Multistage Recognition of Complex Objects with the Active Vision System NAVIS

N. Götze* B. Mertsching† S. Schmalz† S. Drüe*

Abstract

In this paper a biologically motivated active vision system to recognize complex objects within a non-uniform environment is presented. The system is based on simulating the behavior of striate complex cells by extracting oriented contour segments. Object recognition is done in a multistage fashion by first hypothesizing the presence and location of an object and afterwards identifying the object by its parts. A computer controlled pan-tilt unit was used as the experimental platform for evaluating proposed concepts. We show first results gained from a toy world environment.

1 Introduction

Vision research has undergone major changes in the last years. After Marr's influential work [Marr 1982], the scope of vision has been broadened to include non-visual (e.g. proprioceptive) information in the process of exploring and describing the environment. Further, (active) vision is regarded no longer as only passively evaluating images, but instead as actively influencing the way of image acquisition by controlling camera parameters such as zoom or shutter and to use active autonomous processes for tracking visual events that are continuous over time or space [Brooks 1992].
In the field of active object recognition only a few systems already show the complex behaviour typical for natural neural systems. Giefing [Giefing 1992] proposed an active vision system using saccadic camera gaze shifts for explorative scene analysis. An object is represented by a set of foveal images of several object views. Recognition is performed by the correlation of the stored foveal views with the actual foveal image. Rao [Rao 1995] introduced a general active vision architecture based on efficiently computable iconic representations, i.e. high-dimensional feature vectors obtained from an ensemble of Gaussian filters at serveral orientations and scales. Two main visual routines are distinguished: object location by matching a localised set of model features with image features at all possible retinal locations and object identification by comparing a foveal set of image features with all possible model features. Complex visual behaviour is viewed as task-specific sequential compositions of these simple "What" and "Where" strategies.
In this paper an approach for complex object recognition with the *N*eural *A*ctive *Vi*sion *S*ystem NAVIS [Drüe 1994] is presented. Here 2D recognition of previously learned complex objects is accomplished mainly in two stages: pre-attentional object features lead to a hypothesis of the object's presence and location which is successively validated by comparing the forms of several parts of the presented object with the stored ones of the hypothesis. The system uses the capabilities provided by a computer controlled pan-tilt unit and is able to deal with changing situations. Neccessary attention shifts are controlled by special gain and inhibition mechanisms.
After a motivation is given in section 2, the system is described in the 3rd chapter in detail. Experimental results will be discussed in section 4, followed by a conclusion with a comparative view on the systems mentioned above.

2 Physiological and Psychological Background

Research in cognitive psychology resulted in models for a multistage processing of sensory information [Treisman 1986]. Simple features in the field of vision are preattentively extracted in a highly parallel manner. These features are probably grouped according to Gestalt theory's laws [Rock 1990] which state that grouping could occur by construction of new, *emergent features* [Pomerantz 1989]. Hereafter, *selective attention* on a perceptual object or location (see [Yantis 1992] for a discussion) enables to serially differentiate between visual entities defined by a combination of features (e.g. [Koch 1985]), while there is an ongoing dispute about where this selection takes place [Yantis 1990]. Already examined locations are tagged in order to inhibit return of attention to the same place twice [Klein 1988].
Neurophysiological data reveals both functional and structural elements of the primate visual system. Functionally distinct pathways have been reported to operate in the dimensions of colour, form, motion and depth along areas V1 up to V5 [Livingstone 1988]. Higher cortical areas include the polysensoric posterior-parietal (PP) and the infero-temporal (IT) area, where experiments showed the former to process visuo-spatial information, while the latter seems

*Universität GH Paderborn, Pohlweg 47-49, 33098 Paderborn; e-mail: <*lastname*>@get.uni-paderborn.de
†Universität Hamburg, AG IMA, Vogt-Kölln-Str. 30, 22527 Hamburg; e-mail: <*lastname*>@informatik.uni-hamburg.de

to handle object-quality information [Ungerleider 1982, Andersen 1985]. IT-neurons are reported to be stimulated by complex objects such as iconic figures [Fujita 1992, Tanaka 1993] or even faces [Young 1992]. This suggests area IT being the memory of an alphabet of which complex objects are constituted [Miyashita 1993]. PP-neurons, on the other hand, seem to be involved also in attentional processes [Wurtz 1982], together with the superior colliculus and the pulvinar complex [Posner 1990]. Area PP probably also attenuates unattended and amplifies attended stimuli in IT and V4 which would explain that neurons in the latter respond to a particular object when several are presented simultaneously [Wise 1988]. The IT might contribute to the solution of visual search tasks by directing selective attention to specific objects according to what is looked for [Chelazzi 1993].

