This work has been presented at the Cognitive Vision Workshop 2002 in Zürich

Extraction of Object Representations from Stereo Image Sequences Utilizing Statistical and Deterministic Regularities in Visual Data

Norbert Krüger[#], Thomas Jäger[+], Christian Perwass[+]
[#] University of Stirling, Scotland, norbert@cn.stir.ac.uk*
[+] University of Kiel, Germany, chp{thj}@ks.informatik.uni-kiel.de

Abstract

The human visual system is a highly interconnected machinery that aquires its stability through integration of information across modalities and time frames. This integration becomes possible by utilizing regularities in visual data, most importantly motion (especially rigid body motion) and statistical regularities reflected in Gestalt principles such as collinearity.

In this paper we describe an artificial vision system which extracts 3D–information from stereo sequences. This system uses deterministic and statistical regularities to aquire stable representations from unreliable submodalities such as stereo or edge detection. To make use of the above mentioned regularities we have to work within a complex machinery containing sub–modules such as stereo, pose estimation and an accumulation scheme. The interaction of these modules allows to use the statistical and deterministic regularities for feature disambiguation within a process of recurrent predictions.

1 Introduction

Vision, although widely accepted as the most powerful sensorial modality, faces the problem of an extremely high degree of vagueness and uncertainty in its low level processes such as edge detection, optic flow analysis and stereo estimation [1]. However, by integrating information across visual modalities (see, e.g., [9]), the human visual systems acquires visual representations which allows for actions with high precision and certainty within the 3D world even under rather uncontrolled conditions. The power of modality fusion arises from the huge number of intrinsic relations given by deterministic and statistical regularities across visual modalities. The essential need for fusion of visual modalities, beside their improvement as isolated methods, has also been recognised by the computer vision community during the last 10 years (see, e.g., [1, 3]).

Two important regularities in visual data with distinct properties are *motion* (most importantly rigid body motion, RBM, see, e.g., [5]) and *statistical interdependencies* between features such as collinearity and symmetry (see, e.g., [23]).[1] RBM reflects a geometric dependency in the time-space continuum. If the 3D motion between two frames is known then feature prediction can be postulated since the change of the position and the semantic properties of features can be computed (see, e.g., [14]). This can be done by having physical control over the object (as in [14]) or by *computing the RBM* as done in this paper. A computation of the RBM makes the system

[1] There exists evidence that abilities based on rigid body motion are to a much higher degree hard coded in the human visual system than abilities based on statistical interdependencies (for a detailed discussion see [17]).

more flexible since it allows for acquiring object knowledge by watching the object without grasping is. This is also one of the main contributions of this paper compared with [14]. However, having physical control over the object might also have advantageous in specific situation, e.g., when the RBM is controlled in such a way that especially cognitive interesting situation are created.

However, computation of RBM is a non trivial problem. A huge amount of literature is concerned with its estimation from different kinds of feature correspondences (see, e.g., [20, 22]), which are most commonly point and/or line correspondences. In our system, correspondences are established by optic flow. However, one fundamental problem of RBM-estimation is that methods are in general very sensitive to outliers. The pose estimation algorithm we do apply [22] computes the rigid body motion presupposing a 3D model of the object and a number of correspondences of 3D–entities between the object model and their projections in the consecutive frame. In [22] a manually designed 3D model was used for pose estimation. Here, we want to replace this prior knowledge by substituting the manually created model by 3D information extracted from stereo. However, by using stereo we face the above mentioned problems of uncertainty and reliability of visual data as described above. Because of the sensitivity of pose estimation to outliers in the 3D–model we need to compensate these disturbances. We can sort out unreliable 3D–features by applying a grouping mechanism based on *statistical interdependencies* in visual data.

Once the RBM across frames is known (and for the computation of the RBM we need a quite sophisticated machinery) we can utilize and a scheme which uses the *deterministic regularity* RBM to disambiguate 3D entities over consecutive frames [14].

