Accumulated Visual Representation for Cognitive Vision

Nicolas Pugeault[1,2], Florentin Wörgötter[3] and Norbert Krüger[2]
[1] University of Edinburgh, UK.
[2] Syddansk Universitet, Denmark.
[3] (BCCN) University of Göttingen, Germany.

Abstract

In this paper we present a scheme for accumulating local visual information in 3D, under known motion. Information about the object's 3D shape is provided by reconstructing local contour descriptors. This shape information is accumulated over time in three ways: 1) disambiguation: erroneous stereo correspondences that are unsuccessfully tracked are discarded. We make use of aspect cues to increase the data association selectivity. 2) correction: the full pose of the reconstructed features is corrected over time using an Kalman Filter approach. 3) completeness: multiple $2\frac{1}{2}D$ representations become merged, constructing a full 3D representation of the object. The described system is evaluated quantitatively on three different scenarios.

1 Introduction

This paper introduce a system that generates an internal representation of unknown objects under a known motion. First, a set of local contour descriptors, called 2D–primitives, are extracted from the image, providing a first representation of the 2D shapes. Two such representations are matched across two views using stereopsis, and 2D–primitives are matched in both images to reconstruct 3D contours as strings of local features called 3D–primitives, that provide a first representation of the 3D shapes in the scene. At this stage, the representation is merely a collection of 3D–primitives, objects and background are not segmented in any way. By using the motion knowledge provided by the robot, we segment the object from the scene (by selecting primitives that move according to the robot's arm motion) and accumulate the representation. Such an exploratory mechanism enable the system to explore its environment and learn the shape of the objects that inhabit it.

This paper presents the mechanism used for accumulating the visual representation over time. Having control over the object provides a very accurate knowledge of its motion that can be used to track individual 3D primitives. At each frame, new observations are used to correct the 3D primitives' full pose and to enrich the representation with new aspects of the object (e.g., parts that were previously occluded). The mechanism presented herein improves the representation in three respects: (i) Accuracy: The representation is corrected over time using new observations. (ii) Reliability: Tracking primitives over time, it is possible to re–evaluate their reliability over time, and to discard erroneous ones. Because the tracking is done in 3D space, the likelihood for erroneous

BMVC 2008 doi:10.5244/C.22.98

primitives to be tracked successfully is vanishingly small. (iii) Completeness: Through Manipulation of the object, the system witnesses it under a wide range of viewpoints, and accumulate $2\frac{1}{2}D$ representations into a full $3D$ representation. Solutions to this problem belongs to two groups:

The first group consists of the geometric analytic solutions, including multifocal tensors [6] and bundle adjustment [14]. These techniques are ideal solutions to the problem and are prominent in the strict batch–SFM scenario. They can be designed to be robust to erroneous data association (see [14] for a discussion). One major flaw with these solutions stems from the fact that they are *batch* processes: all views need to be available. This can make the problem intractable for large sequences, and implies a delay in any active system. It has been proposed to split the problem into groups of, e.g., 3 frames, reducing both delay and computational cost. Nonetheless, these approaches face the dead–reckoning problem: small motion errors accumulate over time to lead to large localisation errors. Therefore, they generally need an additional global integration stage.

The second group uses various flavours of the Bayesian filtering theory. This provides an on–line solution by formalising the problem as a Markov process where a state vector combining the current pose and the landmarks' bearing can be formalised as the general Bayesian Tracking problem — see [1] for a review. This theoretical formulation allows for an optimal solution, i.e., a Kalman filter, if the state vector as a multivariate normal distribution and if the prediction and observation processes are linear. Kalman filters and its non–linear derivatives (e.g., Extended or Unscented Kalman filters) have been used extensively to solve the simultaneous localisation and map–building (SLAM) problem (see, e.g., [2, 15, 5, 13, 9]).

