Accumulation of Object Representations utilizing Interaction of Robot Action and Perception

Norbert Krüger, Marcus Ackermann, Gerald Sommer

Lehrstuhl für kognitive Systeme
Institut für Informatik,
Christian–Albrechts–Universität zu Kiel
Preusserstrasse 1-9, 24105 Kiel, Germany
nkr{maa,gs}@ks.informatik.uni-kiel.de

Abstract

We introduce a robotic–vision system which is able to extract object representations autonomously utilizing a tight interaction of visual perception and robotic action within a perception action cycle [10, 17]. Controlled movement of the object grasped by the robot enables us to compute the transformations of entities which are used to represent objects and to find correspondences of entities within an image sequence.

A general accumulation scheme allows to acquire robust information from imperfect and partly missing information extracted from single frames of an image sequence. Here we used this scheme with a preprocessing stage in which 3D-line segments are extracted from stereo images. However, the accumulation scheme can be used with any kind of preprocessing as long as the entities used to represent objects can be brought to correspondence by certain equivalence relations such as 'rigid body motion'.

1 Introduction

Model based vision systems usually apply manually designed object representations (see e.g., [19] or [12]). These methods usually work well but commonly have drawbacks with the need of manual intervention for creating object representations and the fine–tuning of these representations. Here we demonstrate an autonomous extraction of object representations making use of a tight interaction of perception and action: Accumulation of information takes place within a perception–action–cycle [10, 17]. As a challenging perspective we aim at a coupled robotic–vision system which is not equipped with manually designed object representations but the object to be manipulated is given to the robot and a representation is accumulated autonomously (see figure 1).

Feature extraction faces the problem that semantic information extracted by artificial systems from a single image or stereo images even under optimal conditions is necessarily imperfect. For instance, although there exist a large amount of edge detectors none of them is comparable to human performance. One important reason for the extremely good performance of humans on these tasks is that the human visual system applies *constraints* to interpret a certain scene or situation [6, 11]. A situation never stands for itself but is embedded in a time continuum [7]. Therefore an important constraint is the utilization of the coherence of objects during a rigid body motion which allows to accumulate information over time.

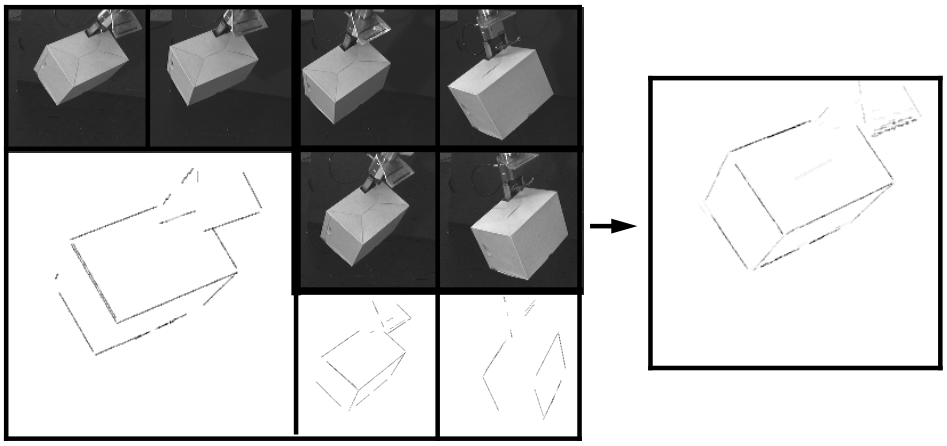

Figure 1: **left)** top: left and right image of an object. bottom: the projected 3D representation extracted from the stereo images. **middle)** Two pairs of stereo images (top: left camera image, middle: right camera image) and the the projected 3D representation (bottom). **right)** Projected 3D Representation accumulated over a set of stereo images. Dark areas represent line segments accumulating high confidences. Grey areas represent line segments accumulating medium or low confidences.

In this paper we suggest to accumulate object representations from image sequences by using the equivalence relation 'rigid body motion'. We account for the vagueness of semantic information extracted from single images by assigning confidences to this information and accumulating this information over an image sequence of a moving object. Although the information extracted from single images contain errors (see the representations on the left hand side of figure 1) a more stable representation can be achieved by combining information from different images (see right hand side of figure 1). Because the object can change its position and orientation — and this change might be wanted because another view of the object gives new information which might not be extractable from another view — we face the correspondence problem: Correspondences between entities describing the object in different images (or 3D interpretations extracted from stereo images) are not known.

