Multi-Pass Feedback Control for Object Recognition

M. Mirmehdi, P. L. Palmer, J. Kittler & H. Dabis*
Vision, Speech & Signal Processing Group
University of Surrey
Guildford GU2 5XH, UK
e-mail: M.Mirmehdi@ee.surrey.ac.uk

22/06/96

Abstract

In order to locate interesting areas of an image we describe a system for focus of attention; this is based on feedback strategies combining low-level features, and a high-level object model to recognise the object and to direct the search for missing information. We aim to improve on established single-pass hypothesis generation and verification approaches by applying our complex feedback strategies to recognise generic classes of objects. By using a complex feedback strategy we produce optimal sets of low level features and reduce the number of hypotheses generated. The system can extract simple and complex objects in a scale and rotation independent manner where the objects may be partially occluded. The method is illustrated for simple cubic objects and the results are expected to be applied for a mobile robot application.

Afin de situer les régions d'intérêt d'une image, nous décrivons un système de centre d'attention basé sur des stratégies de retour de données. Celles-ci comprennent plusieurs caractéristiques de bas-niveau ainsi q'un modèle de haut-niveau lequel reconnaît l'objet et dirige l'opération de recherche des données manquantes. Notre objectif est d'améliorer l'approche de génération et de vérification des hypothèses en utilisant nos stratégies de retour de données afin d'identifier des catégories génériques d'objets. Grâce à ces stratégies nous obtiendrons un ensemble optimal de caractéristiques de bas-niveau et nous réduirons le nombre d'hypothèses générées. Le système est capable de reconnaître des objets simples ou complexes indépendamment de l'échelle, de la rotation ou bien même de l'occlusion partielle d'un objet. Nous démontrerons cette méthode pour de simples objets cubiques. Celle-là sera utilisée pour une application de robot mobile.

*Now at Emrad Ltd., Catteshall Lane, Godalming, Surrey GU7 1NG, UK

1 Introduction

The area of object recognition is of principle importance in computer vision and has been approached through diverse avenues by various researchers [1, 2, 3, 4]. As a first step towards general purpose, human-like perception of real-world objects, we may develop unstrict geometrical constraints and representations that are loose enough to allow some sort of generic object representation.

Although low-level image processing techniques, such as edge detection or line and curve finders, perform well under certain types of noise, the problem of identifying what is good data and what is bad data still remains. The main weakness lies in the fact that the lower levels of processing have no knowledge of the higher level requirements. A resulting feature under-detection or the extraction of spurious structures can seriously affect the success of both the hypothesis generation and verification processes. This suggests that the widely used feedback implicit in the hypothesise-verify approach to image interpretation is rather crude and should be enhanced by more sophisticated strategies.

In this paper we extend the idea of the single-pass feedback framework by employing complex feedback strategies for both more robust hypothesis generation and hypothesis verification. Initially, we introduce control mechanisms for low-level parameter optimisation. Similar purposes are worked upon using scale space theory by Lindeberg [5]. We follow by reviewing control mechanisms for all the general levels of the processing chain from low level feature extraction through high level recognition. Our purpose is to exploit as much information as is available within the visual data. On an opposite scale Draper and Hanson [6] use control strategies to implement a

choice of recognition algorithms to achieve a visual goal.

Furthermore in this paper, we investigate the use of focus of attention on feature extraction; this is essential since in a simple feedback strategy the possibility of failure due to poor quality data is high for an object from a generic class with specific dimensional information. Focus of attention allows us to focus on the region or regions of the image where objects may be and reduce the time required for a more successful verification for multiple types of object and multiple instances of the same class of object.

To focus our attention upon a particular region we employ an interest operator to provide a quantitative measure based on perceptual grouping of low level object features. Perceptual grouping [7] argument that most man-made objects conform to the notion of non-accidentalness is of paramount importance in this work. Once a candidate object is identified, we match against a loose model and this will be used as a basis for searching for missing information, fusing partial hypotheses and deriving a final measure of interest for the outputs.