Because of the on–line constraint, the approach presented in this paper belongs to the second category. The present scenario is quite different, because the motion prediction is very accurate, the framerate is high, and a large proportion of the primitives is visible at any time. On the other hand, because they only encode local contour information, primitives are not very distinctive. SLAM, in contrast, focuses on scenario where motion knowledge is inaccurate, successive frames are generally far apart and landmarks are very distinctive (e.g., SIFT [10]) The differences of the approach presented herein are: (i) tracking in 3D space: the tracking of the primitives is done using homogeneous 3D coordinates, and therefore the robot's arm rigid motion is a linear operation. (ii) data association in stereo 2D space: 3D primitives are re–projected in both image planes to be compared with extracted primitives. This allows to reduce the uncertainty generated by stereopsis. (iii) full pose tracking: because the primitives are local contour descriptors, they encode local position and orientation. Therefore we use a pair of Kalman filters to filter the full pose.

The framework is described in section 2, then evaluated on different scenarios in section 3, before concluding in section 4.

2 Framework presentation

The framework proposed here integrates visual information over time to improve reliability, accuracy and completeness. The scheme presented herein extracts local image features called 3D–primitives (see section 2.1). The system then attempts to tracks 3D–primitives in each new stereo pair of images (see section 2.2.2) using available motion

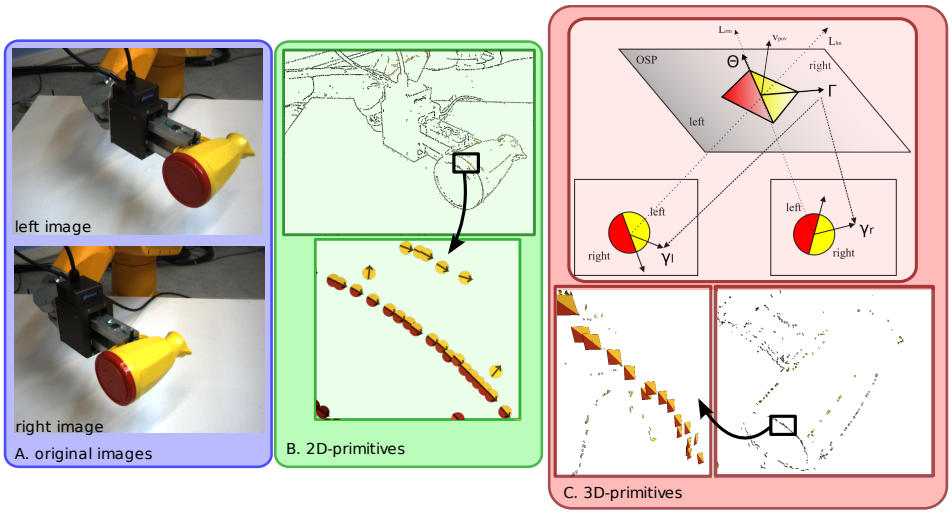


Figure 1: Illustration of the primitive extraction stage. A) original images (left and right); B) extraction of 2D–primitives in both images; and C) stereo reconstruction of 3D–primitives.

knowledge. Matched primitives are corrected over time using Kalman filtering (see section 2.2), and their confidence is re-evaluated depending how successfully they have been tracked since they were first witnessed (see section 2.2.3). Existing primitives whose confidence fall below a set level are then discarded, while new primitives that were not observed before are added to the representation.

2.1 2D and 3D Visual Primitives

In this work we use sparse image descriptors called *primitives*, that exist both in 2D and 3D space, and were discussed in [8]. In the 2D space, those primitives provide a condensed representation of image information sparsely sampled along image contours. In a first stage, linear and non–linear filtering operations are applied to the image (see, e.g., [7]), Positions are detected sparsely with sub–pixel accuracy at places likely to contain edges (see, e.g., [7] for a description), and local information is encoded into the following feature vector:

$$\pi = (\mathbf{x}, \theta, \phi, \mathbf{c}, \mathbf{f}) \tag{1}$$

where $\mathbf{x}$ is the image position determined with sub–pixel accuracy; θ is the local orientation, encoded as a normalised tangent vector; ϕ is the local phase; $\mathbf{c}$ is a dimension 6 vector encoding the HSV colour values on both sides of the contour; and $\mathbf{f}$ is the optic flow vector. In the following we refer to such features as *2D–primitives*.

