Bayesian Visual Feature Integration with Saccadic Eye Movements

Kai Welke, Tamim Asfour and Rüdiger Dillmann
University of Karlsruhe (TH), IAIM, Institute for Anthropomatics
P.O. Box 6980, 76128 Karlsruhe, Germany

{welke,asfour,dillmann}@ira.uka.de

***Abstract*— In order to allow humanoid robots to operate in unstructured environments, behaviors have to be implemented that support exploration of the evironment. In active visual perception, saccadic eye movement is such a behavior that supports the exploration of salient locations within the current scene in a sequential manner.**

The proposed work deals with the integration of visual features extracted at different gazes during saccades executed on an active humanoid head. Using probabilistic methods to account for uncertainties during execution and perception, visual stimuli are integrated in an ego-centric representation. The resulting map stores the regarded stimuli in a consistent fashion. The approach is evaluated using three common types of feature extraction methods.

I. Introduction

The ability to investigate and explore the environment is a crucial task for humanoid robots to become part of our daily life. The integration of sophisticated sensor systems in humanoid platforms allows to study and establish methods that exploit these sensor systems in order to provide a more stable perception.

The work presented here focuses on aspects of visual perception using an active camera system, namely the Karlsruhe Humanoid Head [1]. Most current humanoid robots are equipped with simplified eye-head systems having a small number of degrees of freedom (DoF). The heads of ASIMO [2], HRP-3 [3] and HOAP-2 [4] have two DoF and fixed eyes. In contrast, the Karlsruhe Humanoids Head offers 7 DoF including separately actuated eyes.

Generally, there are two different strategies for controlling the gaze of such an active system: smooth pursuit and saccadic eye movement. In smooth pursuit, a collection of previously perceived stimuli is fixated and centered in the cameras using closed-loop control [5]. While smooth pursuit is the strategy of choice once the agent focuses on an object or interacts with a person, exploratory behavior requires the generation of rapid gaze shifts to only roughly visually perceived stimuli or stimuli from different modalities. These so-called saccadic eye movements allow to implement mechanisms such as attention on humanoid systems. While performing a saccades, the scene is perceived from different viewpoints, each providing distinct visual information. The perception of the scene from different gaze directions augments the visual field of the system.

In contrast to smooth pursuit strategies, saccadic eye movements require a model of the kinematic structure in order to map locations of salient stimuli to the corresponding motor commands. The exact kinematic model of the Karlsruhe Humanoid Head is not known completely from CAD, since the position of the optical centers of the camera system cannot be determined a-priori. In order to derive an approximated model, a visually supported kinematic calibration procedure is performed off-line. Using the calibrated model, saccadic eye movements could be executed with satisfactory accuracy [6]. In order to derive a consistent scene representation, the relation between different gaze directions has to be determined. While the accuracy of the approximated kinematic model is sufficient to implement saccadic eye movements on the head, the perception itself is much more prone to inaccuracies in the underlying transformation. Since both eyes of the head can move independently around their pan axis, small inaccuracies lead to large errors when using stereo vision methods.

In this work we present an approach which allows to integrate visual stimuli from different gazes performed on an active vision system with independently actuated eyes. The proposed method takes into account uncertainties in both execution of motor commands and extraction of stimuli in order to infer a consistent model of the environment. The uncertainties are formulated in a probabilistic fashion. Bayesian methods are proposed in order to infer a map of 3D visual landmarks from the observed stimuli. The proposed method supports online mapping; visual information from each gaze is integrated into the ego-centric model of the scene. As will be shown by the experiments, a consistent map could be generated for different types of visual features.

This paper is organized as follows: The next paragraph gives an overview of related work in the field of visual perception for humanoid robots. In Section II, the components of our method are described in detail, before we present experimental results in Section III.

A. Related Work

The problem of spatial mapping of visual 3D landmarks using cameras is a well-studied research area. Depending on both the domain and the target system, various approaches have been proposed.

In the structure-from-motion (SFM) field, the goal is to determine feature locations and extrinsic camera parameters from multiple observations of a scene at different positions of the camera. Usually only one camera is accounted for. Most commonly, bundle adjustment is performed using minimzation methods ([7], [8]). In [9] SFM is applied in an attentional framework on a mobile robot platform. The approaches proposed in the SFM field are not suitable for our problem, since translational movements are necessary to derive equations that allow for a robust solution. Furthermore, for solving the minimization problem, a set of images has to be regarded at once. Hence, bundle adjustment is not the choice for on-line applications.

The field of auto-calibration or self-calibration is closely related to reconstruction from stereo images. In self-calibration one seeks to recover the fundamental matrix from observations of features in both image planes. The problem of calibrating an active stereo camera system using observations of features is still difficult. Most approaches use simplifying assumptions in the camera model ([10], [11]) or ego-motion [12] in order to derive the fundamental matrix. In our approach we do not seek to guess the exact fundamental matrix from observations. A kinematic calibration is performed offline, which provides an approximation of the fundamental matrix for different actuations of the eye system. The inaccuracy of this calibration is taken into account in terms of noise in the motion model of our approach in order to derive a consistent mapping of 3D features.