Here the correspondence problem is solved within a behavior based paradigm [3, 16]. The parameter of motion are known since the robot manipulates the object and the transformations of entities can be compensated for each frame of the sequence to achieve correspondences. Knowing the correspondences an algorithm can be applied to update and improve the object representation iteratively. This accumulation algorithm is an extension of an algorithm introduced in [11, 14] which has only dealt with 2D representation and translational motion.

2 Extraction of Object Representations from Image Sequences

Our accumulation algorithm can be defined independently of the entities used to represent objects. The algorithm also is independent of the concrete equivalence relation or transformation used to define correspondences. It only requires an object representation by certain entities for which a metric is defined and to which certain transformations or equivalence relations (such as rigid body motion) can be applied. The object establishes itself as an invariant under the equivalence relation, i.e., as an equivalence class. The algorithm in its general form is defined in subsection 2.1. In this paper for the representation of objects we use local three dimensional line

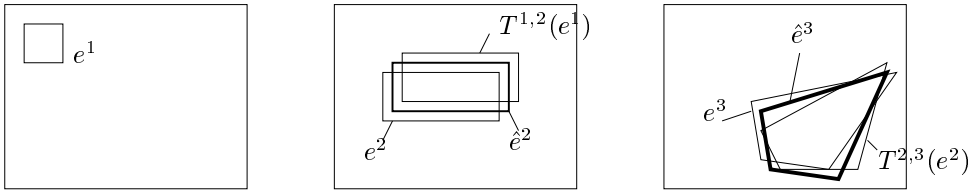

Figure 2: The accumulation scheme. The entity e^1 (here represented as a square) is transformed to $T^{1,2}(e^1)$. Note that without this transformation it is barely unpossible to find a correspondence between the entities e^1 and e^2 because the entities show significant differences in appearance and position. Here a correspondence between $T^{1,2}(e^1)$ and e^2 is found because a similar square can be found close to $T^{1,2}(e^1)$ and both entities are merged to the entity $\hat{e}^2$. The confidence assigned to $\hat{e}^2$ is set to a higher value than the confidence assigned to e^1 indicated by the width of the lines of the square. The same procedure is then applied for the next frame for which again a correspondence has been found. By this scheme information can be accumulated to achieve robust representations.

---

segments only. The extension of the system to other kind of object descriptors such as texture, color or optical flow is part of our current research.

The concrete realization of the accumulation scheme can be divided into two parts, preprocessing (section 2.2.1) and accumulation (section 2.2.2). The algorithm is applied to a stereo image sequence in which the object grasped by the robot is shown to the system in various positions and orientations (see figure 1). A representation is accumulated over the stereo image sequence (see figure 1 right). Although the representations extracted from one stereo image pair shows missing line segments (left) the accumulated representation is more complete (right). Here we give only a condensed description of the algorithm, for details see [1].

2.1 The Accumulation Scheme

Let $e \in E$ be an entity used to describe objects (for instance a 2D–line segment, a structure tensor [9] extracted from an image, 3D–line segments extracted from a stereo image pair or any other kind of object descriptor) and $d(e, e')$ be a distance measure on the space of entities E. Furthermore, let T be a transformation or equivalence relation, for instance a rigid body motion or the projective transformation corresponding to a rigid body motion. If e^i is an entity extracted from frame i of a sequence of events then $T^{i,i+1}(e^i)$ is the transformation $T^{i,i+1}$ from the i–th to the $i+1$–the frame applied to e^i.

Let e^{i+1} be an entity extracted from the i+1–th frame of the sequence we say that e^i and e^{i+1} are likely to correspond to each other if $d(T(e^i), e^{i+1})$ is small. Often it might not be possible to find an exact correspondence with $d(T(e^i), e^{i+1}) = 0$. For example, if we want to compare local image patches in two images knowing the exact projective transformation corresponding to the rigid body motion of an object from the first to the second frame, the corresponding image patches can not be expected to be exactly equal because of factors such as noise during the image acquisition, changing illumination, non–Lambertian surfaces or discretization errors. The problem may even become more severe when we extract more complex entities such as 3D or 2D line segments or 3D–surface patches. Therefore it is advantageous to formalize a confidence of correspondence by a metric.

