences in all 3 images were found.

corresponding structures for our second Homing algorithm (Homing2) we need for each sequence S_1 and S_2 two images, their transformation matrices and the SURF matches in all 4 images. We initialize the structure for an image in each sequence, like we did for S_1 in the Homing1 alternative. Using the so called Arun's algorithm ([13]) we get the transformation matrix between the two domains. 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 SURF matches within 4 images, which is quite difficult to achieve. Figure 3 depicts the principle of the Homing2 algorithm.

In Figure 4 you can see the correspondences needed for the Homing2 algorithm. Figure 4(a) shows the matches between $I_{1.1}$ and $I_{1.2}$, which are needed for the initialization of the point structure in the reference frame $I_{1.1}$. Figure 4(c) represents the matches between $I_{2.1}$ and $I_{2.2}$, which are needed for the initialization of the point structure in the S_2 domain ($I_{2.1}$). Figure 4(c) illustrates the matches between $I_{1.1}$ and $I_{2.1}$, which are finally used to find the correspondences between S_1 and S_2. Applying now Arun's algorithm both domains can be merged.

Which algorithm to use depends therefore strongly on the application and the scene. Since the errors do not vary much (compare the subfigures in Figure 10), mostly the more robust but less accurate Homing1 algorithm is preferable.

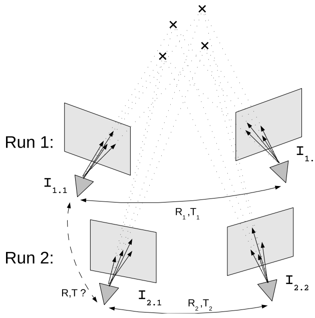
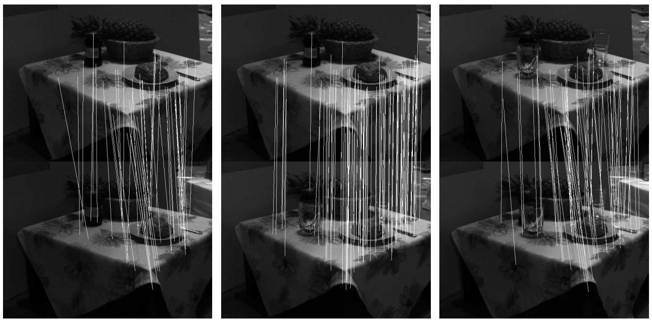


Fig. 3. Visual Homing out of 4 images. The point structure in run 2 is initialized independently of the reference sequence (run 1), so that a higher accuracy is provided at the cost of the robustness (usually less common features are found).



(a) $I_{1.1}$ and $I_{1.2}$

(b) $I_{1.1}$ and $I_{2.1}$

(c) $I_{2.1}$ and $I_{2.2}$

Fig. 4. These figures show the SURF matches for the Homing2 algorithm. The result is used for the surprise trigger in the right image of figure 10. Only 10 common correspondences in all 4 images could be found.

III. SURPRISE DETECTION

In order to predict novel views from an acquired set of reference images, correspondences between the pixels have to be known apart from the position and orientaton of the capturing camera. Hence, in a preprocessing step for view synthesis, per-pixel depth maps are computed for each acquired reference image. On-line view synthesis is done by selecting a small number of reference images (in this work seven) whose image data is warped into the virtual camera, respectively ([6]). Thus, for a given pixel in the virtual image there are usually several color samples giving hint about the true color value which would be captured with a real camera at that position. In order to model the uncertainty about the true color value, we assume that the warped color samples from the reference views are independently drawn from a Gaussian distribution whose mean is identical to the true color value. By maximum-likelihood (ML) estimation, the mean is retrieved from the sample data and written to the given pixel in the virtual image.

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 [3], 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. 5, 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 ([4]), 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:

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

$\mathbf{X}_\mathrm{I} = [x_{\mathrm{I},1}, \ldots, x_{\mathrm{I},7}]$ 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

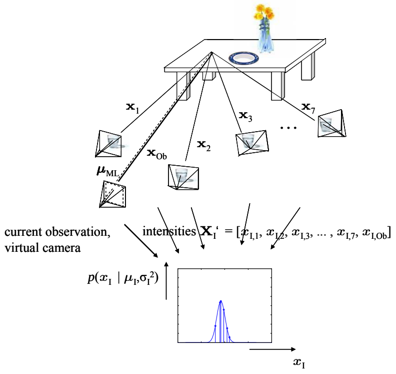
$$p(\lambda_\mathrm{I}) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda_\mathrm{I}^{a_0-1} \exp\left\{-b_0 \lambda_\mathrm{I}\right\}. \quad (2)$$