3 System Architecture

In this section the test platform as well as a system overview is presented.

3.1 Test Platform

A low-cost computer controlled pan-tilt unit and mounted CCD-cameras were used as the experimental basis. In figure 1 a schematic survey is given. The cameras are equipped with zoom, focus and aperture which can be adjusted either automatically or manually through serial I/O connections. In our work described here only one camera and no vergence have been used. In a further step disparity analysis and vergence control which have been developed separately [Trapp 1995] will be integrated.

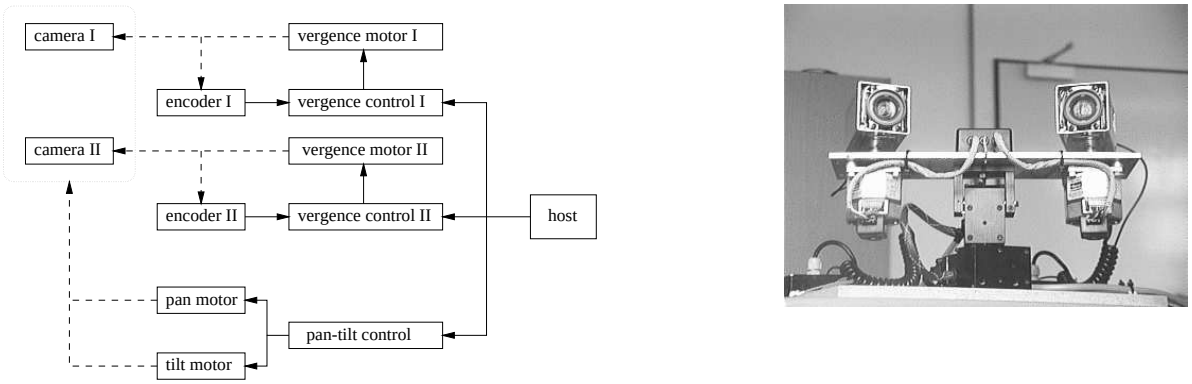

Figure 1: Structure of platform control (left); the pan-tilt unit (right)

3.2 System Survey

The functional and structural elements of the primate vision motivated our approach which is shown in figure 2 in overview. The modules are mainly executed counterclockwise. The recognition of one object takes several cycles. A description of the specific units follows.

3.2.1 Camera Unit

The camera unit is used for interacting with the cameras and the pan-tilt device through serial I/O connections. It has to guarantee a desired spatial orientation of the device. An object is foveated by rotating the camera along the pan and tilt axes where the relative angles are given by the recognition unit. The new desired pan respectively tilt angles are obtained through a simple inverse kinematic solution:

$$\begin{aligned} angle_{pan} &= \arctan\left(\frac{dist_h}{dist_{ap\,h}}\right) \\ angle_{tilt} &= \arctan\left(\frac{dist_v}{dist_{ap\,v}}\right) \end{aligned} \tag{1}$$

$dist_h$	horizontal distance from desired position to imagecenter
$dist_v$	vertical distance from desired position to imagecenter
$dist_{ap\,h}$	distance from focal point to CCD plane (related to the horizontal resolution)
$dist_{ap\,v}$	distance related to the vertical resolution of the CCD element