The field of visual self-localization and mapping (visual SLAM) is closely related to the presented work. Visual SLAM deals with the application of visual information for the SLAM problem [13]. A mature approach for visual SLAM called MonoSLAM has been presented in [14]. MonoSLAM allows to estimate both the motion of a single camera and a map of landmarks from multiple observations in a probabilistic framework. Several applications on robots performing ego-motion have been presented. Since translational motion of the camera is a key cue for retrieving an estimate of the map, MonoSLAM cannot be applied to our problem. Because saccadic eye movements only involve very small translational variations of the camera poses, we solve this problem by introducing an approximated model of the motion of the camera together with appropriate uncertainties. The proposed system does not seek to solve the problem of self localization itself but rather uses the approximated model of its state in order to infer the map.

II. Proposed approach

The proposed approach aims at the accumulation of visual information extracted from the perceived world during saccadic eye movements. The extraction takes place after each gaze shift; during the execution of a saccadic movement the perception is disabled. On a technical system, the interruption of perception helps to cope with motion blur and image synchronization.

Fig. 1 shows a Bayesian network that models the observed and hidden variables in the system. Given this formalization, the goal of the system is to infer the map m from the measurement z captured after each saccadic eye movement. Since the world is not static, we want be able to cope with appearing and disappearing stimuli resulting from interference of other agents. The observed state u is given by the encoder readings from the three eye joints: panning left and right as well as common tilt. The hidden variable x captures the state of the system comprising uncertainties resulting from the positioning of the eye joints.

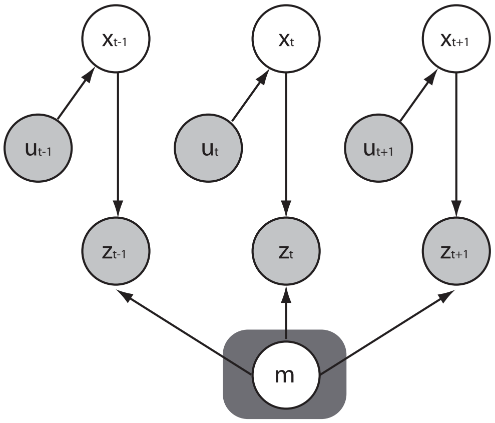


Fig. 1. Graphical model of the proposed system. From the current control u, the system state x is modeled by accounting for the dominant noises in execution and calibration. Each observation of features in the system state x results in measurements z. The goal is to infer map m while capturing measurements during multiple saccadic eye movements.

The head eye system has been visually calibrated using the approach presented in [6]. The resulting approximated kinematic model is made available for the proposed approach. As discussed below, some aspects of the calibration are modeled in both the observed state u and the system state x in order to cope with the remaining uncertainties.

In the following paragraphs we present the motion model, which captures the dominant sources of noise introduced when executing gaze shifts as well as the measurement model, which covers the uncertainties during perception in a probabilistic manner. The organization of the map m is discussed and the deployed probabilistic inference approach is presented.

A. Motion Model

In order to infer a consistent map m, all relevant uncertainties which result from the motion of the eyes have to be formulated in the motion model. In the following, we describe the sources of noise which we identified to be most dominant on our system and derive appropriate representations of the observed state u_t and the hidden state x_t. The motion model is then formulated as the conditional probability $p(x_t|u_t)$ that the system is in state x_t given the observed state u_t.

For our system we identified the following sources of noise to be most dominant:

- **Positioning noise**
  The positioning of the head is very accurate (see [1]). However, depending on the settings of the low-level controllers, small errors in positioning remain. Furthermore a small error in the conversion to joint angles from encoder values has been observed. Both errors are assumed as additive noise with Gaussian uncertainty. While the positioning noise is assumed to be constant over the complete space of joint angles, the conversion noise is modeled as a normal distribution with increasing variance proportional to the actuation of the joint. Considering joint angle

readings $\vec{\theta} = (\theta_{et}, \theta_{epl}, \theta_{epr})^T$ for the eye tilt and both eye pan joints, the uncertainty resulting from inaccurate positioning Σ_p and inaccurate conversion from encoder values to joint angles Σ_c lead to the positioning errors

$$e_{pos,p} \sim \mathcal{N}(0, \Sigma_p) \quad (1)$$

$$e_{pos,c} \sim \mathcal{N}(0, \Sigma_c \vec{\theta}), \quad (2)$$

where Σ_p and Σ_e are diagonal matrices modeling the variance for each joint independently.