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

$$\begin{aligned} p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I}) &\propto p(\mathbf{X}_\mathrm{I} \mid \mu_{\mathrm{I,ML}}, \lambda_\mathrm{I}) \cdot p(\lambda_\mathrm{I}) = \quad (3) \\ &= \lambda_\mathrm{I}^{a_0-1} \lambda_\mathrm{I}^{\frac{7}{2}} \exp\left\{-b_0 \lambda_\mathrm{I} - \tfrac{\lambda_\mathrm{I}}{2} \textstyle\sum_{k=1}^{7} \left(x_{\mathrm{I},k} - \mu_{\mathrm{I,ML}}\right)^2\right\} \end{aligned}$$

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.



$\mathbf{x}_1$
$\mathbf{x}_{\mathrm{Ob}}$
$\mathbf{x}_2$
$\mathbf{x}_3$
$\mathbf{x}_7$
…
$\boldsymbol{\mu}_{\mathrm{ML}}$
current observation,
virtual camera
intensities $\mathbf{X}_\mathrm{I}$' = $[x_{\mathrm{I},1}, x_{\mathrm{I},2}, x_{\mathrm{I},3}, \ldots, x_{\mathrm{I},7}, x_{\mathrm{I,Ob}}]$
$p(x_\mathrm{I} \mid \mu_\mathrm{I}, \sigma_\mathrm{I}^2)$
x_I

Fig. 5. A dissimilarity between the color value of the current observation and the warped colors from previously acquired reference images leads to a surprise trigger. With the data obtained from the homing algorithms, the virtual camera is placed at the current position of the cognitive system's camera.

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

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

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 [5], 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

$$\begin{aligned} &\mathrm{KLD}\left(p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I}');\ p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I})\right) = \\ &= \int_{\lambda_\mathrm{I}} p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I}') \log\left(\frac{p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I}')}{p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I})}\right) \mathrm{d}\lambda_\mathrm{I}. \quad (5) \end{aligned}$$

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

$$\begin{aligned} &\mathrm{KLD}\left(p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I}');\ p(\lambda_\mathrm{I} \mid \mathbf{X}_\mathrm{I})\right) = \\ &= a \cdot \log\left(\frac{b'}{b}\right) + \log\left(\frac{\Gamma(a)}{\Gamma(a')}\right) + b \cdot \frac{a'}{b'} \\ &\quad + (a' - a) \cdot \psi(a') \quad (6) \end{aligned}$$

where $\psi(a') = \frac{\frac{\mathrm{d}}{\mathrm{d}x}\Gamma(x)\big|_{x=a'}}{\Gamma(a')}$ is the digamma function. We evaluate (6) for each pixel in the virtual image and get so a pixel-wise surprise trigger.

IV. INTEGRATION IN DEMONSTRATION SCENARIOS

In principle, our module for visual camera localization and image-based environment modeling can be integrated in any demonstration scenario in CoTeSys. In this section, we show a possible episode for the assistive household scenario, together with experimental results and give an outlook for the integration within JAHIR.

A. Assistive Household Scenario

The episode we envision within the assistive household scenario is the acquisition of an image-based model of a typical household environment which is the basis for cognitive processes. With our module a cognitive robot in the household should be able to update its environment model at any time by surprise detection and classify the objects around it into static or dynamic ones. Moreover, surprise about unexpected events should influence the robot's action plans. Robust visual localization should enable it to retrieve its current position in the environment and to evaluate its current observation with respect to the already acquired reference model. Fig. 6 shows the acquisition of an image sequence S_1 with a stereo camera head (640x480 pixels) mounted on a Pioneer 3-DX robot during AUTOMATICA 2008. The robot went along an approximately circular trajectory around a table set with household objects like glasses, plates etc. with the stereo camera looking towards the objects and capturing 213 pairs of images. The set of images was subsampled by a factor of two.

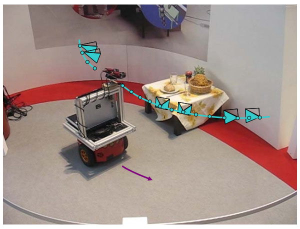


Fig. 6. Acquisition of an image sequence with a stereo camera head mounted on a Pioneer 3-DX.