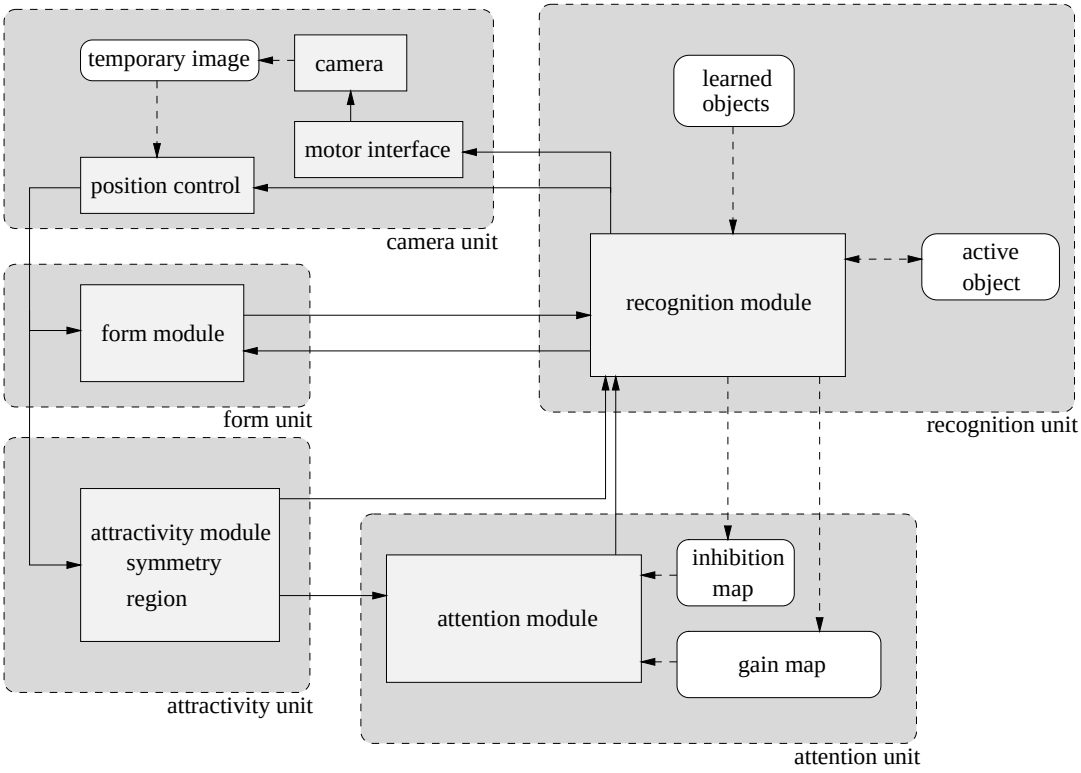

Figure 2: System overview

The necessary conversion from angles to motor steps is done by simply dividing by the angle resolution per step (actual 0.05 degrees). An error estimation gives the maximum position error in a worst-case situation and determines the size of the correlation window for finetuning. Exact positioning is achieved by filtering this window with a Sobel kernel before and after the movement and a correlation of the filtered images together with an appropriate shift of the image. We decided to correct the position error in order to be independent from the accuracy of the pan-tilt unit.

3.2.2 Form Unit

This unit models parts of low-level human visual perception [Hubel 1988]. Simulated response of retinal on-/ off-center ganglion cells is achieved by convolving the grey level input image with a mexican hat shaped kernel. Hereafter, *contour segments* are detected in different orientations in steps of 15 degrees (cf. simple cells; left in figure 3). Two further convolutions with oriented edge connecting kernels suppress noise and enlarge the receptive field sizes which is similar to the behavior of complex cells in primate V1. Object parts, or *forms*, are represented by these contours as a summation over all orientations.

Figure 3: Learned form (left); presented form for matching (right). All orientations are overlayed.

Forms refer to specific parts of an object. The location and size of the forms is specified by the superviser during learning stage. During recognition stage the system tries to match the stored forms with the presented image at the expected locations on the basis of a hypothesis generated by the recognition unit. Recognition uses the binarised activity of the complex neurons (right in figure 3). Through the smoothing and broadening of the edges a certain amount of invariances according to scale, rotation, and translation is achieved.

3.2.3 Attractivity Unit

Here the basis points to form hypothesis are generated from the input image. By measuring the inherent *attractivity* of contours and regions in terms of symmetry and homogeneity discrete *attractivity points*, or *fixation points*, can be extracted which are invariant with respect to small variations in viewpoint or lighting conditions and can be used for the gaze control of the system. At the learning stage these points with their spatial relations coarsely represent the object.

Determination of Symmetry

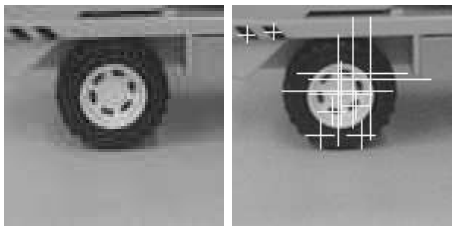

Figure 4: Input image (left); symmetry based attractivity points (right)

Detection of symmetry in relation to a specific image point uses the several oriented edge images and sums up the activity of tangentially oriented contour segments. By restricting the radius to a certain interval symmetrical structures with a determined size can be found. For details see [Drüe 1994]. In figure 4 a demonstration is depicted. White crosses mark the center point of a symmetrical structure and their size refers to the dimension of this structure. But the extracted attractivity points do not neccessarily correspond to meaningful structures in the input image.