The accumulation of information can now simply be achieved by the following update rule: If there exists an entity e^{i+1} in the i+1–th frame for which $d(T(e^i), e^{i+1})$ is small (i.e. a correspondence is likely) then merge $T(e^i)$ and e^{i+1} by some kind

of average operator $\hat{e}^{i+1} = \text{merge}(T(e^i), e^{i+1})$ and set the confidence for $\hat{e}^{i+1}$ to a higher value than the confidence assigned to e^i. If there exists no entity e^{i+1} in the $i+1$–the frame for which $d(T(e^i), e^{i+1})$ is small, the confidence for entity e^i to be part of the object is decreased. In Figure 2 a schematic representation of the algorithm is shown for two iterations.

2.2 Application of the Accumulation Scheme to a Representation with 3D–line segments

In this section we apply the accumulation scheme introduced above to object representations consisting of local 3D line segments. For these entities the change of the transformation (i.e., $T^{i,i+1}(e)$) can be computed explicitly (for details see [1]).

2.2.1 Extraction of a 3D Representation from Stereo Images

In the preprocessing step a 3D representation of the object grasped by the robot and presented at a certain position and orientation is extracted. The orientation of the object differs in each stereo image pair (figure 1). The object representation consists of local 3D–line segments and is extracted using calibrated cameras and epipolar geometry. First, in each single image lines are extracted using the orientation sensitive Hough transformation [15]. The Hough lines are divided into local line segments according to local information indicating evidence for the existence of a local line segment at a certain pixel position in the image by evaluating gradient information. In our implementation the entity 'local line segment' can only be extracted when there is local support (a high magnitude of the gradient) *and* global support (the line segment is part of a Hough line). Second, correspondences of line segments in the two stereo images are found. The epipolar constraint is used to reduce the search problem to a one-dimensional problem. On the epipolar line corresponding to a certain line segment the best match is defined as the corresponding entity. For finding the best match a similarity combining gray level information (by evaluating the correlation of image patches) and semantic information (evaluating the differences in the orientation of the found line segments) are used.[1]

In most cases the correspondence of 2D line segments defines a 3D line segment. In some cases, when the 2D line segments are close to a 'critical plane' [4, 8] the correspondences do not uniquely define a 3D line segment and a 3D representation of parts of the object can not be extracted. Note that by moving the object, 3D–line segments which can not be extracted in one frame (because they are too close to the critical plane) move out of the critical plane so that they can be part of the final representation. Here the haptic control of the object allows the creation of situations in which critical features can be extracted.

The representation extracted from a single stereo image pair usually is not perfect (see figure 1), there are many missing parts (because of the critical plane, correspondences not found, not detected Hough lines or not extracted 2D line segments in one of the two stereo images) and some 'wrong' line segments (because of wrong correspondences or wrong 2D line segments extracted during preprocessing). Here we face the problem that semantic information can not be extracted with sufficient accuracy from single or stereo images which is also one of the the reasons for the need of manually designed object representations in many artificial systems.

To achieve a suitable representation autonomously and to overcome the need

[1] This kind of preprocessing faces the problem that for small edges Hough lines often can not be found so that they do not occur in the object representation extracted from one stereo image pair. Therefore we aim to use the local image operator introduced in [5] to overcome this problem. Furthermore, with this operator we may also integrate matching with semantic and grey value information within one framework.

of manual intervention we accumulate evidence over a self generated stereo image sequence within a perception action cycle as described in the next subsection.

2.2.2 Accumulation of Object Representations in Stereo Image Sequences

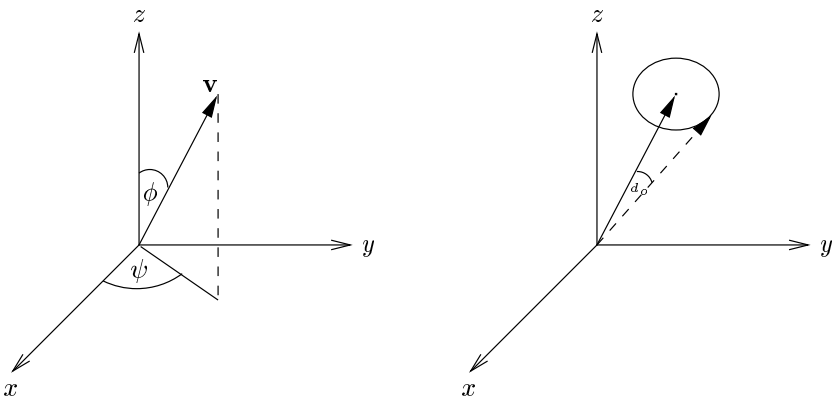

Figure 3: **Left:** Spherical coordinates. **Right:** Difference in orientation d_o

