TOWARDS AN IMAGE UNDERSTANDING ARCHITECTURE FOR A SITUATED ARTIFICIAL COMMUNICATOR*

C. Bauckhage, G.A. Fink, G. Heidemann, N. Jungclaus, F. Kummert, S. Posch, H. Ritter, G. Sagerer, D. Schlüter

University of Bielefeld, Faculty of Technology

P.O.Box 100131, 33501 Bielefeld, Germany

Tel.: +49 521 106 2935, Fax: +49 521 106 2992

e-Mail:

{cbauckha|gernot|gheidema|nils|franz|posch|helge|sagerer|dschluet}@techfak.uni-bielefeld.de

*This research was partially supported by the German Research Foundation (DFG) within the Collaborative Research Center "Situated Artificial Communicators" (SFB 360).

Abstract

In this paper we propose an architecture of an image understanding system for a situated artificial communicator realizing human-machine interaction. Starting with sensor input the processing is initially carried out in separate pathways using different schemes of image segmentation. Subsequently, a hybrid technique for 2D-object recognition is employed. The final model based 3D-reconstruction yields a 3D-scene representation. Intermediate results are linked over time in memory moduls to enhance efficiency of processing on image sequences. Results of the individual moduls will be presented and discussed.

1 Introduction

The goal of the project we are jointly working on with researchers in the field of pattern recognition, artificial intelligence, and linguistics is to study advanced human-machine interaction. The machine should be able to process acoustic and visual input and react meaningfully by producing speech output or by manipulating objects in the environment of the communicating partners. This device is called a *"situated artificial communicator"*.

The domain was chosen to be the cooperative construction of a toy-airplane with parts from a wooden construction-kit for children. In the first phases of the project the artificial situated communicator is to act as service robot carrying out simple tasks specified by the human instructor. It is not yet intended to plan the construction process. It is, however, intended to recognize acoustically or visually referenced complex objects from the task domain, provide information about its current understanding of the environment, and perform basic manipulations.

2 Architecture

A situated artificial communicator as outlined in the last section constitutes a complex system of interacting components combining different modalities as vision, speech, and actuators. The system has to act and react in a situated way within a evolving environment. It has to cope with uncertainty and errors of perception results and world models as well as with different time scales and overlapping capabilities of the various modules. Fig. 1

Figure 1 shows the current architecture of the visual system using a static stereo camera as

sensors. Here we follow the visual pathways proposed by van Essen [1]. The main purpose of our system is the recognition of visible objects including their temporal correspondence and three-dimensional pose. Starting with the sensor input the processing is initially carried out in two separate pathways. The first one uses intensity information for contour based segmentation into straight lines and elliptical arcs and subsequent grouping into more abstract image primitives. The second pathway relies on color information to segment the input images into homogeneous regions. Results of both processing paths are combined matching contour groups to regions when caused by the same event in the scene. The subsequent hybrid 2D-object recognition is based mainly on regions, while 2D object hypotheses with the associated contour groups are exploited by the final model based 3D-reconstruction. At this stage results of the so far separately processed stereo images are combined to enhance 3D reconstruction. For region and 2D object hypotheses a memory was realized linking the results over time. This supports on the one hand the recognition of events and actions and allows on the other hand an efficient processing of image sequences.

3 Realization

The architecture outlined above was implemented as a distributed system consisting of several independent modules. The inter-module communication is realized using the *Distributed Applications Communication System* [2]. In the following we will shortly describe the system's components for color-based region segmentation, contour-based grouping, matching of region and contour information, hybrid 2D-object recognition, and model-based 3D-reconstruction.

Region Segmentation Region segmentation starts by classifying every single pixel of the YUV-image into one of the twelve fixed colors of our domain using a polynomial classifier of sixth degree. To speed up this process a look-up table is used where for every combination of YUV-values the corresponding pre-calculated classification result is stored. Subsequently a smoothing operation is applied deciding for the color with maximum occurance within a window. Finally for every region of identical color a set of form-parameters is calculated.