Such 2D–primitives are extracted on stereo pairs of images and are matched using the epipolar line and similarity constraints (see Fig. 1B, and [12] for an assessment). Pairs of matched 2D–primitives provide enough information to reconstruct the 3–dimensional equivalent of a 2D–primitive, denoted *3D–primitive* in the following (see Fig. 1C). We direct the reader to [4, 6] for a description on classical stereo reconstruction and [11] for

the special case of primitives. A 3D–primitive encodes a scene contour's local position and orientation along with the local contrast and colour on each side.

$$\Pi = (\mathbf{X}, \Theta, \phi, \mathbf{C}) \tag{2}$$

where $\mathbf{X}$ is the position in space; Θ is the 3D orientation of the line; ϕ is the local phase; $\mathbf{C}$ is a dimension 6 vector encoding the HSV colour values on both sides of the contour.

2.2 3D–primitive's pose filtering

In this section we present how a 3D–primitive is tracked and how we apply Bayesian filtering to its pose. The Kalman filter is an optimal derivation of the Bayesian filtering problem that is limited to filtering state vectors with normally distributed noise when the prediction and observation can be modelled by linear functions. The 3D–primitives' full pose is filtered using a pair of Kalman states. The position and orientation in space are encoded by a pair of dimension 4 homogeneous vectors $(\mathbf{X}, \Theta)$. The same filter equations can be used to filter both position and orientation.

2.2.1 State vector and state prediction

Motion knowledge can come from motor knowledge for example when a robot manipulates an object, embarked sensors for a mobile system, or visual estimation. This motion knowledge predicts the pose of the primitive at a later time-step. In this work we only consider a restricted kind of motion, called Rigid Body Motions (RBM). RBM transformation is a linear function in homogeneous spatial coordinates, and therefore can be directly used in a Kalman filter, leading to the following equations:

$$\hat{s}_{t|t-1} = \mathbf{M}_t \cdot \hat{s}_{t-1|t-1} \tag{3}$$

$$\hat{\Sigma}_{t|t-1} = \mathbf{M}_t \cdot \hat{\Sigma}_{t-1|t-1} \cdot \mathbf{M}_t^\top + \mathbf{Q}_t \tag{4}$$

where s signifies the state, and can either be the 3D position $\mathbf{X}$ or orientation Θ as 3D homogeneous (dimension 4) vectors.

In these equations, the RBM knowledge is encoded as the 4×4 matrix $\mathbf{M}_t$, and the prediction uncertainty is $\mathbf{Q}_t = \varepsilon \mathbf{I}_{4\times 4}$, where ε is a small value set to be larger than the motion prediction error. Because this error is very small, it can be approximated as normal and isotropic.

2.2.2 Temporal matching

The matching itself is performed by re–projecting both the accumulated $\hat{\Pi}_{t|t-1}$ and the newly reconstructed Π_t 3D–primitives on both image planes, and match primitives which projections are proximate and similar in both — see Fig. 2 Doing the matching in 2D provides additional robustness to projective deformation and stereo uncertainty while the necessary selectiveness is preserved by ensuring that the primitives are matched in *both* views at each time–step.

The position uncertainty of the 3D–primitives is also re–projected in the image domain, into a 2×2 covariance matrix. Using this covariance matrix we estimate the likelihood for the 3D–primitive to find a match at each location by a normal distribution combined with an uniform distribution (that expresses the chance for a correct 3D–primitive