The object representation computed from the first stereo image pair consists of a list $\mathcal{L}$ of 3D line segments $l = (\mathbf{p}, \mathbf{v})$, i.e., a line segment is described by its position $\mathbf{p} = (x, y, z)$ and by the unit vector $\mathbf{v}$ indicating the orientation of the line segment (see figure 3 left). For these entities a metric $d(l, l')$ can be defined which gives low values for similar line segments and high values for dissimilar line segments.

Definition of the Metric: We define a metric between two line segments by evaluating their orientation difference d_o and spatial difference d_p. Given $l = (\mathbf{p}, \mathbf{v})$ and $l' = (\mathbf{p}', \mathbf{v}')$. The orientation difference is simply defined as

$$d_o(l, l') := \arccos(\mathbf{v} \cdot \mathbf{v}'),$$

i.e. as the angle between $\mathbf{v}$ and $\mathbf{v}'$ (see figure 3 right).

For the distance measure d_p we have, because of the aperture problem (see e.g. [13]), also to take the orientation of a line segment into account: The translation of a line segment along the axis spanned by $\mathbf{v}$ should not increase the distance between two line segments as long as it is less then half of the length of the line segment. In the following we define an elliptical unit sphere, i.e. we allow in the $\mathbf{v}$ direction a larger translation than orthogonal to $\mathbf{v}$ (see figure 4 left).

To compare $\mathbf{p}$ and $\mathbf{p}'$ we need the coordinates of $\hat{\mathbf{p}}'$ of $\mathbf{p}'$ in the coordinate system spanned by $\mathbf{p}$ and $\mathbf{v}$ ($\mathbf{p}$ be the origin and $\mathbf{v}$ the x–axis). For this we first translate $\mathbf{p}'$ by $-\mathbf{p}$ and then perform the rotation which maps $\mathbf{v}$ on the x–axis. This rotation can be well described by quaternions [2]. The rotation axis is

$$\mathbf{q}' = \frac{\mathbf{v} + \mathbf{e}_x}{||\mathbf{v} + \mathbf{e}_x||}$$

with $\mathbf{e}_x$ being the vector $(1, 0, 0)$. The rotation angle is π which yields the quaternion $\mathbf{q} = \cos\frac{\pi}{2} + \mathbf{q}' \sin\frac{\pi}{2}$ i.e., we have a simple reflection. Now we gain the coordinates of $\hat{\mathbf{p}}'$ in the system spanned by $\mathbf{p}$ and $\mathbf{v}$ by the formula

$$\hat{\mathbf{p}}' = \mathbf{q}(\mathbf{p}' - \mathbf{p})\bar{\mathbf{q}} = \begin{pmatrix} \hat{x}' \\ \hat{y}' \\ \hat{z}' \end{pmatrix}.$$

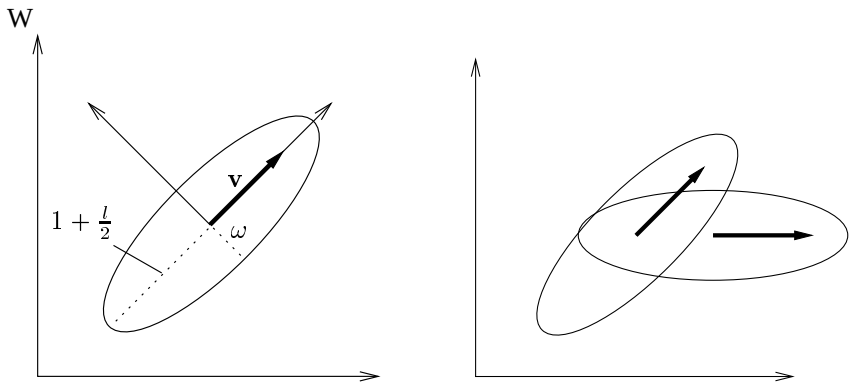

Figure 4: **Left:** Elliptical sphere in 2D. **Right:** Symmetry does not hold for f_p: The midpoint of the diagonal elliptical area is in the horizontal elliptical area but not the other way round.