2 Visual Sub-modalities

Our system acquires stable representations from stereo image sequences by integrating the following visual sub–modalities: edge detection based on the monogenic signal [6], a new stereo algorithm which makes use of geometric and appearance based information [15], optic flow [19], pose estimation [22] and an accumulation scheme which extracts stable representations from disturbed data over consecutive frames [14]. An overview of the system is given in figure 1.

At this point, we want to stress the difference between two different sources of disturbances:

- Outliers: 3D entities caused by wrong stereo correspondences. They have an irregular non–Gaussian distribution (see figure 3 (top row))
- Feature inaccuracy: Deviation of parameters of estimated 3D entities (e.g., 3D orientation and 3D position) caused by unreliable position and orientation estimates in images. This kind of disturbance can be expected to have Gaussian like distribution with its mean close to the true value.

Both kinds of disturbances have distinct distribution and the visual modules have a different sensitivity to both errors: for example, while outliers can lead to a completely wrong estimation of pose, feature inaccuracy would not distort the results of pose estimation that seriously.

We will deal with these two kinds of disturbances in distinct ways: *Outliers* are sorted out by a filtering algorithm utilizing the statistical interdependency "collinearity" in 3D and by a process of recurrent predictions based on rigid body motion estimation. Both processes *modify confidences* associated to features. *Feature inaccuracy* becomes reduced by *merging* corresponding 3D line segments over consecutive frames. During the merging process semantic parameter (here 3D–position and 3D–orientation) are iteratievely adapted.

In the following we briefly introduce the applied sub–modalities and their specific role within the whole system.

Feature extraction: Edge detection and orientation estimation is based on the isotropic

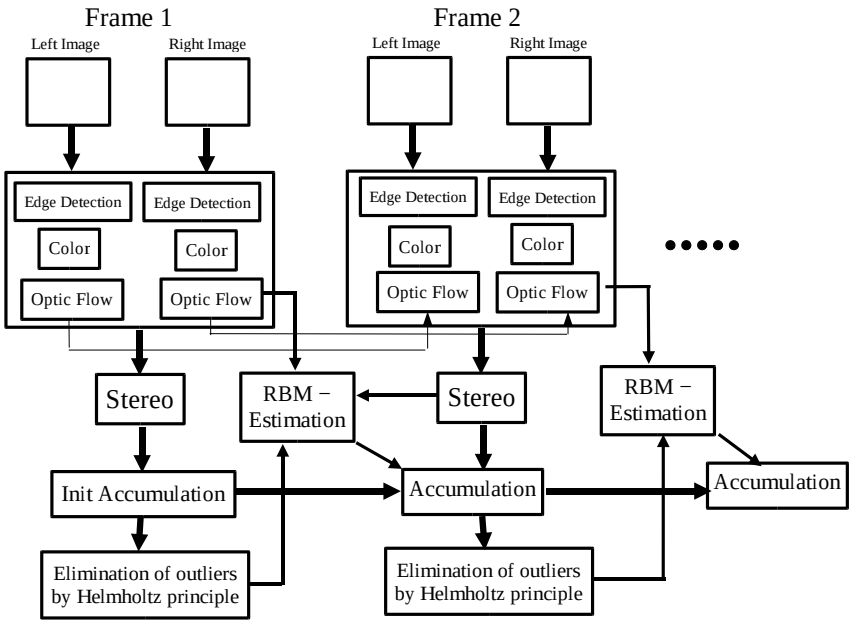

Figure 1. Scheme of Interaction of visual sub–modalities

linear filter [6] and on phase congruence over neighbouring frequency bands (see e.g., [13]). The applied filter [6] performs a *split of identity*: it orthogonally divides an intrinsically one-dimensional bandpass filtered signal in its energy information (indicating the likelihood of the presence of a structure), its geometric information (orientation) and its contrast transition expressed in the phase (called 'structure' in [6]). Furthermore, we use structural information in form of color averaged at the left and right side of the edge separately. Figure 2 shows the results of preprocessing.

