Accumulation of 3D Object Representations within a Perception–Action–Cycle

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. Controlled movement of the object grasped by the robot enables us to find correspondences within an image sequence. Analogies to human perception and the possibilities of more flexible applications with less need for manual intervention are discussed.

1 Introduction

Model based vision systems usually apply manually designed object representations (see e.g., [Yuille, 1991] or [Lanitis et al., 1997]). Often these representations are constructed by CAD–tools (see e.g., [Hansen and Henderson, 1989]).

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 [Koenderink, 1992, Sommer, 1997]. 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 by it utilizing robot–controlled movement (see figure 1 and figure 3).

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. A situation never stands for itself but is embedded in a time continuum [Gibson, 1979]. Therefore an important constraint is the utilization of the coherence of objects during a rigid body motion which allows to accumulate information over time. Furthermore as additional constraint the statistical relations of the occurrence of events (Gestalt principles, see e.g., [Ellis, 1938]) are used by the human visual system to correct errors occurring on different levels of visual processing.

In this paper we suggest to accumulate object representations from image sequences applying both constraints (rigid–body motion and the Gestalt principle collinearity) mentioned above. We account for the vagueness of semantic information extracted from single images by as-

signing 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.

In this paper the correspondence problem is solved by making use of interaction of action and perception, the parameter of motion are known because the robot manipulates the object. Knowing the correspondences an algorithm can be applied to update and improve the object representation iteratively. The algorithm is an extension of an algorithm introduced in [Krüger, 1998, Pötzsch et al., 1999] which only has dealt with 2D representation and translational motion. At the end of the paper we describe analogies to human object perception and the perspective of more flexible applications of robots.

2 Extraction of Object Representations

Our algorithm can be divided into two parts, preprocessing and accumulation. 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 images in 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). In the following we give a short description of the algorithm, for details see [Ackermann, 2000].

2.1 Extraction of a 3D Representation from Stereo Images

In the preprocessing step a 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 [Princen et al., 1990]. 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 (see figure 2) by evaluating gradient information. In this way the Gestalt principle collinearity is realized: 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. 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' [Faugeras, 1993, Hahn, 1999] the correspondences do not uniquely define a 3D line segment and a 3D representation of parts of the object can not 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

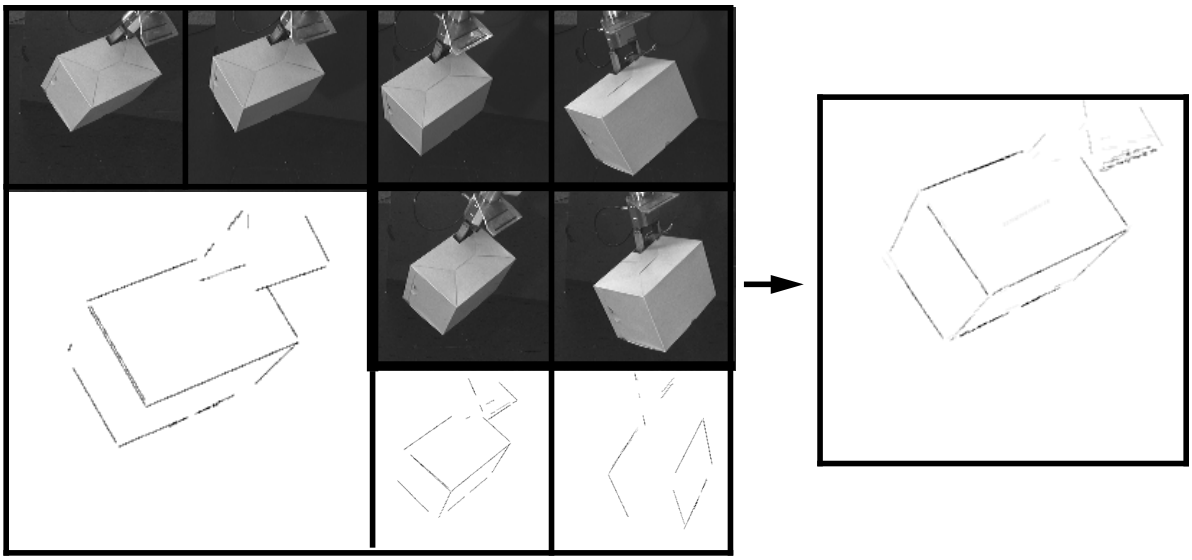

Figure 1: **left)** top: left and right image of an object. bottom: the projected 3D representation extracted from the stereo images. **middle)** Two pairs of stereo images (top: left camera image, middle: right camera image) and the the projected 3D representation (bottom). **right)** Projected 3D Representation 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.