- **Calibration noise**
  As aforementioned, an approximated kinematic model of the eye system has been calibrated offline. While the orientation of relevant coordinate systems can be determined very accurately during offline calibration, the exact position of the rotation axes is hard to derive from the visual calibration procedure. In order to cope with this inaccuracy, we model the calibration error e_{cal} that is induced to the position of rotation axes for different calibrations using additive Gaussian noise. The position of each rotation axis is described with its translation $\vec{r}$ relative to the last joint. The calibrated translation parameters are subsumed in $\vec{d} = (\vec{r}_{et}, \vec{r}_{epl}, \vec{r}_{epr})^T$. The uncertainty about the position of the axes is modeled using the normal distribution

$$e_{cal} \sim \mathcal{N}(0, \Sigma_{cal}). \quad (3)$$

Given the above considerations, the observed state u_t comprises the currently measured actuation of the joints $\vec{\theta}_t$ and the calibrated joint axes translations $\vec{d}$. Using the formulated errors, the conditional probability of being in the hidden state x_t given an observed state $u_t = (\vec{\theta}_t, \vec{d})^T$ can be formulated by

$$p(x_t|u_t) = u_t + \begin{pmatrix} e_{pos,p} + e_{pos,c} \\ e_{cal} \end{pmatrix}, \quad (4)$$

where all errors are assumed to be independent.

B. Map Representation

In the presented work we only consider movements of the 3 DoF of the eye system. Therefore, an ego-centric reference coordinate frame is established which corresponds to the initial position of the left camera. Our approach aims to derive a map of the environment which contains visual features in this reference frame. The resulting map expresses uncertainties about the position and the existence of features within the scene.

The deployment of a scene representation which consists of features motivates a landmark based representation of the map. Each landmark in the system consists of its 3D position $\vec{l}_n = (x, y, z)$ accompanied with the position uncertainty $\Sigma_{l,n}$. Together with the signature s_n, which describes the appearance within a defined region, and the log probability i_n for the existence of the landmark, each landmark L_n can be described by

$$L_n = (\vec{l}_n, \Sigma_{l,n}, s_n, i_n). \quad (5)$$

The map m contains an entry for each landmark

$$m = \{L_1, \cdots, L_N\}, \quad (6)$$

where N is the number of landmarks in the map.

Reviewing the graphical model in Fig. 1 the uncertainty of landmarks is independent. The uncertainty of the feature position is represented using a normal distribution. The probability of a landmark n being at a position within the map is expressed using the equation

$$p(m_n) = \mathcal{N}(\vec{l}_n, \Sigma_{l,n}). \quad (7)$$

C. Measurement Model

In the following section the measurement model for our approach is derived. Both cameras of the active vision system are modeled as a single sensor that measures 3D positions of features. As will be shown, this allows intuitive inference of 3D maps. The measurement model is defined by the conditional probability $p(z_t|x_t, m)$ that a measurement z_t is observed given the system state x_t and the map m. Let each measurement z_t be composed of K measured features $z_{k,t}$. We derive the above conditional probability for the measurement of a single feature $z_{k,t}$. First a model for the uncertainty of the 3D position resulting from inaccuracies in the 2D position of features in the image plane is derived. Then the resulting 3D localization uncertainty is expressed with respect to the system state x_t.

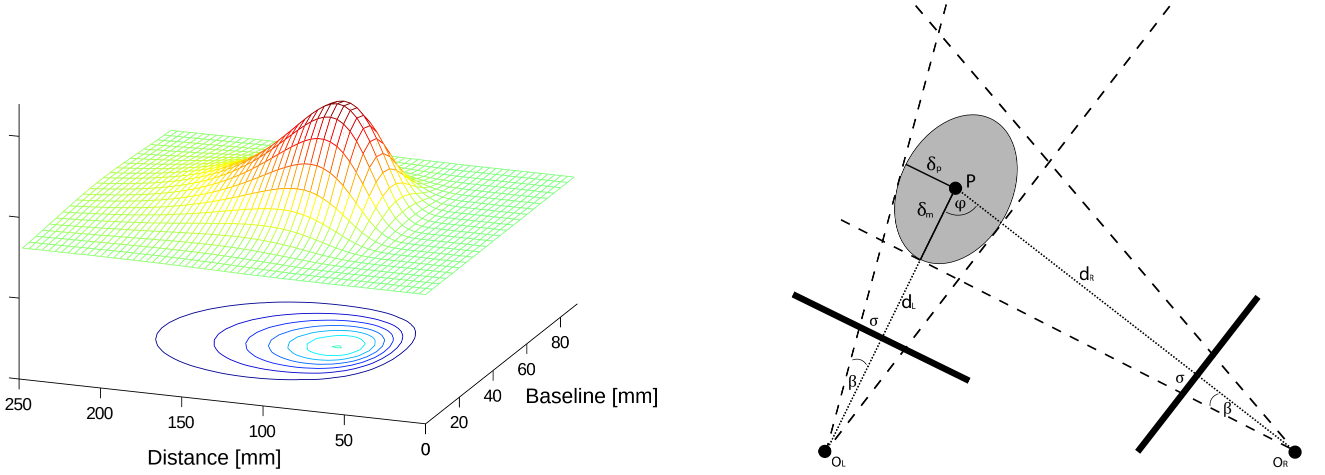
Features are extracted in the camera images, whereas each feature detector identifies positions $\vec{u}_l, \vec{u}_r$ in the left and right image plane. It is assumed that for each applied feature extraction method the uncertainty of determining its position in the image plane I can be approximated with the normal distribution