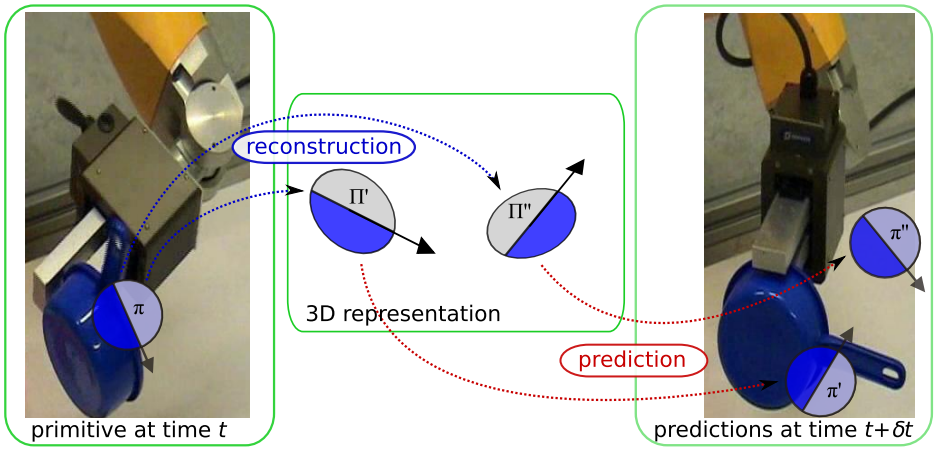


Figure 2: Illustration of the primitives temporal matching.

not to be matched). We will write the fact that a primitive Π_i that predicts a primitive $\hat{\Pi}_i^t$ at time t is matched (as described above) as $\mu_t(\hat{\Pi}_i)$ and evaluate its likelihood as:

$$p[\mu_t] = \frac{1}{(2\pi)\sqrt{|\Sigma_x|}} e^{\frac{1}{2}(\Delta x)^\top \Sigma_\Delta^{-1}(\Delta x)} \tag{5}$$

The matrix $\Sigma_\Delta = \hat{\Sigma}_x + \Sigma_x$ is the sum of the re–projected position uncertainty for both primitives in this image. Also, (Δx) is the difference between the position of the two re–projected primitives $\Delta x = \hat{x}_{t|t-1} - x_t$ where $\hat{x}_{t|t-1}$ is the predicted position and x_t is the position of the potential match. If the confidence $p[\mu_t|(\Delta x), \Pi]$ is larger than the chance value $o = 0.1$, the the match is considered valid. Furthermore, the similarity between the primitives (in orientation, phase, and colour) is also considered, and matches with a too low similarity (lower than $\tau = 0.9$) are disregarded.

2.2.3 Confidence re–evaluation from tracking

The third corrective mechanism re–evaluates the confidence in the 3D–hypotheses depending on their resilience. This is justified by the continuity assumption, which states that 1) any given object or contour of the scene should not appear and disappear in and out of the field of view (FoV) but move gracefully in and out according to the estimated ego–motion, and 2) that the position and orientation of such a contour at any point in time is fully defined by the knowledge of its position at a previous point in time and of the motion of this object between these two instants.

As we exclude from this work the case of independent moving object, and as the ego–motion is known, all conditions are satisfied and we can trace the position of a contour extracted at any instant t at any later stage $t + \Delta t$, as well as predict the instant when it will disappear from the FoV.

We define the tracking history of a primitive Π_i from its apparition at time 0 until time t as:

$$\mu(\Pi_i) = \left(\mu_t(\hat{\Pi}_i), \mu_{t-1}(\hat{\Pi}_i), \cdots, \mu_0(\hat{\Pi}_i)\right)^T \tag{6}$$

thus, applying Bayes formula:

$$p\left[\Pi_i|\mu(\hat{\Pi}_i)\right] = \frac{p\left[\mu(\hat{\Pi}_i)|\Pi\right]p\left[\Pi\right]}{p\left[\mu(\hat{\Pi}_i)|\Pi\right]p\left[\Pi\right] + p\left[\bar{\mu}(\hat{\Pi}_i)|\bar{\Pi}\right]p\left[\bar{\Pi}\right]} \tag{7}$$

where Π and $\bar{\Pi}$ are correct and erroneous primitives, respectively.

Furthermore, if we assume independence between the matches we have, and assuming that Π exists since n iterations and has been matched successfully m times, we have:

$$\begin{aligned} p\left[\mu(\hat{\Pi}_i)|\Pi\right] &= \prod_t p\left[\mu_t(\hat{\Pi}_i)|\Pi\right] \\ &= p\left[\mu_t(\hat{\Pi}_i)=1|\Pi\right]^m p\left[\mu_t(\hat{\Pi}_i)=0|\Pi\right]^{n-m} \end{aligned} \tag{8}$$