Determination of Homogeneous Regions

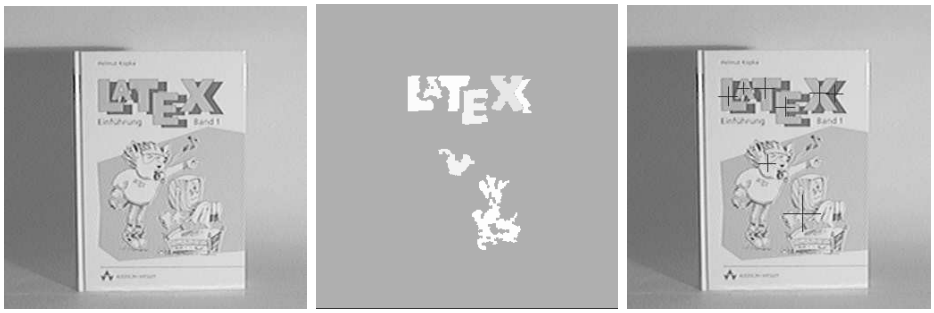

Figure 5: Input image (left); segmented regions (middle); region-based attractivity points (right)

Another strategy to determine the inherent attractivity of structures is segmentation based using homogeneous pixel intensities. Region growing is based on the maximum of the horizontally and vertically Sobel-filtered input image [Schlüter 1995]. If the value at a location is lower than a specified threshold the pixel at this location is merged with the currently considered region. In a second step adjacent regions having similar means and variances in grey values are merged. The value of attractivity is evaluated according to the variance within a region. Size information about each region can be drawn from the maximum of the points' distances from the region's center of gravity (the actual location of the attractivity point). In figure 5 the resulting attractivity points for an example are given. The points are marked by black crosses where their size corresponds again to the dimension of a region.

3.2.4 Attention Unit

The attention unit selects one attractivity point as the *point of attention* p_{att} which is the point with maximum priority among all candidate points p_{cand} calculated from all fixation points p_{fix}. This point serves as the reference point to form a hypothesis. Furthermore the attention point is used as the next spatial position the pan-tilt unit moves to. The attention unit contains two maps: The gain map $\mathcal{GM}$ emphasizes attractivity points that belong to the hypothesized object during validation, while the inhibition map $\mathcal{IM}$ is used to guarantee that an already examined attractivity point is not selected again for a time interval by decaying the priority values:

$$v(p_{cand}) = v(p_{fix}) \cdot \prod_{p_{IM} \in \mathcal{IM}} (1 - v(p_{IM}) \cdot d_1(p_{fix}, p_{IM})) \cdot \prod_{p_{GM} \in \mathcal{GM}} (1 + v(p_{GM}) \cdot d_2(p_{fix}, p_{GM})) \qquad (2)$$

$v(x)$ denotes the priority of a point x in $[0,1]$
$d(x,y)$ denotes a distance measure for x and y in $[0,1]$ (1 for identical locations)

3.2.5 Recognition Unit

The recognition unit is responsible for the generation of an object hypothesis and its validation by matching with stored forms. Furthermore it manages the gain and inhibition map.

Hypothesis Generation

A hypothesis is generated based on extracted attractivity points. The point of attention from the attention unit is taken as a reference point for the comparison with the stored attractivity configuration (see figure 6).
Let $\mathcal{LAP}$ and $\mathcal{PAP}$ denote the sets of learned and presented attractivity points and let $p_m \in \mathcal{LAP}$ be the learned attractivity point that matches the point of attention. The match value between presented attractivity points and those learned for one object is calculated in equations 3 and 4 (d and v are functions according to equation 2).

The fraction n/m in the interval $[0,1]$ is a measure of how good the set of learned attractivity points for one object matches with those presented. The highest value of n/m for all possible $p_m \in \mathcal{LAT}$ denotes the supposed position of a learned object. The highest fraction between all learned objects finally determines the object which will be taken as the next hypothesis.

$$n = \sum_{p_{LAP} \in \mathcal{LAP} \setminus \{p_m\}} v(p_{LAP}) \cdot v(p_1) \cdot d(p_{LAP}, p_1) \quad (3)$$
$$\text{with } p_1 \in \mathcal{PAP} \text{ such that}$$
$$d(p_{LAP}, p_1) \geq d(p_{LAP}, p_2)\ \forall p_2 \in \mathcal{PAP}$$

$$m = \sum_{p_{LAP} \in \mathcal{LAP} \setminus \{p_m\}} v(p_{LAP}) \cdot f \quad (4)$$

$$f = \begin{cases} v(p_3) & p_3 \in \mathcal{PAP} \text{ such that } 0 < \\ & d(p_{LAP}, p_3) \geq d(p_{LAP}, p_4)\ \forall p_4 \in \mathcal{PAP} \\ v(p_{LAP}) & \text{else} \end{cases}$$