We define a distance measure between l and l' (taking into account an elliptical deformation) by

$$f_p(l,l') := \sqrt{\left(\frac{1}{1+\frac{l}{2}}\hat{x}'\right)^2 + \hat{y}'^2 + \hat{z}'^2}.$$

Since this measure is not symmetric (see figure 4 right) we define

$$d_p(l,l') := \min\{f_p(l,l'), f_p(l',l)\}$$

as the final metric. Now we are able to say that line segment $\vec{l}$ and $\vec{l'}$ do correspond to each other when $d_o(\vec{l},\vec{l'})$ and $d_p(\vec{l},\vec{l'})$ are smaller then certain thresholds s_o and s_p.

Accumulation: A rigid body movement M of the robot can be described by six parameters $\vec{\beta} \in I\!R^6$, three describing translation and the others describing rotation. Let $M^{\vec{\beta}}(\mathcal{L})$ be the list of local line segments $\mathcal{L}$ representing the object moved by $M^{\vec{\beta}}$. Let $\mathcal{L}'$ be the list of local line segments extracted from a new stereo image pair. In this image pair the object is shown after a movement whose parameters $\vec{\beta}$ are known. For our algorithm the correspondences between the representations $\mathcal{L}'$ and $\mathcal{L}$ can easily be achieved by applying the rigid body motion $M^{\vec{\beta}}$ to the stored representation $\mathcal{L}$: $M^{\vec{\beta}}(\mathcal{L}) \approx \mathcal{L}'$ and comparison of the line segments by applying the above defined metric.

After achieving correspondences the two representations $M^{\vec{\beta}}(\mathcal{L})$ and $\tilde{\mathcal{L}}'$ can be merged by the accumulation scheme defined above: For each line segment l_j in $M^{\vec{\beta}}(\mathcal{L})$ we search for a line segment l'_k in $\mathcal{L}'$ which is close to l_j according to our metric d. If such a corresponding line segment has been found a value c_j, indicating the confidence of the system that l_j is part of the object, is increased, otherwise it is decreased. Line segments in $\mathcal{L}'$ to which no correspondences in $M^{\vec{\beta}}(\mathcal{L})$ do exist are included in the accumulated representation with only low confidences. After a couple of iterations with different views of the object the accumulated representation becomes more and more stable (see figure 1 right). It is even possible to segment objects from the background: Since the background is fixed and not changing according to the equivalent relation rigid body motion line segments corresponding to the background do vanish after a few iterations (see figure 5) and only line segments corresponding to the object and gripper remain.

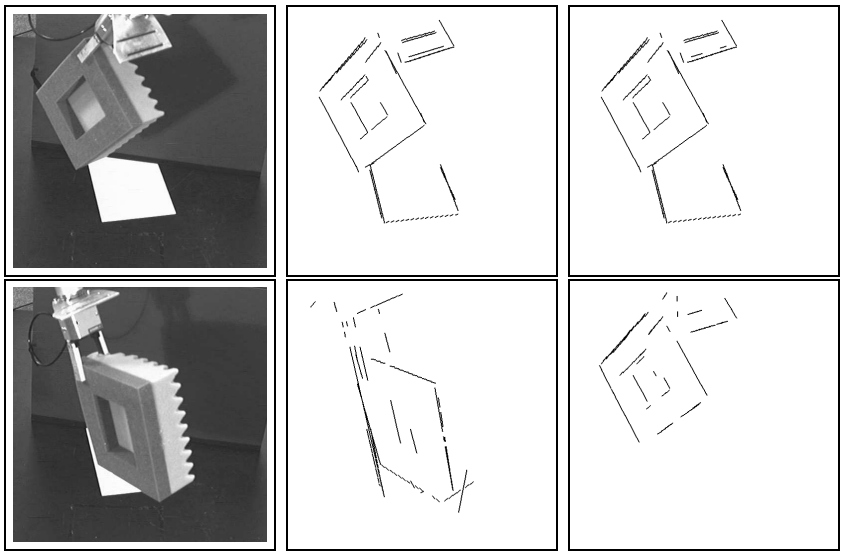

Figure 5: Accumulation of an object representation (first and fifth iteration). Line segments corresponding to the background vanish after a few iterations. Left: one of the stero images. Middle: Representation extracted from one stereo image pair. Right: Accumulated representation.