Stereo: In stereo processing with calibrated cameras we can reconstruct 3D points from two corresponding 2D points by computing the point of intersection of the two projective lines generated by the corresponding image points and the optical centers of the camera. However, most meaningful image structure is intrinsically one-dimensional [24], i.e., is dominated by edges or lines. Orientation at intrinsically one-dimensional image structures can be estimated robustly and precisely by various methods (see, e.g., [11]). Therefore, it makes sense to use orientation information also for the representation of visual scenes: from two corresponding 2D points with associated orientation we can reconstruct a 3D point with associated 3D orientation (in the following called '3D–line segment'). A more detailed description can be found in [14].

To find stereo correspondences in the left and right image we can use geometrical as well as structural information in form of phase and color. In [15] we can show that both factors are important for stereo–matching and that the optimal result is achieved by *combining* both kinds of information.

Note that out stereo algorithm does not make use of the ordering constraint (such as , e.g., in [21]) but that the system can also be used in case of depth continuities. The only restriction is the occurence of local line segments in the scene.

The basic feature we extract from the stereo module is a 3D line segment coded by its mid–point (x_1, x_2, x_3) and its 3D orientation coded by two parameter (θ, ϕ). Furthermore a confidene c is associated to the parametric description of the 3D entity. We therefore can formal-

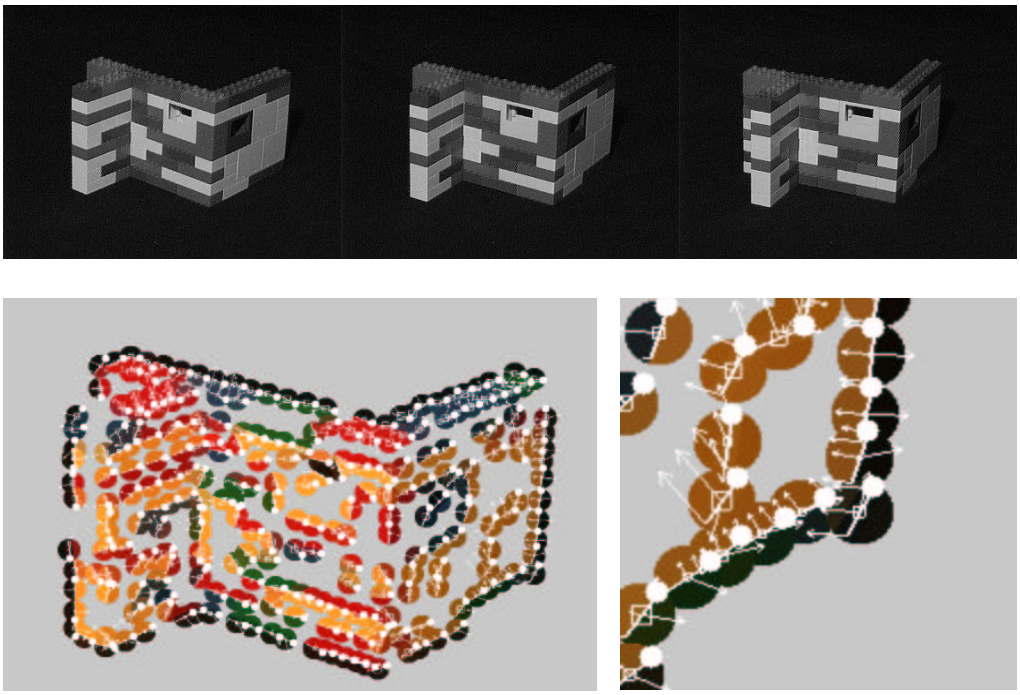

Figure 2. Top: Three images of an image sequence. Bottom: Feature processing (left: complete image, right: Sub-area). Shown are the orientation (center line), phase (single arrow), color (half moon on the left and right side of the edge) and optic flow (three parallel arrows).

ize a 3D–line segment by

$$\mathbf{l} = ((x_1, x_2, x_3), (\theta, \phi); c). \qquad (1)$$

All parameters are subject to modifications by contextual information (as described below) utilizing the Gestalt law Collinearity and the regularity RBM across frames.