In this case the probabilities for μ_t are equiprobable for all t, and therefore we define the quantities $\alpha = p\left[\Pi\right]$, $\beta = p\left[\mu_t(\hat{\Pi})=1|\Pi\right]$ and $\gamma = p\left[\mu_t(\hat{\Pi})=1|\bar{\Pi}\right]$. We can then rewrite Eq. (7) as follows:

$$p\left[\Pi_i|\bar{\mu}(\hat{\Pi}_i)\right] = \frac{\beta^m(1-\beta)^{n-m}\alpha}{\beta^m(1-\beta)^{n-m}\alpha + \gamma^m(1-\gamma)^{n-m}(1-\alpha)} \tag{9}$$

We measured these prior and conditional probabilities using a video sequence with known motion and depth ground truth obtained via range scanner. We found values of $\alpha = 0.2$, $\beta = 0.4$ and $\gamma = 0.1$. This means that, in these examples, the prior likelihood for a stereo hypothesis to be correct is 20%, the likelihood for a correct hypothesis to be confirmed is 40% whereas for an erroneous hypothesis it is of 10%. These probabilities show that Bayesian inference can be used to identify correct correspondences from erroneous ones. When a confidence in an accumulated primitive rises above 0.9, it is preserved and its confidence is not updated anymore. If it falls below 0.1, its is discarded.

2.2.4 Pose correction

The primitives position and orientation are corrected using the newly reconstructed match s_t. The correction is a direct application of the Kalman update step. Because both, the predicted and the observed states lie in the same homogeneous 3D space, the Kalman observation function is the identity, and the difference between predicted and observed states is:

$$\Delta s = s_t - \hat{s}_{t|t-1} \tag{10}$$

The innovation matrix is given by:

$$S_k = \Sigma_{t|t-1} + \hat{\Sigma}_t \tag{11}$$

And the optimal Kalman gain is:

$$K_t = \Sigma_{t|t-1} \cdot S_k^{-1} \tag{12}$$

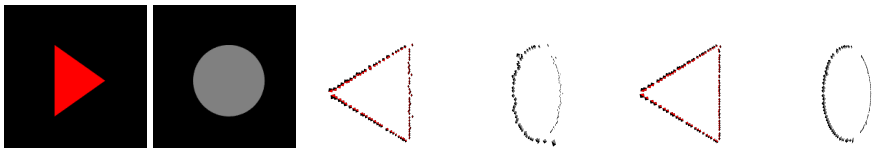
Finally the updated state is:

$$s_{t|t} = s_{t|t-1} + K_t \cdot \Delta s \tag{13}$$

With an updated covariance:

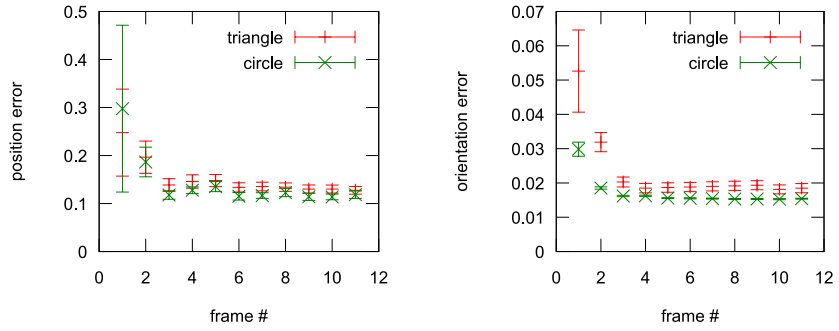
$$\Sigma_{t|t} = (I - K_t) \cdot \Sigma_{t|t-1} \tag{14}$$

Those formulae are applied similarly to the position X and the orientation T vectors.



(a) images (b) stereo (c) accumulation 10 frames

Figure 3: Pose correction over time, for a pure lateral translation.



(a) position (b) orientation

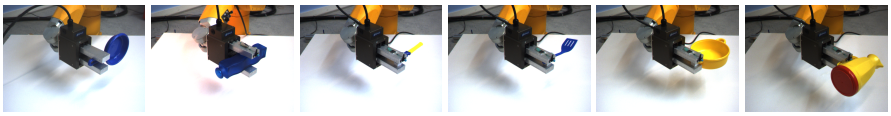
Figure 4: Pose correction over time, for a pure lateral translation, see Fig. 3.