$$p(I) = \mathcal{N}\left(\vec{u}, \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}\right). \quad (8)$$

The 3D position is determined using epipolar geometry. For each feature in the left image, the epipolar line in the right image is calculated and the best match in proximity of the epipolar line is selected. The resulting uncertainty of the 3D position using epipolar geometry can not be derived in a closed form from a pair of uncertain 2D positions in the left and right camera images. Fig. 2(a) shows the true distribution of the 3D uncertainty resulting from reconstruction using uncertain 2D features according to (8). While the result is clearly not normal distributed, the Gaussian approximation still yields good results in practice ([15], [16]).

The geometry of stereo triangulation is depicted in Fig. 2(b). The dotted lines denote the points corresponding to the standard deviation σ_x of the feature localization. The distribution is the product of the uncertainties of the left and the right camera for each point in the epipolar plane. In [17] and [16], a Gaussian approximation of the resulting distribution is derived for cameras with parallel optical axis. In the following this model is adapted for actuated cameras.

The error is decomposed in two components as proposed in [18]. The pointing error δ_p is derived from the uncertainty



(a) Reconstruction uncertainty in the epipolar plane for cameras with baseline of 9cm and actuation of 25°. The plot shows the uncertainty of a point at $(45, 100)$ with a standard deviation in the image plane of $\sigma_x = 0.6mm$.

(b) Approximated 3D reconstruction error. The uncertainty ellipse is derived using the pointing error δ_p and the matching error δ_m.

Fig. 2. True distribution and proposed model of the 3D reconstruction error.

of localization in the left camera image and is assumed to be orthogonal to the projection axis $\overline{PO}_L$. The projection of the axis $\overline{PO}_L$ into the right camera (epipolar line) is the basis for the identification of matching correspondences. Consequently, the matching error δ_m is measured along the projection axis. Both errors can be derived using the following equations:

$$\begin{aligned}\delta_p &= \frac{d_L\sigma}{f}\\ \delta_m &= \frac{sin(d_r\beta)}{sin(180-\varphi-\beta)}\end{aligned}$$

The extension to the third dimension is straight forward using the standard deviation σ_y in y direction.

Using the above approximation the uncertainty of a 3D measurement is calculated in the following way:

$$\Sigma_{3D} = H_{\overline{PO}_L}\begin{pmatrix}\delta^2_{p,x} & 0 & 0\\ 0 & \delta^2_{p,y} & 0\\ 0 & 0 & \delta^2_m\end{pmatrix}H^T_{\overline{PO}_L}, \qquad (9)$$

where $H_{\overline{PO}_L}$ describes the rotation of the ellipsoid around the point P according to the projection axis in the left image and $\delta_{p,x}$ and $\delta_{p,y}$ describe the pointing error in x and y direction.

The proposed approximation of the uncertainty reflects the important property of the true distribution, that uncertainty in depth increases with decreasing angles φ between the projection axes of the left and right camera [8]. Note that we assume the camera planes to be orthogonal to the projection axes, which results in an optimistic approximation of the error.

In order to incorporate the system state x_t in our measurement model we define the functions $F_{k,t}(z_{k,t}, x_t)$ and $G_{k,t}(\Sigma_{3D,k,t}, x_t)$ which update the position $z_{k,t}$ and the uncertainty $\Sigma_{3D,k,t}$ of the k-th measurement using the calibrated model of the eye system and the current system state x_t.

Altogether, the measurement model for a single feature $z_{k,t}$ with respect to the n-th landmark and the system state x_t is given by

$$p(z_{k,t}|x_t, m_n) = det(2\pi G_{k,t})^{-\frac{1}{2}}$$
$$exp\left(-\frac{1}{2}(F_{k,t}-\vec{l}_n)^T G^{-1}_{k,t}(F_{k,t}-\vec{l}_n)\right).$$

D. Bayesian Inference

In order to infer the hidden variables m_t and x_t of our system, we use a Rao-Blackwellized particle filter approach [19]. The method we choose is a slightly adapted version of FASTSlam [20]. In the following, a short overview of the aspects specific to our approach are given. For further details, the reader is referred to the cited work.