Pose estimation: To be able to predict 3D–features in consecutive frames, we want to track an object in a stereo image sequence. More precisely, we want to find the rigid body motion from one frame to the consecutive frame. To compute the rigid body motion we apply the pose estimation algorithm [7] which requires a 3D model of the object as well as correspondences of image entities (e.g., 2D line segments) with 3D object entities (e.g., 3D line segments)[2]. A 2D–3D line correspondence defines a constraint on the set of possible rigid body motions that (using a linear approximation of a rigid body motion [7]) can be expressed by two linear equations. In combination with other constraints we get a set of linear equations for which a good solution can be found iteratively [7] using a standard least square optimization algorithm.

Optic flow: The 3D model of the object is extracted by our stereo algorithm. Correspondences between 3D entities (more precisely by their 2D projections respectively) and 2D line segments in the consecutive frame are found by the optic flow. After some tests with different optic flow algorithms (see [10]) we have chosen the algorithm developed by Nagel [19] which showed good results especially at intrinsically 1D structures. Correspondences are established by simply moving a local line segment according to its associated optic flow vector.

[2]This pose estimation algorithm has the nice property that it can combine different kinds of correspondences (e.g., 3D point–3D point, 2D point–3D point, and 2D line–3D line correspondences) within one system of equations. This flexible use of correspondences makes it especially attractive for sophisticated vision systems which process multiple kinds of features such as 2D junctions, 2D line segments or 3D points.

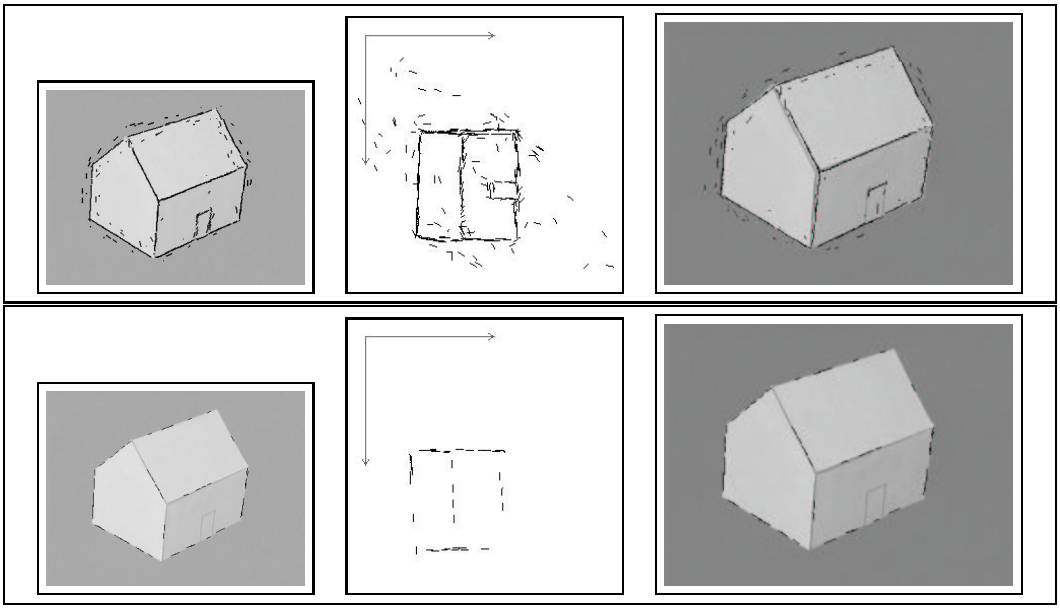

Figure 3. Top: Using the stereo module without ellimination procedure. Left: Projection onto the image. Middle: Projection onto the xz plane. Note the large number of outliers. Right: Pose estimation with this representation. Note the deviation of pose from the correct position. Bottom: The same after the ellimination process. Note that all outliers could be elliminated by our collinearity criterion and that pose estimation does improve.