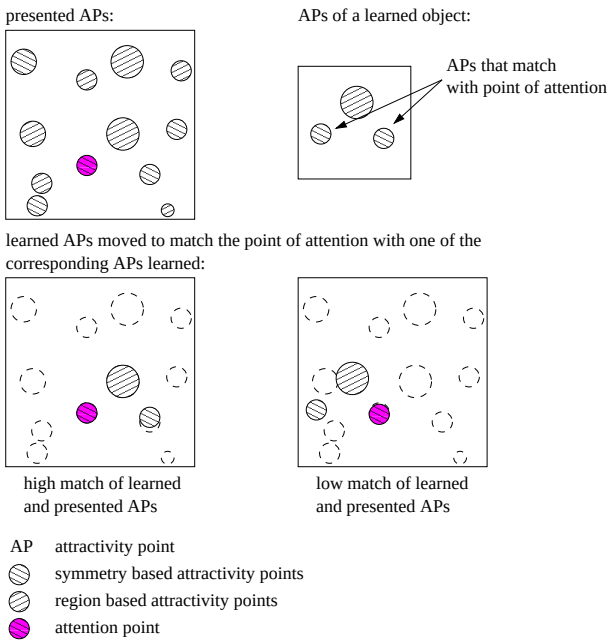

Figure 6: Matching of attractivity points during hypothesis generation.

Hypothesis Validation

During the validation of a hypothesis the unit tries to match the stored forms with the presented image. The quality of the match between the learned and the presented form is determined by dividing the numbers of matching elements by the sum of all elements learned:

$$match = \frac{\sum_{image} presented \wedge learned}{\sum_{image} learned} \quad (5)$$

The successfully recognized forms are maintained in an active object storage and the relative position of the next form is given to the camera unit for consecutive foveation. The learned attractivity points of the hypothesized object are put into the gain map to ensure that the same object is paid attention to in the next cycle. On the other hand examined points are marked in the inhibition map to avoid further influence. The whole recognition process is shown in overview in figure 7.

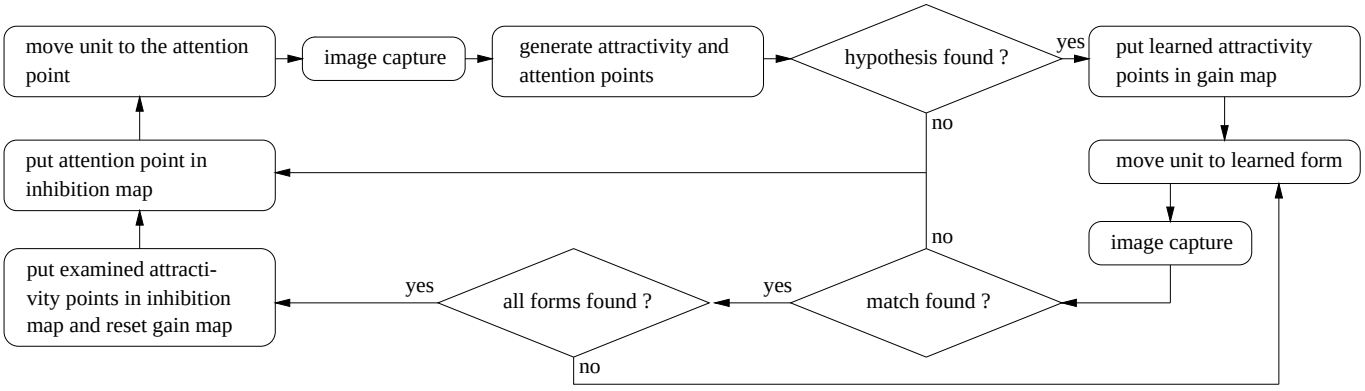

Figure 7: Main steps during recognition.

4 Experimental Results

A toy world problem is now presented to demonstrate the performance of the described approach.

4.1 Learned Objects

The system has already been trained with six objects in advance. Four toy vehicles were chosen, furthermore two other objects were trained whose silhouettes contrast with those of the vehicles. In figure 8 the objects are shown together with the attractivity points chosen autonomously by the system for later hypothesis generation (black and white crosses) and the forms selected by the superviser for later matching (black rectangles).

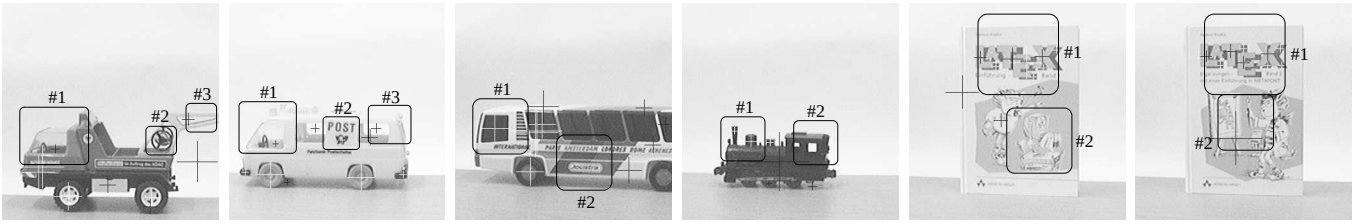

Figure 8: Learned objects (#1 until #6 from left to right) with marked attractivity points and forms.