Contour-Based Grouping In addition to region segmentation the system exploits discontinuities of the intensity image in a separate pathway. Subsequent to initial edge detection, straight line segments and elliptical arcs are approximated using the method of Leonardis [3]. These are used to define a hierarchy of grouping hypotheses with growing complexity using the Gestalt laws of proximity, good continuation, symmetry, and closure. In the lowest level contour segments are combined to form collinear and curvilinear groups which are again approximated with straight line segments or elliptical arcs. In addition, pairs of contour segments or linear groups may form a proximity grouping. The next 2×1D level deals with pairs of symmetric or parallel linear groups. The last level organizes linear groups into closed contours.

In the first stage of the grouping process, grouping hypotheses are generated within this hierarchy taking only local evidence into account. For the first two levels of the hierarchy we define the concept of *Areas of Perceptual Attentiveness* introduced in [4]. These areas are derived from a hand labelled training set of our domain and restrict relative distance and position when grouping two contour segments or groupings. In addition, local criteria like orientation difference are employed to restrict potential grouping. To hypothesize closed contours of the highest level of the hierarchy, a *proximity graph* is constructed from all contour segments, col- and curvilinear groupings as nodes and proximities between them as edges. From this graph closed contours are generated searching for simple cycles where the underlying contour segments do not intersect.

To obtain a set of global consistent groupings the results are judged in a global context using a Markov Random Field in the second stage (see [5]). Each grouping hypothesis corresponds to a node (or site) of this graph with an associated random variable representing the significance (or correctness) of the associated interpretation. Therefore, in contrast to most other approaches using MRFs, different sites may interpret a common subset of the image data, while the neighborhood system of the MRF represents support and competition of the groupings hypotheses within the hierarchy. With appropriately defined clique potential the energy of the MRF is minimized yielding a global judgement of the grouping hypotheses.

Matching of Regions and Contours Matching of regions to contour groupings and vice versa is integrated to resolve the drawbacks inherent to both segmentation schemes and to enhance the performance of the object recognition and contour-based grouping. Matches describe a relation between primitives of both segmentation techniques in the meaning of being caused by the same event in the scene, e.g. the projection of the same surface of an object. They are distinguished into boundary and structural matches. Boundary matches establish relations with groupings approximating the boundary of a region, whereas structural matches relate groupings based on the structure (or texture) of an object surface to the region approximating that surface. Generally, the matching process is controlled by calculating the endpoint distance and average distance of the contour group to the region boundary [6]. To this end, a parameter set is estimated on a hand labeled training set for both types of matches.

Within the grouping process only boundary matches are used. The groupings matching to the same region boundary facilitate the generation of additional grouping hypotheses, e.g. proximity groupings along the region boundary, and augment the judgment in the Markov Random Field. For object recognition, closed contours and boundary matches serve as alternative contour based regions to resolve errorneous region segmentation. Additionally, structural matches will be included to enhance the object recognition in cases, where the color classification based region segmentation looses the internal structure on an object surface.

Object Recognition For the recognition of isolated parts from our construction-kit we use a hybrid approach combining the semantic network language ERNEST [7, 8] with a holistic object recognition module. The latter generates hypotheses on the object's identities in mainly three processing steps as described in [9]: First, the objects are located by a colour segmentation. Then from the segmented regions features are extracted using an optimized set of Gabor-filters. In the third stage, these features are classified by a neural network of the Local-Linear-Map (LLM) type. For each competing LLM-hypothesis a so-called holistic instance inside the semantic network is created which are stored in competing search tree nodes. Dependent on the object type detected by the LLM-network an appropriate specialization is selected to verify the object hypothesis according to the structural knowledge stored in the semantic network [10].

During this instantiation process restrictions for position, color, and shape are propagated in a model–driven way. To increase the robustness of our analysis and to speed up processing we extended this approach to image sequences where on the basis of previous structural results the current image is analyzed [11]. So temporally linked results are calculated supporting the recognition of events and actions. Additionally, we developed a cyclic semantic network modeling assembled objects of our construction scenario. On the one hand this simple model guarantees the representation of every possible assembly. On the other hand it allows a straight recognition of assembled objects. The processing is based on the results of our hybrid object recognition for single objects and on the examination of the topological arrangement of regions in a cluster [12].