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

2.2 Accumulation of Object Representations in Stereo Image Sequences

The object representation computed from the first stereo image pair consists of a list $\mathcal{L}$ of 3D line segments $l = (x, y, z, \alpha_1, \alpha_2, e)$, i.e., a line segment is described by its position (x, y, z) and two angles α_1 and α_2 describing its orientation by azimuth angles and its elongation e. 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. Here similarity is measured in space and orientation, i.e. in the parameters (x, y, z) and (α_1, α_2) and elongation e.

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 representing the object representation $\mathcal{L}$ moved by $\vec{\beta}$. Let $\tilde{\mathcal{L}}$ 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 $\tilde{\mathcal{L}}$ and $\mathcal{L}$ have to be known. This can be easily achieved by applying the rigid body motion $M^{\vec{\beta}}$ to the stored representation $\mathcal{L}$: $M^{\vec{\beta}}(\mathcal{L}) = \tilde{\mathcal{L}}$

After achieving correspondences the two representations $M^{\vec{\beta}}(\mathcal{L})$ and $\tilde{\mathcal{L}}$ can be merged by a simple update rule. Roughly speaking, for each line segment l_i in $M^{\vec{\beta}}(\mathcal{L})$ we search for a line segment $\tilde{l}_j$ in $\tilde{\mathcal{L}}$ which is close to l_i according to our metric d. If such a corresponding line segment has been found a value c_i, indicating the confidence of the system that l_i is part of the object, is increased, otherwise it is decreased. Line segments in $\tilde{\mathcal{L}}$ 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

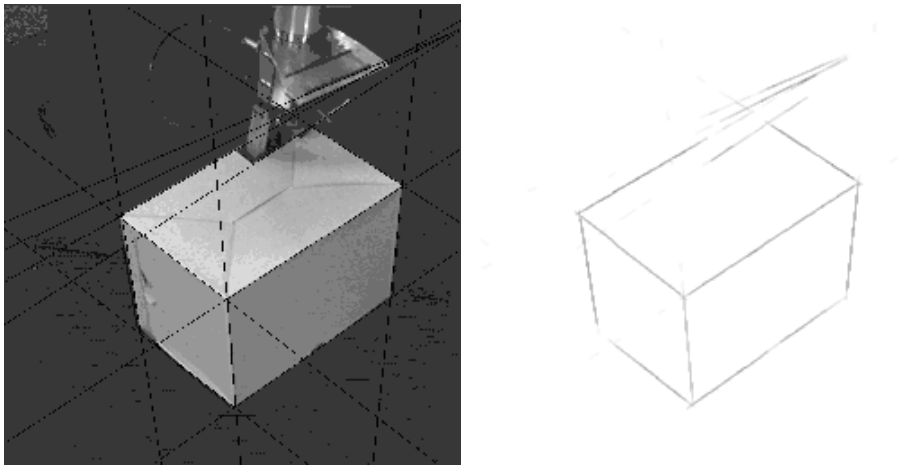

Figure 2: Left: Extracted Hough lines applying orientation sensitive Hough transformation. Right: Local line segments representing the object.

object the accumulated representation becomes more and more stable (see figure 1 right and figure 3 and 4).

3 Analogies to Human Perception

Our system shows an interesting analogy to human object learning. After approximately six months a baby is able to perform visually controlled movements of its arms. Then babies are in a position to produce controlled training data as we did with our robot and camera. Interestingly enough, the infants' concept of objects changes dramatically at this stage of development: babies younger than six months perceive an object as "something at a certain position" or "something moving with a certain velocity". Objects have no "above, below, left, right, in front or behind" [Bower, 1971]. After approximately six months the representation of objects starts to be based on form [Bower, 1971] and objects aquire permancy, i.e., objects continue to exist for the baby while being occluded [Piaget, 1976]. To our knowledge the relationship between self–controlled movements of objects and internal object representation has not been investigated in any detail. Nevertheless, the ability to create a situation in which an object appears under controlled conditions and in which correspondences can be achieved by making use of also body movement information may help, as in our system, to extract suitable representations of objects.

4 Conclusion

We showed that our algorithm is able to accumulate autonomously representations utilizing self–controlled movements. 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.