3 Conclusion and Outlook

We showed that our algorithm is able to accumulate autonomously representations utilizing self–controlled movements within a perception–action cycle. For the future a robot systems equipped with the ability to extract efficient object representations in a normal environment promises more flexible applications of robot vision systems. Instead of being equipped with manually defined representations the robot may use its own ability as a basis for manipulation and recognition. An important pre stage of our algorithm would be a behavior which allows to achieve haptic control over new objects and which positions the robot arm and the camera such that the accumulation process can start. We are currently implementing such a basic competence.

Also the integration of additional cues such as optic flow, color, texture and haptic cues is part of our current and future research. Our accumulation algorithm can be applied to all of these entities as long as certain the equivalance relation can be applied to them and a metric can be defined for them. A further important step is the application of the accumulated representation for matching or tracking tasks. In this contex, a promising method for pose estimation has been defined in our group [18] which we aim to apply to our representations.

Finally we aim to build a system equipped with some basic competencies (such as the introduced accumulation scheme) which starts a bootstrapping process in which knowledge of the world is extracted by own motivation and experience. We think that gaining haptic control over the object is one essential prerequisite of such a system in which perception and action have to be closely connected to support each other.

Acknowledgment

We would like to thank Daniel Grest, Marco Hahn, Bodo Rosenhahn and Daniel Wendorff whose work at the software library KiViGraP was very helpful for our simulations. For technical support we would like to thank Gerd Diesner and Henrik Schmidt.

References

[1] M. Ackermann. Akkumulieren von Objektrepräsentationen im Wahrnehmungs–Handlungs Zyklus. *Christian–Albrechts Universität zu Kiel, Institut für Informatik und praktische Mathematik (Diplomarbeit)*, 2000.

[2] W. Blaschke. *Kinematik und Quaternionen.* VEB Deutscher Verlag der Wissenschaften, 1960.

[3] R.A. Brooks. Intelligence without reason. *International Joint Conference on Artificial Intelligence*, pages 569–595, 1991.

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

[5] Michael Felsberg and Gerald Sommer. Structure multivector for local analysis of images. *Christian Albrechts Universität Kiel, Institut für Informatik. Technical Report no. 2001*, 2000.

[6] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1995.

[7] J.J. Gibson. *The ecological approach to visual perception.* Boston, MA: Houghton Mifflin, 1979.

[8] Marco Hahn. Semiglobale Verfahren zur Generierung von Eckpunkthypothesen in 2D und 3D. *Christian–Albrechts Universität zu Kiel, Institut für Informatik und praktische Mathematik (Diplomarbeit)*, 1999.

[9] B. Jähne, editor. *Digitale Bildverarbeitung.* Springer, 1997.

[10] J.J Koenderink. Wechsler's vision: An essay review of computational vision by Harry Wechsler. *Ecological Psychology*, 4:121—128, 1992.

[11] N. Krüger. *Visual Learning with a priori Constraints (Phd Thesis).* Shaker Verlag, Germany, 1998.

[12] A. Lanitis, C.J. Taylor, and T.F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, July:743–756, 1997.

[13] H.A. Mallot. *Sehen und die Verarbeitung visueller Information.* vieweg, 1998.

[14] M. Pötzsch, N. Krüger, and C. von der Malsburg. A procedure for automatic analysis of images and image sequences based on two–dimensional shape primitives. *U.S. Patent Application*, 1999.

[15] J. Princen, J. Illingworth, and J. Kittler. An optimizing line finder using a hough transform algorithm. *Computer Vision, Graphics, and Image Processing*, 52:57–77, 1990.

[16] G. Sommer. Verhaltensbasierter Entwurf technischer visueller Systeme. *Künstliche Intelligenz*, 3:42–45, 1995.

[17] G. Sommer. Algebraic aspects of designing behaviour based systems. In G. Sommer and J.J. Koenderink, editors, *Algebraic Frames for the Perception and Action Cycle*, pages 1–28. Springer Verlag, 1997.

[18] G. Sommer, B. Rosenhahn, and Y. Zang. Pose estimation using geometric algebra. Dagstuhl–Seminar, to appear 2000.

[19] Alan L. Yuille. Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70, 1991.