Model-Based 3D-Reconstruction As described earlier, the system is to interact in a 3D world with a human instructor. Therefore, information about 3D poses of objects is useful in order to supply the robotics part of the communicator with metric 3D information e.g. for grasping objects and also to interpret spatial relations of an utterance. This 3D-reconstruction is facilitated by simple geometric models for the object of our domain using points, circles, and line segments. The projection of these model features into the image plane is explicitly modeled using a pin-hole camera model. The resulting image features are constituted by contour groups and correspondence between model and image features is determined using previously found matchings between object regions and contour groups as well as knowledge supplied by the object models. Given these correspondences, we define a multivariate cost function measuring the distance of projected model features to detected image features. This cost function in turn is minimized using a monitored Levenberg-Marquardt-Method [13]. If more than one view of the scene is available, as for stereo images, this information can be exploited to enhance the accuracy of reconstruction results, given the correspondence of the objects in the different views.

Memory and Scene Representation The calculation of a 2D- or 3D-representation of the actual contents of a scene provides enough information for a dialog about the scene as long

as the scene is relatively static. In real construction scenes however, it is necessary to keep knowledge about temporary hidden objects (e.g. by a robot arm) or some history of objects to provide a stable scene representation over time to allow robust dialogs. Therefore, we integrated a universal memory module that links the results from the recognition modules over time using adaptive associative mechanisms [14]. The memory module is able to recognize unchanged objects (that need not necessarily be processed by the following modules) as well as changes and events that should be propagated in the system resulting in a more efficient processing of image sequences. Furthermore, the module provides multiple communication links to act as a buffer for simultaneous top-down and bottom-up evaluation.

4 Results

In this section we discuss various aspects of the system's performance and show exemplary results for a typical scene of our domain. Fig. 2

As the results of the polynomial classifier are pre-computed the color-based region segmentation works extremely efficient. At full image resolution (502×566 pixels) it takes on average 800 ms to do the pixel-wise classification, apply the smoothing, and calculate regions with associated form parameters[1]. Reduction of image resolution results in an almost linear speedup. When sub-sampling by a factor of 3 only 95 ms are required to generate the region segmentation for a single image. Results for a typical scene from our domain are shown in Fig. 2(a).

Significant contour segments and co- and curvilinear groupings are shown in Fig. 2(b). In conjunction with the detected closed contours, not being displayed, they describe the relevant structures in the image. Additionally, some noisy contours are detected mainly for elliptical structures. In most cases the grouping process is also able to overcome fragmentation of contour segments due to occlusion or low contrast, which do not occur in the example of Fig. 2. Closed contours are detected for the outer contour of most objects including occlusion, except for cubes and rings, which is caused by overlapping elliptical arcs or missing proximities. Summarizing, the significant grouping hypotheses contain the salient parts of the projected objects in the images. Computation times for this example are 17.6 sec for initial segmentation and 3.5 sec

[1] All experiments were carried out on a DIGITAL AlphaStation 500/400 (SPECint95 12.3, SPECfp95 14.1).

for generation and judgment of the grouping hypotheses.

The runtimes of contour grouping and region segmentation vary significantly. Therefore, matching of contours and regions is only applied to regions, which have been detected with a certain stability during the computations of the grouping process, since otherwise the results would already be outdated. Exploiting these matches for additional proximities to generate closed contours yields some important additional image structures. The added computational needs for matchings and generation of hypotheses using these matches is only about 0.5 sec, thus negligible compared to the overall runtime.

Hybrid object recognition was evaluated on 50 color images containing 12 objects on the average. The scenes were arranged by five persons who had no knowledge about the processing strategy of the system. Although there were no overlaps on the working platform some occlusions occurred due to the viewpoint of the camera. With region segmentation carried out at full image resolution an object recognition rate of 93.5% is achieved (for an example see Fig.2(c)). In contrast to the region segmentation process the speed of the object recognition depends only on the number of objects in the image. On average 84 ms are needed to recognize a scene object.

Qualitative results of the 3D-reconstruction based on recognition and grouping results are shown in Fig. 2(d). The 3D poses are very well reconstructed, except for the tires, where substantial deviations occur due to mismatches of image to model features. A typical figure for the quantitative accuracy is a mean error of about 3.5 mm for the relative distance and 4.5 degrees for the relative orientation between any two objects in the scene. The accuracy is mainly determined by the localization accuracy of the image features and the number of features available.