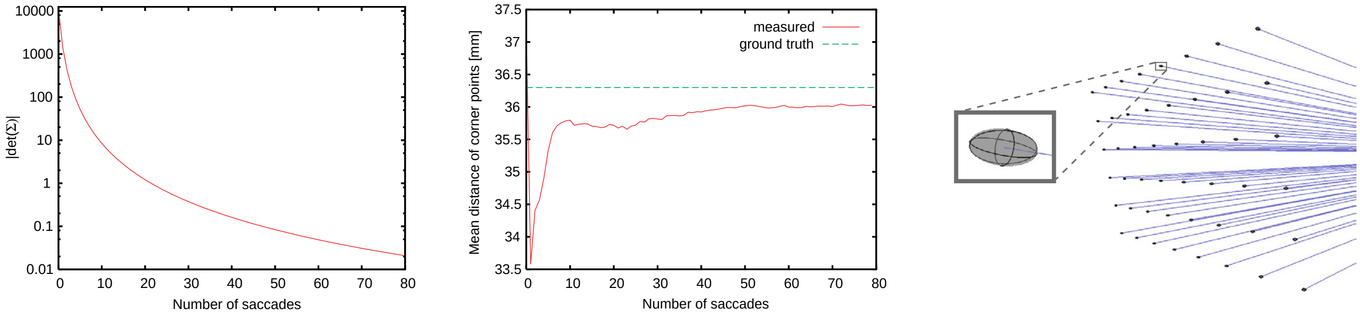
The update of the system state is performed by sampling the system state x_t according to the motion model (4). Note that for our approach we do not consider the state of the system at time $t-1$. Furthermore, each sampled particle Y^w_t contains a map representation m_w. The maps are sampled from the last particle filter iteration according to their probability $p(Y^w_t)$, which will be introduced later (13). The w-th particle is defined by

$$Y^w_t = (\vec{x}^w_t, N^w, L^w_1, \cdots, L^w_N). \qquad (10)$$

For each particle in the set of W particles, the map is updated using the current measurement z_t. As the correspondence between measured features and landmarks is not known in our approach, we solve the correspondence problem using a maximum likelihood (ML) estimator [21]. For each ML correspondence between a landmark $L^w_{n,t}$ and an observation $z_{k,t}$, if the ML probability p_n is above p_{new}, the belief is updated using a Kalman filter approach.

Let the belief of a the map m be defined by

$$bel(m) = p(m|z_{1\ldots t}, u_{1\ldots t}). \qquad (11)$$



(a) Mean of the uncertainty ellipsoid volume for all chessboard corners during 80 saccadic eye movements in log scale.

(b) Mean distance between neighbored corner points during 80 saccadic eye movements. The ground truth was measured with 36.3mm.

(c) Best map of chessboard corner features after 80 iterations.

Fig. 3. Results of the mapping of chessboard corner points after 80 saccades.

Each Kalman filter updates the belief associated with the landmark n using the Bayes update rule

$$bel(m_n^w) = \eta p(z_{k,t}|x_t^w, m_n^w,)bel(m_n^w). \qquad (12)$$

For measurements where no correspondence can be established, a new landmark L_{N+1}^w is generated using both measurement uncertainty and position corresponding to the system state stored in the particle x_t^w. The landmark probability is initialized to $p_{N+1} = p_{new}$.

Furthermore, for observed landmarks $L_{n,t}^w$ the log probability of their existence i_n^w is increased by a constant amount. For landmarks in the map that are visible but do not have a corresponding measurement, the log probability is decreased accordingly. Landmarks with $i_n^w \leq 0$ are removed from the map. In order to allow fast removal of features that disappeared from the scene, the log probability is restricted to lie below i_{max}.

For each particle, the associated probability is calculated with the following equation

$$p(Y_r^w) \propto (p_{miss})^u \prod_{n=1}^{N_{new}} p_n q(s_n), \qquad (13)$$

where p_{miss} describes the probability of not observing a landmark, u amounts to the number of visible but unobserved landmarks, N_{new} is the updated number of landmarks and $q(s_n)$ derives a probability based on a similarity metric for the stored signature s_n and the observed feature.