Acknowledgement

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

[Ackermann, 2000] Ackermann, M. (2000). Akkumulieren von Objektrepräsentationen

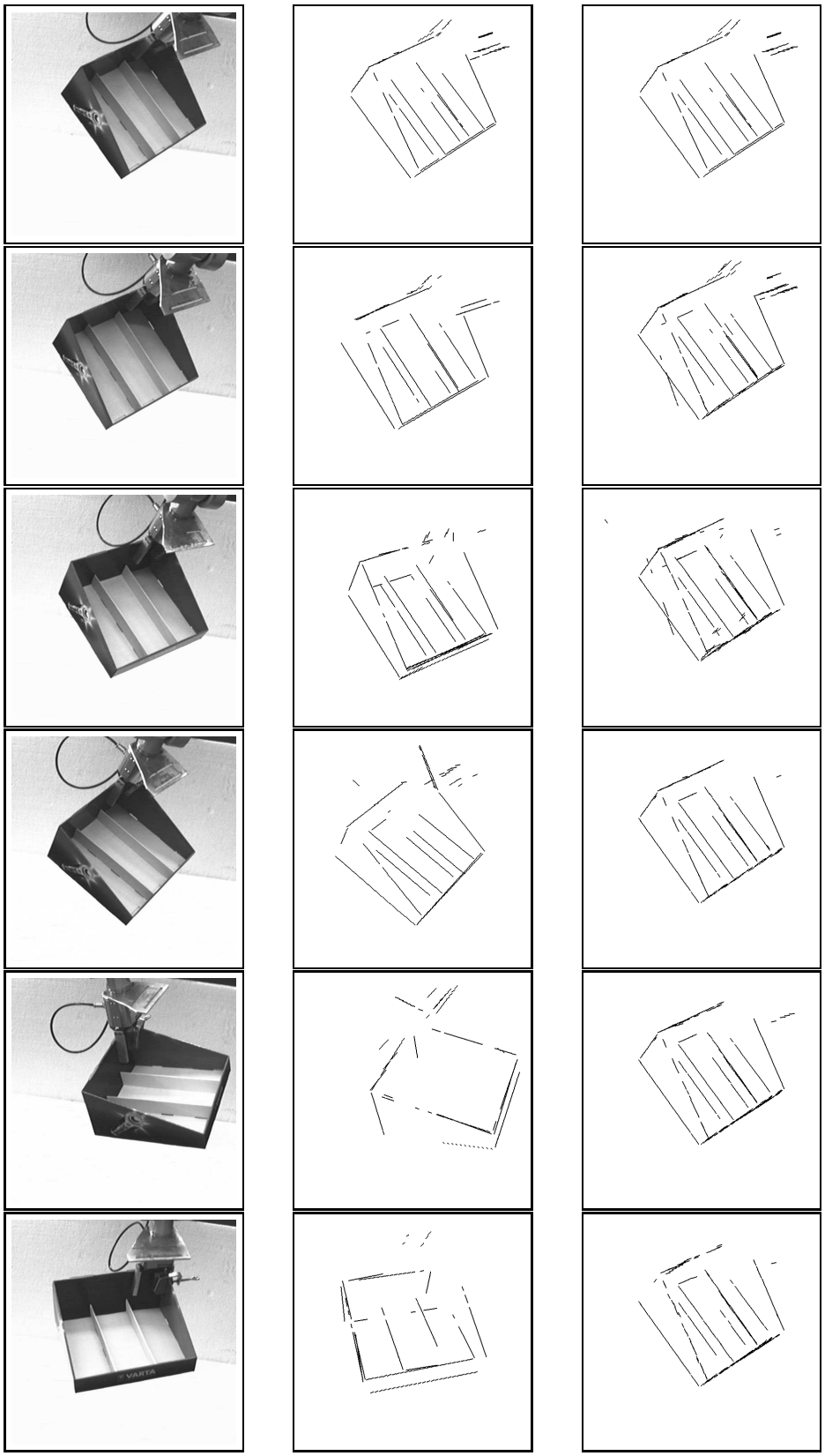

Figure 3: Learning of an object representation (Frame 1–6). Left: one of the stero images. Middle: Representation extracted from one stereo image pair. Right: Accumulated representation.