Using collinearity in 3D to eliminate outliers: The pose estimation algorithm is sensitive to outliers since these outliers can dominate the over–all error in the objective function associated with the equations established by the geometric constraints. We therefore have to ensure that no outliers are used for the pose estimation.

According to the *Helmholtz Principle*, every large deviation from a "uniform noise" image should be perceivable, provided this large deviation corresponds to an *a priori* fixed list of geometric structures (see [4]). The *a priori* geometric structure we do apply to eliminate wrong 3D–correspondences are *collinear structures in 3D*: We assume that (according to the Helmholtz principle) a local 3D line segment that has many neighbouring collinear 3D line segments is very unlikely to be an outlier and we only use those line segments for which we find at least a couple of collinear neighbours. More precisely, we lower the confidence c in (1) for all line segments that have only few collinear neighbours. Figure 3 (middle) shows the results of the elimination process for a certain stereo image). We can show that the elimination process improves pose estimation (see figure 3 (right). For a more in depth discussion about applying Gestalt principles within our system see [18].

Acquisition of object representations across frames: Having extracted a 3D representation by the stereo module and having estimated the RBM between two frames we can apply an accumulation scheme (for details see [14]) which uses correspondences across frames to accumulate confidences for visual entities. Our accumulation scheme is of a rather general nature. Confidences associated to visual entities are increased when correspondences over consecutive frames are found and decreased if that is not the case. By this scheme, only entities which are validated over a larger number of frames (or for which predictions are often ful-

filled) are considered as existent while outliers can be detected by low confidences (in Figure 4 a schematic representation of the algorithm for two iterations is shown). Since the change of features under an RBM can be computed explicitly (e.g., the transformation of the square to the rectangle from the first to the second frame), the rigid body motion can be used to predict the correspondences (see also [14]).

This accumulation scheme presupposes a metrical organisation of the feature space. If we want to compare visual entities derived from two frames even when we know the exact transformation corresponding to the rigid body motion, the corresponding entities cannot be expected to be exactly the same (the two squares in figure 4 are only similar not equal) because of factors such as noise during the image acquisition, changing illumination, non–Lambertian surfaces or discretization errors. Therefore it is advantageous to formalize a measure for the likelihood of correspondence by using a metric (for details see [14]). Once a correspondence is established we apply an update rule on the confidence c as well as the semantic parameters (x_1, x_2, x_3) and (θ, ϕ). for the confidence and semantic properties of the line segment (for details see [14] and [10]). That means that by the accumulation scheme our 3D line segments are embedded in the time domain, *they represent features in 3D-space and time.*

Integration of visual sub-modalities: The recurrent process based on the sub-modalities described above is organised as shown in figure 1. For each frame we perform feature extraction (edge detection, optic flow) in the left and right image. Then we apply the stereo algorithm and the elimination process based on the Helmholtz principle. Using the improved accumulated model (i.e., after eliminating outliers), we apply the pose estimation module which uses the stereo as well as the optic flow information. Once the correct pose is computed, i.e., the RBM between the frames is known we transform the 3D entities extracted from one frame to the consecutive frame based on the known RBM (for details see [14]).Then we are able to perform one further iteration of the accumulation scheme.

We have applied our system to different image sequences, one of them is shown in figure 2. Figure 5 (left) shows the results. At the top the extracted stereo representation at the first frame is shown while at the bottom the accumulated representation after 6 frame is shown. We see that the number of outliers can be reduced significantly. In figure 5 (right) the mean difference of the semantic parameters (3D–position and 3D–orientation) from a ground truth (manually measured beforehand) is shown. We see that the difference between the extracted representation (consisting of line segments with high confidence) compared to the ground truth for position and orientation decreases during accumulation. Further simulations can be found in [10].