III. Experiments

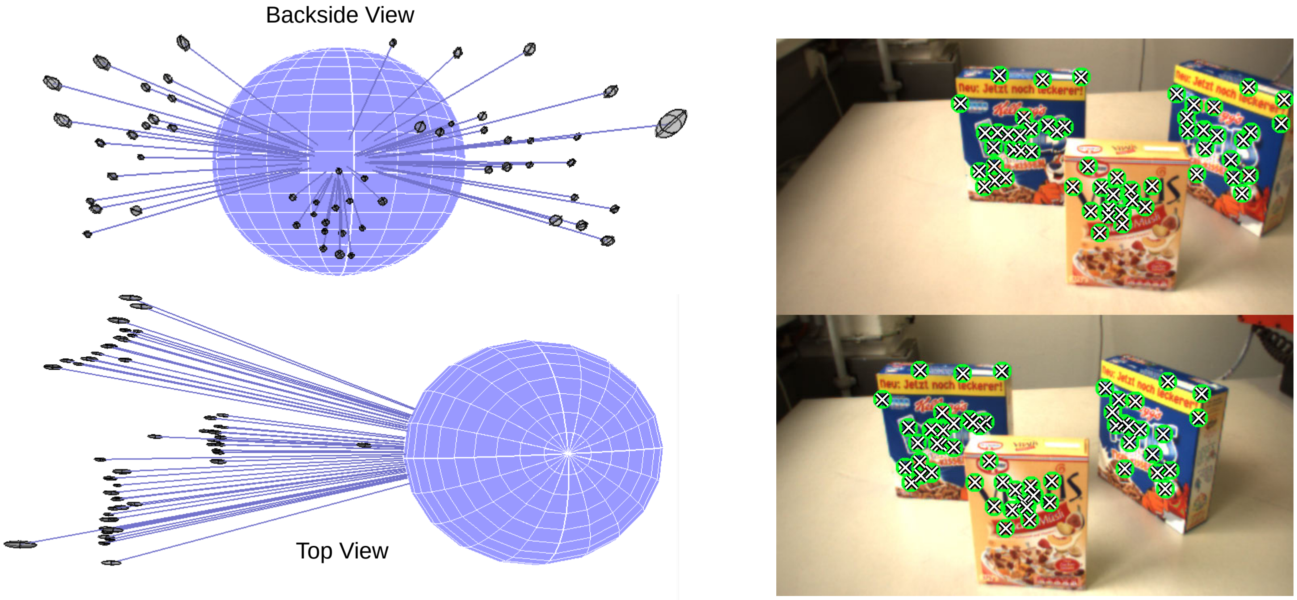
All experiments were carried out on the Karlsruhe Humanoid Head, equipped with 640×480 resolution FireWire cameras and lenses of focal length $f = 4mm$. The saccadic eye movements were performed using direct kinematics with random actuations for each eye joint. For eye pan actuations we selected values from the interval $\theta_{epl}, \theta_{epr} \in [-10°; 10°]$, eye tilt actuations were generated in the interval $\theta_{et} \in [-10°; 10°]$. In all experiments we used the same motion model. The positioning noise Σ_p was determined using the results from [6]. The standard deviation of all joints was set to the conservative value $\sigma_p = 0.15°$. The joint angle conversion error was determined empirically with $\sigma_e = 0.1°$. The calibration error could be derived by performing multiple calibrations and analyzing the Cartesian position of the rotation axis. The observed standard deviation was about $\sigma_{cal} = 1.0mm$. Furthermore, the maximum log probability for the existence of a landmark was set to $i_{max} = 5$ in order to react fast to changes in the scene. For all experiments we used $W = 200$ particles.

In the following, three experiments are presented using different features to generate landmarks.

A. Mapping with ground truth

In order to test the convergence of the map towards the observed scene, a chessboard rig with known size was deployed for the first experiment. As ground truth, the distance of chessboard corners was drawn on. For chessboard patterns, out-of-the-shelf corner point detectors can be applied with subpixel accuracy (see [22]). For this experiment we did not use signatures to describe the features. This corresponds to a similarity measure of $q(s_n) = 1$ for all landmarks in (13). Using the values from the cited work we chose a standard deviation $\sigma_x = \sigma_y = 0.5$ pixel for the uncertainty of corner point localization. 80 random saccadic eye movements were performed and the covariance matrix of all landmarks together with the distance of neighbored corner points was recorded.

The 48 landmarks could be tracked through all 80 saccadic eye movements. Figure 3(a) illustrates how the mean volume of the uncertainty ellipsoid $|det(\Sigma)|$ converges over the iterations of the particle filter. After 80 saccadic movements, the volume of the ellipsoid amounts to about $0.021mm^3$. In Figure 3(b), the mean distance of neighbored corner points over the 80 saccadic eye movements is illustrated. The manually measured distance of corner points on the chessboard amounts to $36.3mm$. The mean distance as calculated from the particle with the highest probability converged to about $36.0mm$. The difference between the mean distance and the ground truth of about 0.8% results from unmodeled phenomena such as inaccurate intrinsic camera parameters. Figure 3(c) illustrates the map after 80 iterations together with the uncertainties in landmark localization.



(a) Best map of SIFT features ofter 20 saccadic eye movements in the ego-centric coordinate frame. The resulting landmarks lie along the visible plane of the object.

(b) Reprojection of the best map onto the images of left and right camera.

Fig. 4. Results for mapping using SIFT features.