4.2 Experiments

The objects were presented in front of a uniform white background as well as in front of a intense structured background (newspaper). The system searches actively for the learned objects in its environment (see figure 9). Several other experiments were performed which cannot be shown here due to space limitations. In table 1 an experiment summary regarding the number of steps until recognition, the number of rejected hypothesises and the match values is given. As expected the recognition tends to need more steps in a non-uniform environment. On the other hand the resulting match values are relatively independent of background structures. At present the system needs approximately 58 seconds to generate a valid hypothesis for one of the objects in figure 9 on a SPARCstation 20 with two 60 MHz processors. The time needed for matching depends on the size of the form which was determined by the superviser during learning stage. It takes 2:17 minutes for a 200×200 pixel wide form.

Table 1: Experiment summary.

#	description	steps til recogn.	rejected hypoth.	match values
E1	Object #1 in front of a white background.	5	0	0,865
E2	Same object as in E1 in front of a newspaper.	6	0	0,947
E3	Object #5 in front of a white background.	3	0	0,855
E4	Same object as in E3 in front of a newspaper.	14	0	0,840
E5	Objects #1 and #5 in front of a white background.	10	1	0,941 resp. 0,910
E6	Same object as in E5 in front of a newspaper (see figure 9).	9	0	0,697 resp. 0,878

5 Conclusions

Our experiments show that the system is capable of recognizing previously learned complex objects within a non-uniform environment. The handling of object occlusion is not demonstrated here but the approach described enables a straight-forward solution: An object is accepted if a certain number of the object's forms can be found.
Some limitations are encountered, too. First, fixation is restricted to objects with inherent symmetrical structures or homogeneous regions, because no hypothesis could be generated otherwise. An extension can be done by incorporating further features for fixation as e.g. done in the *key-point* approach [Heitger 1993, Henricsson 1994]. Second, recognition is neither invariant with respect to distances nor to changes in perspective, mainly because a set of attractivity points with their fixed relative positions is tried to be compared. An estimation of the objects distance through a disparity analysis (as already proposed in section 3.1) should allow to scale the attractivity point sets accordingly.
In comparision to other active recognition systems (e.g. [Giefing 1992, Rao 1995]) one can determine several things in common: object location/recognition dichotomy; use of motorized units to explore actively; managment of the