5 Conclusion

We presented the architecture of a vision system for a "situated artificial communicator". It consists of several components realized as individual modules within a large distributed processing system. Currently the control strategy is mainly data driven but in the next future we will integrate the propagation of regions of interest determined by other modalities like

speech or robotics. Furthermore, the declarative knowledge in our semantic network about possible assemblies should be exploited to a greater extend to guide the recognition of occluded objects.

The results of evaluating the system's modules individually as well as in their combination demonstrate the effectiveness of our approach. However, the processing time needed to perform all steps from region segmentation to 3D-reconstruction sequentially would be prohibitive. Therefore, the system architecture is designed to facilitate the distibuted processing of individual modules on a cluster of general purpose work-stations. Additionally, lines of processing running at widely differing time scales can be recombined resulting in a good overall responsiveness of the complete vision system.

5 REFERENCES

[1] D. C. Van Essen and J. L. Gallant. Neural mechanisms of form and motion processing in primate visual systems. *Neuron*, 13(1):1–10, 1994.

[2] Gernot A. Fink, Nils Jungclaus, Helge Ritter, and Gerhard Sagerer. A Communication Framework for Heterogeneous Distributed Pattern Analysis. In *International Conference on Algorithms And Architectures for Parallel Processing*, pages 881–890, Brisbane, 1995.

[3] Aleš Leonardis. *Image Analysis Using Parametric Models.* PhD thesis, University of Ljubljana, 1993.

[4] A. Maßmann and S. Posch. Mask-oriented grouping operations in a contour-based approach. In *Proc. 2nd Asian Conference on Computer Vision*, volume 3, pages 58–61, Singapore, 1995. IEEE Computer Society Press.

[5] A. Maßmann, S. Posch, G. Sagerer, and D. Schlüter. Using markov random fields for contour-based grouping. In *Int. Conf. on Image Processing*, volume II, pages 207–210, Santa Barbara, CA, 1997. IEEE Computer Society Press.

[6] Daniel Schlüter and Stefan Posch. Combinig contour and region information for perceptual grouping. In P. Levi, R.-J. Ahlers, F. May, and M. Schanz, editors, *Mustererkennung 1998,*

20. DAGM-Symposium, Informatik Aktuell, pages 393–401. Springer-Verlag, Heidelberg, 1998.

[7] F. Kummert, H. Niemann, R. Prechtel, and G. Sagerer. Control and Explanation in a Signal Understanding Environment. *Signal Processing, special issue on 'Intelligent Systems for Signal and Image Understanding'*, 32:111–145, 1993.

[8] G. Sagerer and H. Niemann. *Semantic Networks for Understanding Scenes.* Advances in Computer Vision and Machine Intelligence. Plenum Publishing Corporation, New York, 1997.

[9] G. Heidemann and H. Ritter. A Neural 3-D Object Recognition Architecture Using Optimized Gabor Filters. In *Proceedings of the 13th International Conference on Pattern Recognition, Vienna*, volume IV, pages 70–74. IEEE Computer Society Press, 1996.

[10] F. Kummert. *Interpretation von Bild- und Sprachsignalen — Ein hybrider Ansatz.* Shaker Verlag, Aachen, 1998.

[11] F. Kummert, G. A. Fink, G. Sagerer, and E. Braun. Hybrid Object Recognition in Image Sequences. In *14th International Conference on Pattern Recognition*, volume 2, pages 1165–1170, Brisbane, Australia, 1998.

[12] C. Bauckhage, F. Kummert, and G. Sagerer. Modeling and Recognition of Assembled Objects. In *24th Annual Conference of the IEEE Industrial Electronics Society, IECON'98*, volume 4, pages 2051–2056, Aachen, 1998.

[13] G. Socher, T. Merz, and S. Posch. 3-d reconstruction and camera calibration from images with known objetcs. In *Proc. 6th British Machine Vision Converence*, pages 167–176, 1995.