B. Mapping of Texture Features

In the second experiment we focused on a more realistic mapping task. As stimuli, a slight modification of the widely used SIFT features has been used [23]. The feature points were extracted using the Harris corner detector; as descriptor the SIFT approach was chosen. For the experiments we applied a standard deviation of $\sigma_x = \sigma_y = 1.5$ pixel for the uncertainty of 2D localization. For the similarity of SIFT descriptors $q(s_n)$ we use the Euclidean cross correlation.

Since no ground truth is available concerning the absolute or relative position of these features, examples of resulting mappings are provided in Figure 4(a). The figure illustrates how all three objects produced a set of 3D landmarks which have only a small deviation from the common plane after 20 saccadic eye movements. Four outliers have been mapped which result from erroneous correspondences between the left and right images. Figure 4(b) shows the projection of the resulting map to the images of the left and right camera.

C. Mapping of Coarse Features

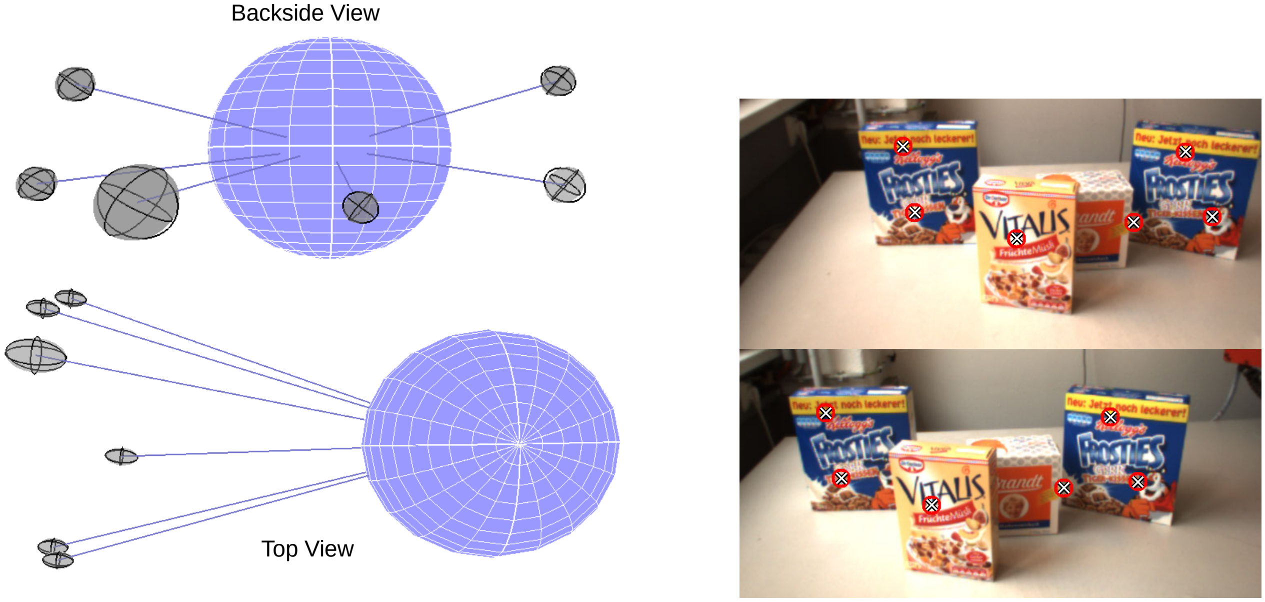
The third experiment deals with a slightly different scenario. Visual stimuli were generated at locations which correspond to an instance of a searched object. As features, receptive field cooccurrence histograms (RFCH) were used, which are a very coarse descriptor of the object. In the experiment, RFCHs only cover the hue channel of the image, the similarity measure $q(s_n)$ was calculated using histogram intersection. Since RFCHs were extracted using a window technique, the uncertainty of perception covers areas equal to the window size in the image plane. For the experiment the minimum window size was set to 32×48 pixel. The RFCH search produced visual stimuli which are a multiple of the minimum window size. The deviations of the 2D localization error σ_x and σ_y were set according to the window size.

Figure 5(a) shows the resulting map after 20 saccadic eye movements in a search task for the Frosties cereal box. Due to the coarse feature extraction method, uncertainties were much higher than in the previous two experiments. As can be seen in the back projection of the map (see Figure 5(b)), both instances of the object have two associated features which correspond to regions with different RCFH signatures. Due to the coarse description of objects using RFCHs, invalid hypotheses for object positions were integrated in the map. Such invalid hypotheses can be eliminated by further verification using foveated vision (see [24]).

IV. Conclusion

In this work we presented an approach for the integration of visual 3D features acquired while performing saccadic eye movements on an active camera system. We showed that a consistent ego-centric map of the environment with respect to the deployed features could be built and tracked using our approach.

While this work focuses on a static head performing eye movements with the goal to explore the surrounding environment, the proposed methods can also be applied on moving platforms by extending the proposed motion model. However, in contrast to related approaches, translational movement is not a prerequisite for the generation of 3D maps. This allows the exploration of salient regions within the augmented visual field of view resulting from eye movements.