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 by removing the two glasses. The task of the cognitive system is to detect these changes. One image of the second set, which is the current observation of the cognitive system, was localized with respect to the world coordinate system of the first set. For pose estimation we "manually" looked for a similar image from the first set. This image is depicted in Fig. 7 (left) together with a photorealistic virtual image rendered from reference images which were selected only from the first set of images (right). Note that there is no real camera image from the first set which was acquired exactly at the position of the observation.

Applying our algorithm from Section III on the luminance signals of the two images, we obtained the surprise trigger shown in Fig. 8 (left). The figure clearly shows a region of high KLD values around the missing glasses. The right part of Fig. 8 shows the pixel-wise absolute difference between the two luminance signals, a method which is still widely used in image change detection. Obviously, our method behaves more robust around the knife and at the edge of the table where the two images are not identical due to geometry and pose inaccuracies.

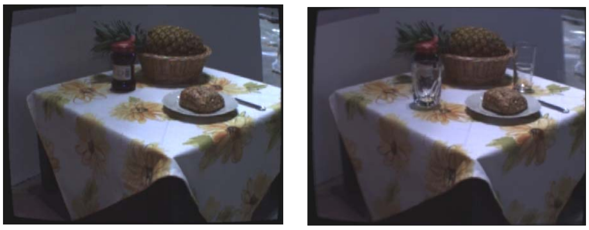


Fig. 7. Observation of the cognitive system (left) and virtual image rendered from the set of reference images (right) at the current position of the observing camera.

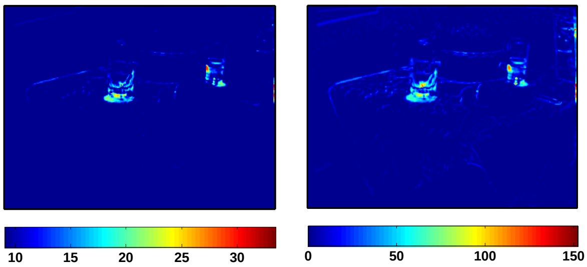


Fig. 8. (Left) Surprise trigger obtained from the pixel-wise calculation of the KLD between prior and posterior distribution over the precision of the color samples. (Right) Surprise trigger obtained from simple differencing.

Among the luminance and the color opponency pathways we only obtained a significant surprise trigger for the luminance since the glasses do not convey much color information, as Fig. 9 shows.

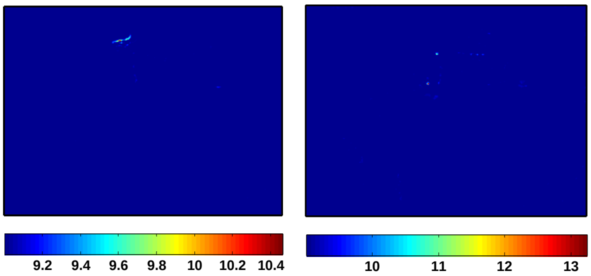


Fig. 9. (Left) Surprise trigger calculated in the red-green opponency pathway. (Right) Surprise trigger obtained in the blue-yellow pathway. Both figures show a much lower surprise trigger around the glasses than in the luminance pathway.

Fig. 10 shows our results for surprise detection applying the two strategies for the homing problem described in Section II. 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. 8 (left). However, in the region around the missing glasses, the surprise trigger is still much higher than in the rest of the surprise map.

B. JAHIR / Cognitive Factory

Further integration of our module into the demonstration scenario cognitive factory is planned in terms of the JAHIR project. An episode which demonstrates joint action is planned in JAHIR. The goal is that a robot mounts a product while communicating with a human worker by speech and

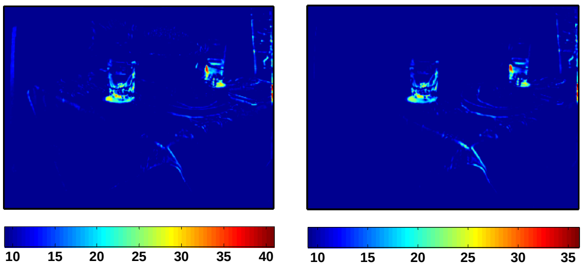


Fig. 10. Bayesian surprise trigger: The cognitive system automatically localizes itself using the Homing1 algorithm described in section II-A (left) and the Homing2 algorithm presented in section II-B. Even if the results for the Homing2 algorithm seem to be more accurate, the Homing1 method is preferable, due to its robustness.