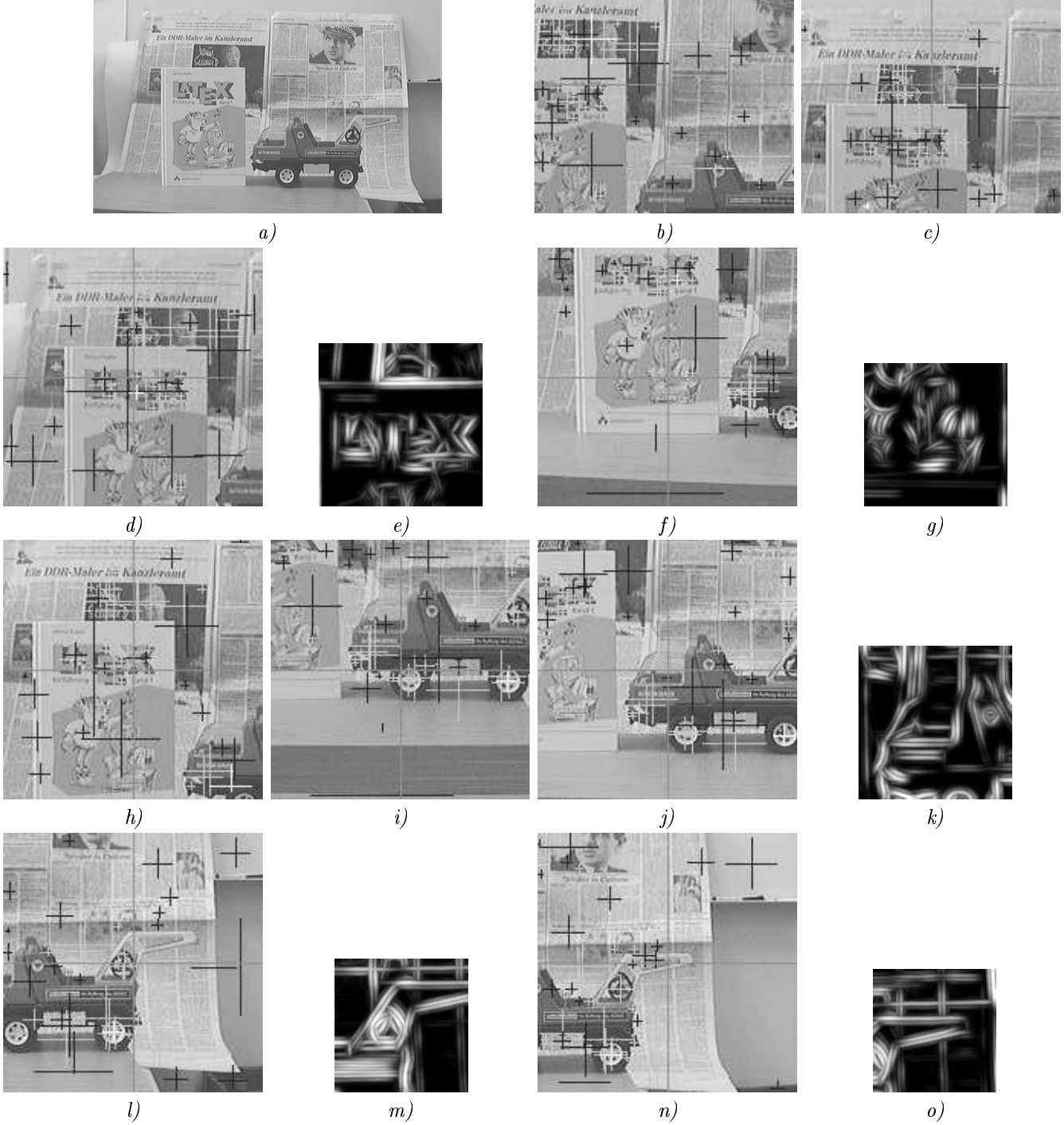

Figure 9: Performed Experiment E6:
a) Test environment to be explored by the system. Object #1 and #5 are presented in front of a newspaper within the laboratory. *b)* First evaluation of the environment by the active vision system. Black crosses represent region based attractivity points while grey crosses represent symmetry based ones. The attention point is marked white. Furthermore the reticle shows the image center. No hypothesis can be made for this scene. *c)* After the pan-tilt unit moved to the attention point, the new visible scene is explored. Now the hypothesis for object #5 can be computed (certainty of 78 %). *d)* The pan-tilt unit is oriented to the first form of object #5 (compare figure 8). *e)* The contour data can be matched with the stored form (97,6 %). *f)* Now the unit is oriented to the second form of object #5. *g)* Object #5 can be accepted through matching of the second form (94,8 %). *h)* The unit searches for new objects after having stored all examined points in the inhibition map. So a new attention point outside object #5 is found. *i)* The system can establish a new hypothesis for object #1 after doing a new fixation. *j)* The pan-tilt unit is now oriented to the first form of object #1. *k)* The first form can be matched (91,8 %). The unit moves to the second form. *l)* The second form of object #1 is fixated. *m)* A match of 88,8 % is calculated. *n)* Form #3 is searched through pan respectively tilt movements. *o)* The whole object #1 can be accepted after having produced a match of 82,9 % for the third form.

environment in special interest, or gain/inhibition maps. Differences exists concerning some implementational aspects. [Rao 1995] always work on the entire image. The resulting intense amount of computations has to be compensated by using specialized hardware (MaxVideo boards). Furthermore objects are recognized in a holistic way. Thereby problems according to occlusion and very large objects arise. But Rao already uses stereo vision for object-background separation. [Giefing 1992] includes the position error as a negative component during object matching in opposition to an error correction done here. In general we not only refer to attractivity points as points with a certain grey value feature but also as points representing eminent global structural properties.
The NAVIS system has been designed to be used as part of an autonomic, mobile, and multisensoric robot. To improve its performance future work will concentrate on the incorporation of strong invariances in order to enable the system to learn new objects completely unsupervised and to recognize objetcs from different perpectives. Furthermore specialized hardware has to be used to adapt the system to realtime conditions.

References

[Andersen 1985] Andersen, R.; Essick, G.; Siegel, R.: Encoding of spatial location by posterior parietal neurons. In: *Science*. 1985, 230, S. 456–458

[Brooks 1992] Brooks, R. A.: Foreword. In: Blake, A.; Yuille, A. (ed.): *Active Vision*. MIT. 1992

[Chelazzi 1993] Chelazzi, L.; Miller, E.; Duncan, J.; Desimone, R.: A neural basis for visual search in inferior temporal cortex. In: *Nature*. 1993, 363, S. 345–347

[Drüe 1994] Drüe, S.; Hoischen, R.; Trapp, R.: Tolerante Objekterkennung durch das Neuronale Active-Vision-System NAVIS. In: Bischof, H.; Kropatsch, W. G. (ed.): *Mustererkennung 1994*. Informatik Xpress 5. 1994, S. 251–264