[14] Nils Jungclaus, Markus von der Heyde, Helge Ritter, and Gerhard Sagerer. An architecture for distributed visual memory. *Zeitschrift für Naturforschung C: A Journal of Biosciences*, Vol. 53, Special Issue: Natural Organisms, Artificial Organisms, and Their Brains (7/8):550–559, 1998.

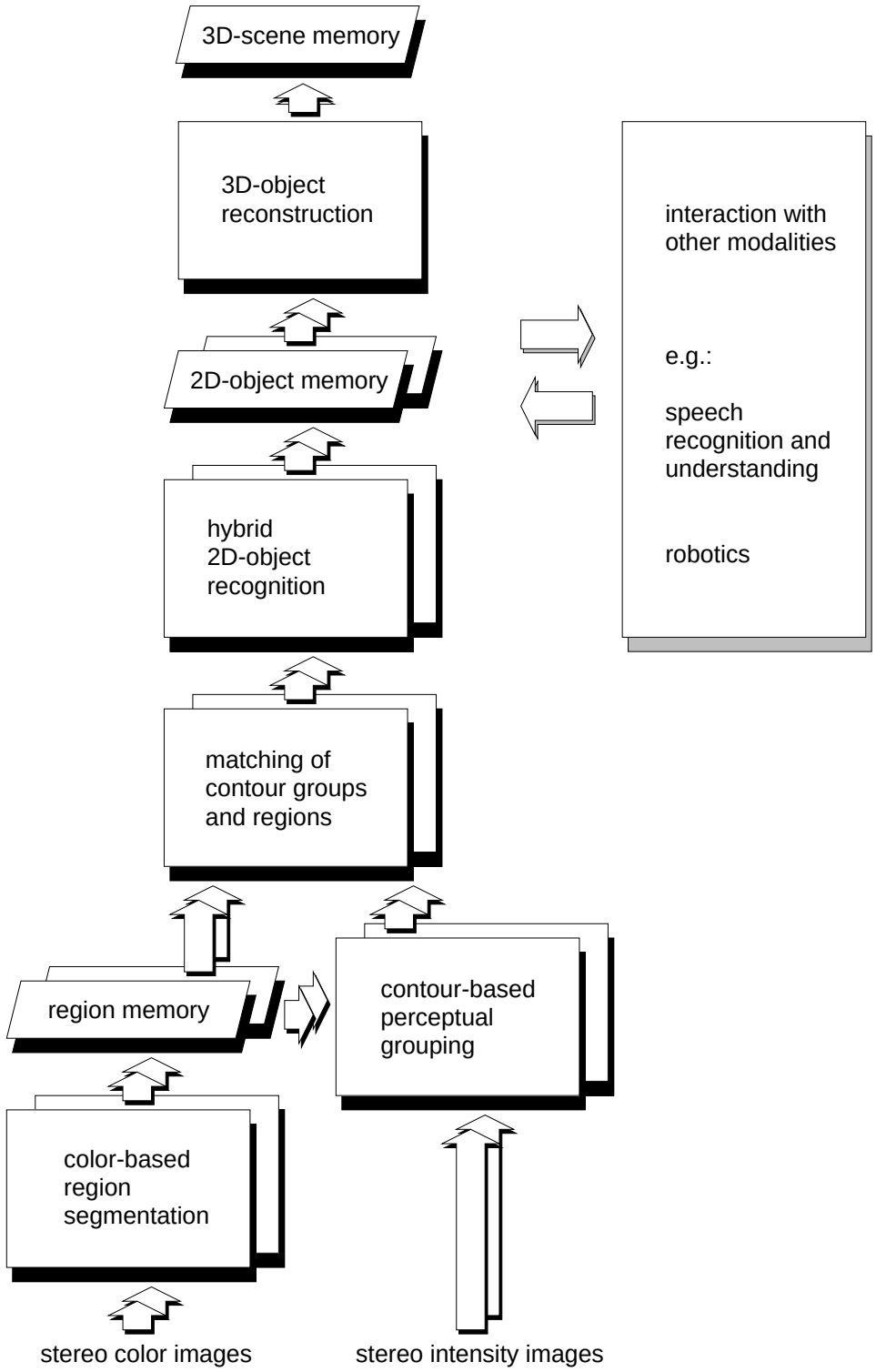