3 Summary and Discussion

We have shown that through integration of different visual modalities we are able to extract reliable object representation from disturbed low level processes. Since we want to make use of the regularity RBM across frames we need to use a complex mechanism (which uses different sub–modules) that allows to compute the RBM. This mechanism also made use of statistical regularities to elliminate outliers for pose estimation. Our feature representation allows for a modification of features depending on contextual information. The confidence c codes the likelihood of the existence of the visual entity while semantic parameters describe properties of the entity. Both kind of descriptors are subject to modification by contextual information, i.e., by the statistical and deterministic regularities coded within the system.

Our system has some interesting properties compared to other systems. Firstly, differering to classical structure from motion approaches (see, e.g., [12, 5]) we do not intend to aquire 3D–information only but we are interested in attributes that are relevant for perceptive tasks. Here, our representations consist

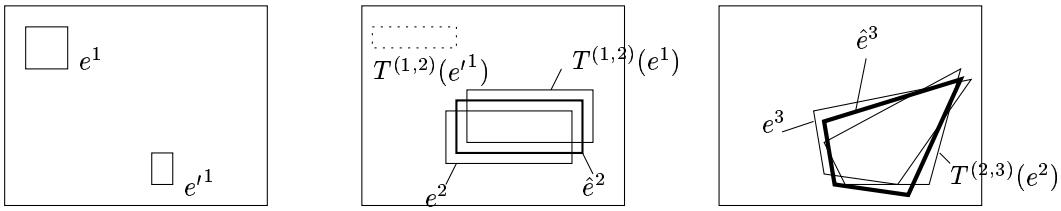

Figure 4. The accumulation scheme. The entity e^1 (here represented as a square) is transformed to $T^{(1,2)}(e^1)$. Note that without this transformation it is barely possible to find a correspondence between the entities e^1 and e^2 because the entities show significant differences in appearance and 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. In contrast, the confidence assigned to e'^1 is decreased because no correspondence in the second frame is found. The same procedure is then applied for the next frame for which again a correspondence for e^1 has been found while no correspondence for e'^1 could be found. The confidence assigned to e^1 is increased once again while the confidence assigned to e'^1 is once again decreased (the entity has disappeared). By this scheme information can be accumulated to achieve robust representations.

of line segments. We use these representation for the task of tracking of objects. Representations based on line segments have also been used for object recognition (see, e.g., [16]). We have applied our accumulation scheme to geometric 3D entities. However, this scheme is generic and we intend to apply our accumulation scheme to other visual domains (such as color, texture or other appearance based information) to extract richer and more powerfull object representations.

Secondly, in our representations semantic properties of features and their reliability are explicitely coded. Both, semantic properties and the reliability are subject to contextual influences. The integration of contextual information and its modelling by recurrent processes that modify reliabilities is the central aim of our current project and our method differs in that respect to classical structure from motion algorithms (see, e.g., [8]). Here, we have used the reliability information to improve the RBM estimation by picking out certain 'good' feature constellations from the big feature pool. This way to handle outliers works complementary to other well established methods such as RANSAC ([2]). Coding information with its reliability allows to keep hypotheses that are (looking at them locally) unlikely but may become likely taking the context into account.

Acknowledgement: We would like to thank Gerald Sommer and Florentin Wörgötter for fruitful discussions. Furthermore, we would like to express our thankfulness for the work of the students at the Univeristy of Kiel who have been involved in this project.

References

[1] J. Aloimonos and D. Shulman. *Integration of Visual Modules — An extension of the Marr Paradigm.* Academic Press, London, 1989.

[2] Fischlerand Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 1981.

[3] A. Cozzi and F. Wörgötter. Comvis: A communication framework for computer vision. *International Journal of Computer Vision*, 41:183–194, 2001.

[4] A. Desolneux, L. Moisan, and J.M. Morel. Edge detection by the Helmholtz principle. *JMIV*, 14(3):271–284, 2001.