3 Results

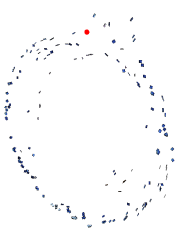
The framework presented in the previous section was evaluated in three different scenarios. The first, artificial scenario features a simple geometric scene, of a moving circle, where the primitives' accuracy could be evaluated with precision (section 3.1). The motion is known. The second scenario features a robot context, where the robot's arm has grasped an object and is inspecting it from different viewpoints (section 3.2). The motion is known from the arm's actuators with good precision. Finally, in a third scenario stereo cameras were mounted on a driving car (section 3.3). In this case the car's motion is approximated by on-board instruments, but is known to be more unreliable.

3.1 Artificial scenario

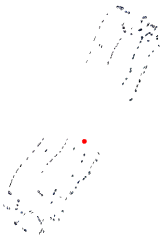
As a first evaluation of the scheme, we used two simple artificial sequences generated using openGL. In this case the motion is known with accuracy, as well as the exact shape of the object (a circle and a triangle). We used those sequences to evaluate the evolution of the 3D–primitives accuracy during accumulation. The images and the reconstructed 3D–primitives are illustrated in Fig. 3, and the primitives' position and orientation accuracy is recorded in Fig. 4. It is clear in these plots that the position and orientation error reduces quickly with accumulation, as well as the variance.



(a) original images



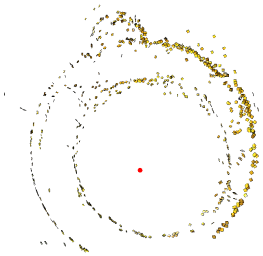
(b) pan



(c) vase

(d) knife

(e) spatula



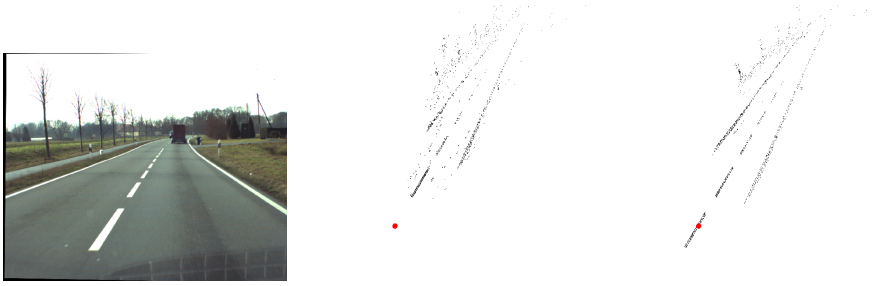
(f) pot

(g) jug

Figure 5: Accumulated objects (after 70 frames, spanning a whole rotation of the object).

3.2 Robotic scenario

The accumulation scheme was applied to a robotic manipulation scenario. In this scenario, we have a fixed camera and a robotic arm holding an unknown object. The object is manipulated in front of the cameras such that all its sides become visible, The robot's arm motion is known from the actuators with very good precision, making primitive tracking over time easy. On the other hand, the curved surfaces of the objects make the accumulation more difficult. Also the rotational movement means only half of the object is visible at anytime, leading to large occlusions. The results for several toy kitchen objects is shown in Fig. 5. Fig. 5(a) shows one image of each object, being held by the robot. Fig. 5(b-g) show the accumulated representation of the objects. One can see there that the full 3D-shape is accumulated, with few outliers. Note that the segmentation of the object from the background is done implicitly because the background do not move according to the predicted motion. The robot's hand is discarded using a simple bounding box, as its position is always known exactly (namely, at the end of the arm). The blank space in the representations correspond to the place where the object were held, and could be filled by manipulating the object a second time, holding if from a different point.



(a) image (b) stereo (c) accumulation 8 frames

Figure 6: Accumulation of visual information in a driving scenario. (a) shows one image from the video sequence processed; (b) shows the 3D–primitives reconstructed by stereo only; and (c) shows the accumulated 3D–primitives after 8 frames.