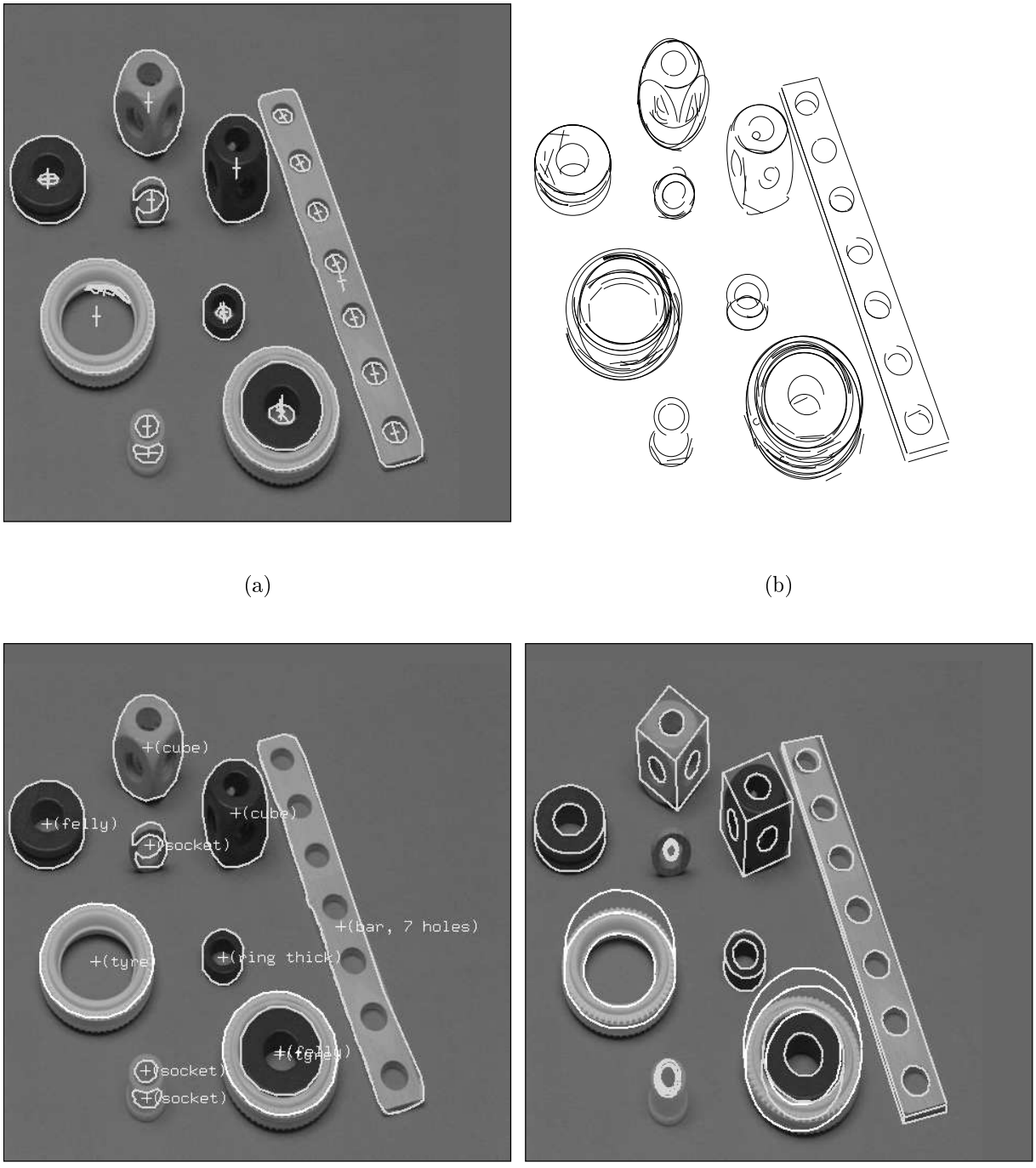


Fig. 1: (p. 2) Architecture of the vision system



(a) (b)

(c) (d)

Fig. 2: p. 7) Results for an example from the construction domain (a) Regions resulting from color segmentation; (b) Contour hypotheses; (c) 2D-object hypotheses generated using the hybrid approach; (d) 3D reconstruction