virtual buttons which are projected onto the working table. We are going to integrate our algorithms described in this work during the quality assessment step after the product is mounted. To this end, an image sequence of the mounted error-free product is first captured which is processed for a reference image-based model. Our algorithm for surprise detection detects unforeseen changes in the product which might be due to failures in the single production steps. It is desirable that the inspecting camera is localized with respect to the product and not with respect to the robot's coordinate system since this does not require that the product is exactly at the same position as during the acquisition of the reference model. We plan to solve this issue with our homing algorithms. The cognitive robot uses this surprise trigger for making decisions about repair strategies and the next production steps. Fig. 11 shows a pair of stereo images captured from a workpiece with the JAHIR robot. The left image also shows features extracted with our algorithm for localization.

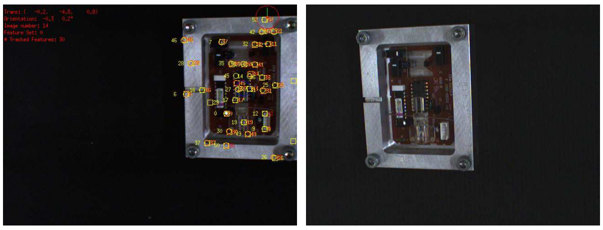


Fig. 11. A stereo image pair of an image sequence acquired using the JAHIR robot. On the left image also the tracking and visual navigation results are displayed.

V. CONCLUSION AND FUTURE WORK

In this work, we presented an approach for visual surprise detection which is based on image-based models of a cognitive system's environent. Bayesian probabilistic inference allows computing pixel-wise surprise triggers which give hints about unexpected changes in dynamic environments. Experimental results show that this method provides more robust results than simple differencing. 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. Furthermore, we outlined the possible integration of our work into the demonstration scenarios in CoTeSys.

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 a understanding of the environment and execution of tasks on higher cognitive levels. We plan to further increase the accuracy of our homing algorithms such that robust and reliable surrpise triggers can be generated.

VI. ACKNOWLEDGEMENT

We thank the German Aerospace Center (DLR) for making available a Pioneer 3-DX for our experiments. We also thank Alexander Bannat, Jürgen Gast and Dr. Frank Wallhoff (TUM) from the JAHIR team for their cooperation in terms of the acquisition of image sequences with the JAHIR 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] 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.

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

[4] 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.

[5] 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.

[6] E. Mair, W. Maier, D. Burschka and E. Steinbach, "Image-Based Environment Perception for Cognitive Technical Systems", submitted to *1st International Workshop on Cognition for Technical Systems*, Munich, Germany, 2008.

[7] Bruce D. Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. *in International Joint Conference on Artificial Intelligence*, pages 674-679, 1981.

[8] Carlo Tomasi and Takeo Kanade. Detection and Tracking of Point Features. *in Carnegie Mellon University Technical Report CMU-CS-91-132*, April 1991.

[9] Jianbo Shi and Carlo Tomasi. Good Features to Track. *in IEEE Conference on Computer Vision and Pattern Recognition*, pages 593-600, 1994.

[10] Stan Birchfield. Derivation of Kanade-Lucas-Tomasi Tracking Equation. *Unpublished*, May 1996.

[11] Herbert Bay, Tinne Tuytelaars, Luc Van Gool, "SURF: Speeded Up Robust Features", *in Proc. of the ninth European Conference on Computer Vision*, May 2006.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *in IJCV*, 60(2):91110, 2004.

[13] Arun, K. S. and Huang, T. S. and Blostein, S. D.. Least-squares fitting of two 3-D point sets. *in IEEE Trans. Pattern Anal. Mach. Intell.*, pages 698-700, 1987.

[14] Möller R., Vardy A.. Local visual homing by matched-filter descent in image distances. *in Biological cybernetics*, vol. 95, n. 5, pp. 413-430, 2006.

[15] Stürzl W., Mallot H.A.. Efficient visual homing based on Fourier transformed panoramic images. *in Robotics and Autonomous Systems*, 54:300-313,2006.

[16] Vardy A., Möller R.. Biologically plausible visual homing methods based on optical flow techniques. *Connection Science*, 17:47-89, 2005.

[17] Andrew Vardy. Low-level visual homing. *in Advances in Artificial Life - Proceedings of the 7th European Conference on Artificial Life (ECAL)*, 875-884, 2003.