3.3 Driving scenario

We applied this scheme to a driving scenario. In this scenario, two stereo cameras were fixed on a car driving on a country road. Some motion knowledge is provided by the car's instruments, but the recorded motion is limited to the planar components (velocity and rotation angle) and is known to be inaccurate. Nonetheless, we approximated the RBM from this motion knowledge. The correction of the 3D–primitives over time prevents the motion error from breaking the tracking. In this case the motion knowledge is too inaccurate to provide a good correction (the motion uncertainty is larger than the reconstruction uncertainty), but sufficient for tracking and outlier removal.

In Fig. 6(a) shows one image from a sequence recorded from the inside of a car, while driving. In Fig. 6(b) shows the 3D–primitives reconstructed using stereo, with some outliers due to texture on the road. Fig. 6(c) shows the 3D–primitives accumulated over 8 frames. In the accumulated representation, spurious primitives and erroneous stereo matches have been discarded. At the same time, 3D–primitives that were visible at different frames are integrated into one representation, showing a longer stretch of the road. The primitives were matched and tracked accurately despite the motion inaccuracy.

4 Conclusion

Low level image processes suffer from a large amount of noise, uncertainty, and ambiguity. In this paper, we presented a framework accumulating stereo information over time and improving the quality of the representation in three respects; 1) the accuracy of the 3D–primitives' position and orientation is corrected over time; 2) spurious primitives and erroneous stereo correspondences are discarded; and 3) 3D–primitives visible under different perspectives are integrated.

Moreover, using an accumulated representation improves the stability of the representation over time. When new parts of the scene becomes visible they are naturally integrated in the scene, and when parts become occluded they become preserved in memory.

Because of all that, we believe that temporally accumulated visual representation is a

better basis for high level vision processes to be based on.

Acknowledgements This work is supported by the European project DRIVSCO (FP6-IST-FET) [3].

References

[1] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A Tutorial on Particle Filters for Online Nonlinear/Non–Gaussian Bayesian Tracking. *IEEE Transactions on Signal Processing*, 2002.

[2] P. Dissanayake, P. Newman, H.F. Durrant-Whyte, S. Clark, and M. Csorba. A solution to the simultaneous localisation and mapping (SLAM) problem. *IEEE Transactions in Robotics and Automation*, 17(3):229–241, 2001.

[3] DrivSco. DrivSco: learning to emulate perception–action cycles in a driving school scenario. FP6-IST-FET, contract 016276-2, 2006.

[4] O.D. Faugeras. *Three–Dimensional Computer Vision*. MIT Press, 1993.

[5] J.E. Guivant and E.M. Nebot. Optimization of the Simultaneous Localization and Map–Building Algorithm for Real–Time Implementation. *IEEE Transactions on Robotics and Automation*, 17(3):242–257, 2001.

[6] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2000.

[7] B. Jähne. *Digital Image Processing*. Springer, 2002.

[8] N. Krüger, M. Lappe, and F. Wörgötter. Biologically motivated multi-modal processing of visual primitives. *Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour (AISB)*, 1(5):417–427, 2004.

[9] T. Lemaire, C. Berger, I-K. Jung, and S. Lacroix. Vision–Based SLAM: Stereo and Monocular Approaches. *International Journal of Computer Vision*, 74(3):343–364, 2007.

[10] D.G. Lowe. Distinctive Image Features from Scale–Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[11] N. Pugeault. *Early Cognitive Vision: Feedback Mechanisms for the Disambiguation of Early Visual Representation*. PhD thesis, Georg-August-Universität Göttingen, 2008.

[12] N. Pugeault and N. Krüger. Multi–modal matching applied to stereo. *Proceedings of the BMVC 2003*, pages 271–280, 2003.

[13] S. Thrun, Y. Liu, D. Koller, A.Y. Ng, Z. Ghahramani, and H.F. Durrant-Whyte. Simultaneous Localization and Mapping with Sparse Extended Information Filters. *International Journal of Robotics Research*, 23(7–8):693–716, 2004.

[14] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – A modern synthesis. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, LNCS, pages 298–375. Springer Verlag, 2000.

[15] R. van der Merwe, A. Doucet, N. de Freitas, and E. Wan. The Unscented Particle Filter. Technical Report CUED/F-INFENG/TR 380, Cambridge University Engineering Department, 2000.