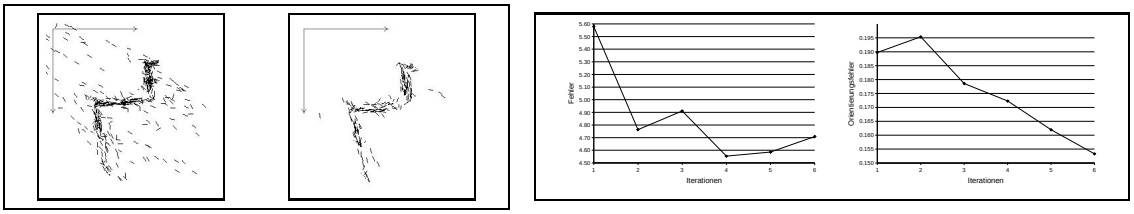

Figure 5. Left: Projection of representations (extracted from the image sequence shown in figure 2) on xz-plane at the beginning of accumulation (left) and after 6 steps of accumulation (right). Right: Deviation of 3D position (left) and 3D orientation (right) from the ground truth during accumulation. Estimation of both semantic parameters improves during accumulation.

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

[6] M. Felsberg and G. Sommer. The monogenic signal. *IEEE Transactions on Signal Processing*, 41(12), 2001.

[7] O. Granert. Poseschaetzung kinematischer ketten. *Diploma Thesis, Universität Kiel*, 2002.

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

[9] D.D. Hoffman, editor. *Visual Intelligence: How we create what we see*. W.W. Norton and Company, 1980.

[10] Thomas Jäger. Interaktion verschiedener visueller Modalitäten zur stabilen Extraktion von objektrepräsentationen. *Diploma thesis (University of Kiel)*, 2002.

[11] B. Jähne. *Digital Image Processing – Concepts, Algorithms, and Scientific Applications*. Springer, 1997.

[12] R. Klette, K. Schlüns, and A. Koschan. *Computer Vision - Three-Dimensional Data from Images*. Springer, 1998.

[13] P. Kovesi. Image features from phase congruency. *Videre: Journal of Computer Vision Research*, 1(3):1–26, 1999.

[14] N. Krüger, M. Ackermann, and G. Sommer. Accumulation of object representations utilizing interaction of robot action and perception. *Knowledge Based Systems*, 13(2):111–118, 2002.

[15] N. Krüger, M. Felsberg, C. Gebken, and M. Pörksen. An explicit and compact coding of geometric and structural information applied to stereo processing. *Proceedings of the workshop 'Vision, Modeling and VISUALIZATION 2002'*, 2002.

[16] N. Krüger and G. Peters. Orassyll: Object recognition with autonomously learned and sparse symbolic representations based on metrically organized local line detectors. *Computer Vision and Image Understanding*, 77:49–77, 2000.

[17] N. Krüger and F. Wörgötter. Different degree of genetical prestructuring in the ontogenesis of visual abilities based on deterministic and statistical regularities. *Proceedings of the Ẃorkshop On Growing up Artifacts that LiveŚAB 2002*, 2002.

[18] N. Krüger and F. Wörgötter. Multi modal estimation of collinearity and parallelism in natural image sequences. *to appear in Network: Computation in Neural Systems*, 2002.

[19] H.-H. Nagel. On the estimation of optic flow: Relations between different approaches and some new results. *Artificial Intelligence*, 33:299–324, 1987.

[20] T.Q. Phong, R. Haraud, A. Yassine, and P.T. Tao. Object pose from 2-d to 3-d point and line correspondences. *International Journal of Computer Vision*, 15:225–243, 1995.

[21] M. Pollefeys, R. Koch, and L. van Gool. Automated reconstruction of 3d scenes from sequences of images. *Isprs Journal Of Photogrammetry And Remote Sensing*, 55(4):251–267, 2000.

[22] B. Rosenhahn, N. Krüger, T. Rabsch, and G. Sommer. Automatic tracking with a novel pose estimation algorithm. *Robot Vision 2001*, 2001.

[23] S. Sarkar and K.L. Boyer. *Computing Perceptual Organization in Computer Vision*. World Scientific, 1994.

[24] C. Zetzsche and E. Barth. Fundamental limits of linear filters in the visual processing of two dimensional signals. *Vision Research*, 30, 1990.