[Fujita 1992] Fujita, I.; Tanaka, K.; Ito, M.; Cheng, K.: Columns for visual features of objects in monkey inferotemporal cortex. In: *Nature*. 1992, 360, S. 343–346

[Giefing 1992] Giefing, G.-J.; Janßen, H.; Mallot, H.: Saccadic object recognition with an active vision system. In: Neumann, B. (ed.): *Proceedings of the ECAI 92*. Wiley and Sons. 1992, S. 803–805

[Heitger 1993] Heitger, F.; von der Heydt, R.: A computational model of neural contour processing: Figure-ground segregation and illusory contours. In: *Proceedings of the Fourth International Conference on Computer Vision*. 1993, S. 32–40

[Henricsson 1994] Henricsson, O.; Heitger, F.: The role of key-points in finding contours. In: Eklundh, J. (ed.): *Computer Vision-ECCV '94*. Springer. 1994, S. 371–382

[Hubel 1988] Hubel, D.: Eye, brain and vision. Scientific American Library. 1988

[Klein 1988] Klein, R.: Inhibitory tagging system facilitates visual search. In: *Nature*. 1988, 334, S. 430–431

[Koch 1985] Koch, C.; Ullman, S.: Shifts in selective visual attention: Towards the underlying neural circuitry. In: *Human Neurobiology*. 1985, 4, S. 219–227

[Livingstone 1988] Livingstone, M.; Hubel, D.: Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. In: *Science*. 1988, 240, S. 740–749

[Marr 1982] Marr, D.: Vision. W.H. Freeman and Company. 1982

[Miyashita 1993] Miyashita, Y.: Inferior temporal cortex: Where visual perception meets memory. In: *Annual Review Neuroscience*. 1993, 16, S. 254–263

[Pomerantz 1989] Pomerantz, J.; Pristach, E.: Emergent features, attention, and perceptual glue in visual form perception. In: *Journal of Experimental Psychology: Human Perception and Performance*. 1989, 15, S. 635–649

[Posner 1990] Posner, M.; Petersen, S.: The attention system of the human brain. In: *Annual Review Neuroscience*. 1990, 13, S. 25–42

[Rao 1995] Rao, R. P. N.; Ballard, D. H.: An active vision architecture based on iconic representations. In: *Artificial Intelligence*. 1995, 78, S. 461–505

[Rock 1990] Rock, I.; Palmer, S.: The legacy of gestalt psychology. In: *Scientific American*. 1990, (263), S. 48–61

[Schlüter 1995] Schlüter, N.: Entwicklung und Simulation einer flächenbasierten Fovealisierungstrategie zur Mustererkennung in Grauwertbildern. Studienarbeit, Universität-Gesamthochschule Paderborn. 1995

[Tanaka 1993] Tanaka, K.: Neuronal mechanisms of object recognition. In: *Science*. 1993, 262, S. 685–688

[Trapp 1995] Trapp, R.; Drüe, S.; Mertsching, B.: Korrespondenz in der Stereoskopie bei räumlich verteilten Merkmalsrepräsentationen im Neuronalen-Active-Vision-System NAVIS. In: Sagerer, G.; Posch, S.; Kummert, F. (ed.): *Mustererkennung 1995*. Springer-Verlag, Berlin, Heidelberg. 1995, S. 494–499

[Treisman 1986] Treisman, A.: Features and objects in visual processing. In: *Scientific American*. 1986, (255), S. 114–125

[Ungerleider 1982] Ungerleider, L.; Mishkin, M.: Object vision and spatial vision: Two cortical pathways. In: *Trends in Neuroscience*. 1982, 6, S. 414–417

[Wise 1988] Wise, S.; Robert, D.: Behavioural neurophysiology: Insights into seeing and grasping. In: *Science*. 1988, 242, S. 736–741

[Wurtz 1982] Wurtz, R.; Goldberg, M.; Robinson, D.: Brain mechanisms of visual attention. In: *Scientific American*. 1982, (246), S. 100–107

[Yantis 1992] Yantis, S.: Multielement visual tracking: Attention and perceptual organization. In: *Cognitive Psychology*. 1992, (24), S. 295–340

[Yantis 1990] Yantis, S.; Johnston, J.: On the locus of visual selection: Evidence from focused attention tasks. In: *Journal of Experimental Psychology: Human Perception and Performance*. 1990, 16, S. 135–149

[Young 1992] Young, M.; Yamane, S.: Sparse population coding of faces in the inferotemporal cortex. In: *Science*. 1992, 256, S. 1327–1330