(a) Best map of RFCH features ofter 20 saccadic eye movements in the ego-centric coordinate frame.

(b) Reprojection of the best map onto the images of left and right camera.

Fig. 5. Results for mapping using RFCH features.

Acknowledgment

The work described in this paper was partially conducted within the EU Cognitive Systems project PACO-PLUS (FP6-2004-IST-4-027657) funded by the European Commission.

References

[1] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, “The Karlsruhe Humanoid Head,” in *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, Dec. 2008, pp. 447–453.

[2] Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, and K. Fujimura, “The intelligent ASIMO: system overview and integration,” in *Intelligent Robots and System, 2002. IEEE/RSJ International Conference on*, vol. 3, 2002, pp. 2478–2483 vol.3.

[3] K. Akachi, K. Kaneko, N. Kanehira, S. Ota, G. Miyamori, M. Hirata, S. Kajita, and F. Kanehiro, “Development of humanoid robot HRP-3,” in *IEEE/RAS International Conference on Humanoid Robotics (HUMANOIDS)*, 2005.

[4] “Fujitsu, humanoid robot hoap-2,” 2003.

[5] A. Ude, C. Gaskett, and G. Cheng, “Foveated vision systems with two cameras per eye,” in *Proceedings of the 2006 IEEE International Conference on Robotics and Automation (ICRA)*, 2006, pp. 3457–3462.

[6] K. Welke, M. Przybylski, T. Asfour, and R. Dillmann, “Kinematic calibration for saccadic eye movements,” Institute for Anthropomatics, University of Karlsruhe (TH), Tech. Rep., 2008.

[7] O. Faugeras and Q. Luong, *The Geometry of Multiple Images*. MIT Press, 2004.

[8] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[9] S. Frintrop, P. Jensfelt, and H. I. Christensen, “Attentional robot localization and mapping,” *ICVS Workshop on Computational Attention and Applications*, 2007.

[10] H. Yu and Y. Wang, “An improved self-calibration method for active stereo camera,” in *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on*, vol. 1, 2006, pp. 5186–5190.

[11] D. Liebowitz and A. Zisserman, “Metric rectification for perspective images of planes,” in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, Jun 1998, pp. 482–488.

[12] M. J. Brooks, L. d. Agapito, D. Q. Huynh, and L. Baumela, “Direct methods for self-calibration of a moving stereo head,” in *ECCV ’96: Proceedings of the 4th European Conference on Computer Vision-Volume II*. London, UK: Springer-Verlag, 1996, pp. 415–426.

[13] L. Goncalves, E. di Bernardo, D. Benson, M. Svedman, J. Ostrowski, N. Karlsson, and P. Pirjanian, “A visual front-end for simultaneous localization and mapping,” in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, April 2005, pp. 44–49.

[14] A. Davison, I. Reid, N. Molton, and O. Stasse, “Monoslam: Real-time single camera slam,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 1052–1067, June 2007.

[15] C. F. Olson, L. H. Matthies, M. Schoppers, and M. W. Maimone, “Rover navigation using stereo ego-motion,” *Robotics and Autonomous Systems*, vol. 43, no. 4, pp. 215 – 229, 2003.

[16] L. Matthies and S. Shafer, “Error modeling in stereo navigation,” *Robotics and Automation, IEEE Journal of*, vol. 3, no. 3, pp. 239–248, June 1987.

[17] Z. Chen, D.-C. Tseng, and J.-Y. Lin, “A simple vision algorithm for 3-d position determination using a single calibration object,” *Pattern Recognition*, vol. 22, no. 2, pp. 173 – 187, 1989.

[18] D. R. Murray, “Patchlets: a method of interpreting correlation stereo three-dimensional data,” Ph.D. dissertation, 2004.

[19] K. Murphy and S. Russell, “Rao-blackwellized particle filtering for dynamic bayesian networks,” in *n Sequential Monte Carlo Methods in Practice*, 2001.

[20] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, “Fastslam: A factored solution to the simultaneous localization and mapping problem,” in *In Proceedings of the AAAI National Conference on Artificial Intelligence*. AAAI, 2002, pp. 593–598.

[21] Thrun, *Probabilistic Robotics*. Cambridge, MA: MIT Press, 2005.

[22] J. Mallon and P. F. Whelan, “Which pattern? biasing aspects of planar calibration patterns and detection methods,” *Pattern Recognition Letters*, vol. 28, no. 8, pp. 921 – 930, 2007.

[23] P. Azad, T. Asfour, and R. Dillman, “Combining harris interest points and the sift descriptor for fast scale-invariant object recognition,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.

[24] K. Welke, T. Asfour, and R. Dillmann, “Active multi-view object search on a humanoid head,” in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2009.