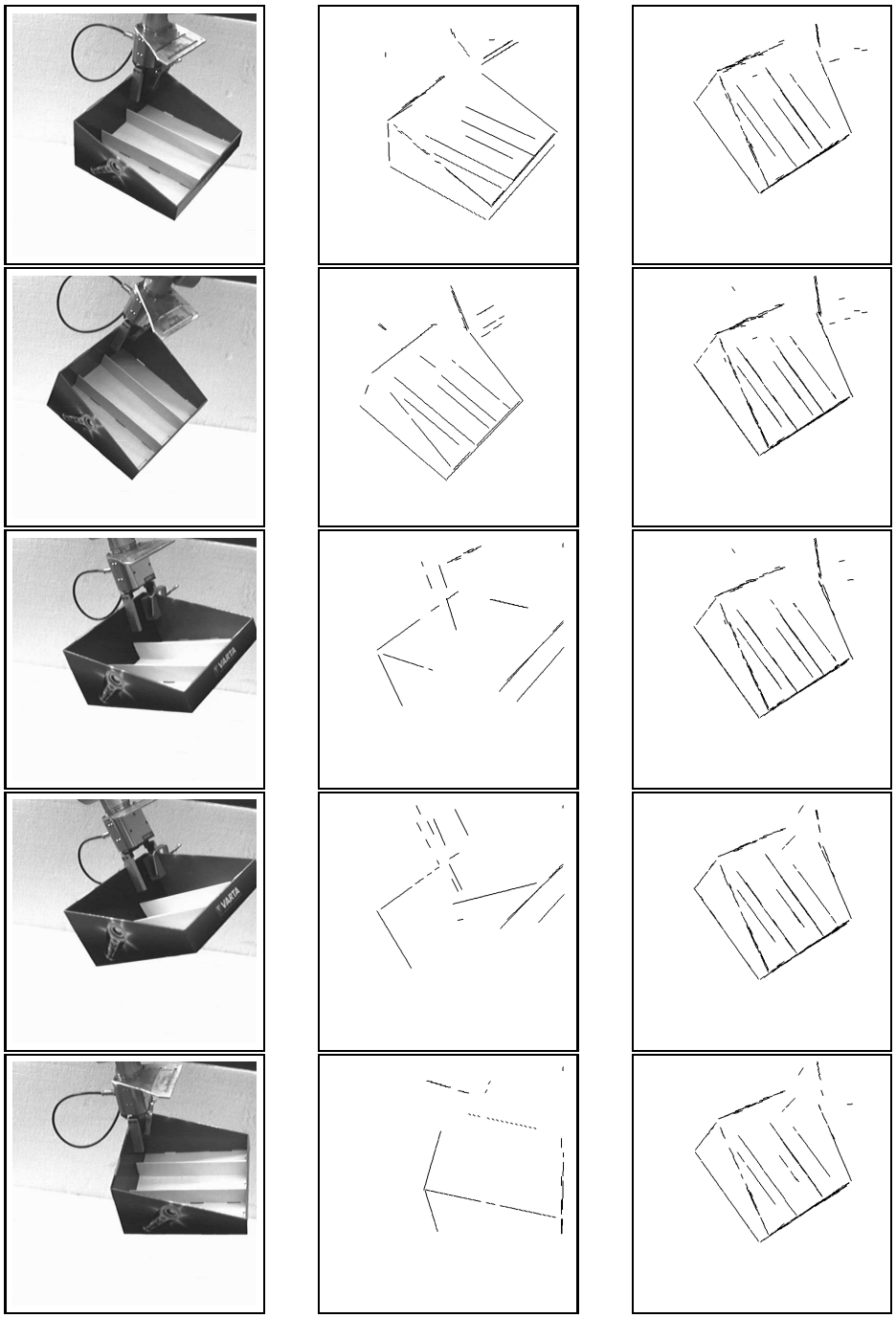

Figure 4: Learning of an object representation (Frame 7–11). Left: one of the stero images. Middle: Representation extracted from one stereo image pair. Right: Accumulated representation.

im Wahrnehmungs–Handlungs Zyklus. *Christian–Albrechts Universität zu Kiel, Institut für Informatik und praktische Mathematik (Diplomarbeit)*.

[Bower, 1971] Bower, T. (1971). The object in the world of the infant. *Scientific American*, 225:30–38.

[Ellis, 1938] Ellis, W., editor (1938). *Gestalt Theory, A source book for Gestalt Psychology*.

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

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

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

[Hansen and Henderson, 1989] Hansen, C. and Henderson, T. (1989). Cagd–based computer vision. *PAMI*, 11(11):1181–1193.

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

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

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

[Piaget, 1976] Piaget, J. (1976). *The psychology of intelligence*.

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

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

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

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