Although we design our system to be as modular and independent of the actual application, in this paper we shall consider the feedback strategies required for the location and recognition of geometrically simple, cubic or box-shaped objects. We include in this category objects such as books, video tapes and computers, as well as packaged goods which may have labels and other markings on the box which may potentially confuse the vision system. Although a wealth of work has already been attempted in this area [8, 1, 9, 2, 3, 4] it is still a major problem in computer vision. The difference in our work can be re-stated at this point to be the investigation of complex feedback strategies for the recognition of objects, be they simple or complex. In this particular instance, the purpose of investigating such simple objects is that we intend to use the results for an indoor mobile robot project which is expected to move within the boundaries of laboratories and corridors and perform object recognition as well as collision avoidance. The system is intended to be simple enough to allow it to respond to events in real-time. However, we have also applied our system to a completely different application, but based entirely on the same feedback and recognition principles, for the identification of bridges in IR images [10]. This shows the applicability of our system for the recognition of highly complex structures within very noisy and cluttered scenes.

In the following sections we first consider our vision system and then review the issue of feedback at various levels of processing. Finally we present our results.

2 The Control Structure

A flow-diagram of the vision system developed is shown in Figure 1; the system consists of three main sections: the low level, the intermediate levels, and the high level modules. These are all manipulated by our feedback control modules. The system is designed to be modular enough to allow us to plug in different techniques at each processing level as well as being flexible if we need to alter the corresponding feedback control modules for different tasks.

The design of the low level modules is very closely linked to the modeling and representation of the target objects. Our application example of boxes has a well defined target object. The complexity in this problem comes from the fact that there is a large amount of clutter, that some of the edges of the boxes have very low contrast so information is missing, and that many boxes have designs printed on them that leads to confusion and multiple interpretations. Finally, another major issue is the extra line segments arising due to shadows.

We perform edge detection [11], linking and a Hough transform algorithm [12] for locating straight line segments in the Low Level processing stages. Our representation of the target object will only include those straight line segments that are connected via junctions. The junction finder [13] is part of the Intermediate Level stages along with perceptual grouping of object features. The High Level processing stage consists of model matching and search for missing information. The feedback control modules generate and test hypotheses, and create search strategies for the list of possibilities where missing information may be found.

The feedback control system will need to have a bottom-up description of the target object based upon a set of object features derived from low level processing. We shall refer to this set of object features as the *signature* of the target object. Given that not all of the low level features may be accessible, e.g. due to partial occlusion or shadows, it is therefore necessary that the control structure also has a top-down model of the target object. This is viewed, in our paper, as a *stereotype* of the target object and is only a crude description of the target object, but it serves as a stepping stone for the search for missing data, and the decision on final termination; this being a principle function of the control mechanism.

Another aspect of the control structure, related to the above, will be to make quantitative measures of the performance of the system. This will be required at the low level stages where parameter settings on the various algorithms will need to be adjusted. It will also be needed to determine to which level of the processing chain feedback should be directed to and what form the feedback should take, and finally it will be required to determine confidence measures for the final output.

3 Feedback at Lower Levels of Processing

One of the first issues for the feedback process is the settings of the various parameters in the low level algorithms. We ran the low level modules at four different parameter settings ranging from very low to very high values; then we took as the optimal parameter setting the value at the turning point of a cubic fitted to that data. In total we optimised five parameters: the edge detection mask width, the lower and upper hysteresis thresholds for linking, and finally the kernel widths in ρ and θ for the Hough transform. This approach is similar to that of Peak Holding in Control System theory [14].

To determine if the parameters are well chosen we use a performance measure to assess the performance of the low level procedures. We can feedback for different parameter settings for optimisation if the initial results are not satisfactory. Details of the performance quality measure can be found in [15]; it only suffices here to stress that it is based upon the results coming out of the entire low level stage - edge detection, linking and Hough transform. This is a different concept to measuring performance of the individual procedures within the processing chain [16].

The quality measure is firstly applied to the whole image and then again, on the first iteration, to the initial set of regions of interest (see Figure 1). We have mechanisms to apply the quality measure on every iteration but found that in most cases, after the first iteration, the regions have reduced to a fairly small size and quality differences are insignificant.

4 Feedback at Intermediate Levels of Processing

The features obtained from the low level stage need to be grouped into structures which resemble the signature of the target object. Initially, our junction finder [13] groups together sets of lines to form second order (V2) junctions, third order (V3) junctions and occluded (T2) junctions. Next, we use perceptual grouping [7] to group lines and junctions which are likely to belong to the target object. This is designed to look for the signature of the target object among the object features derived from the Hough transform and the junction finder. The strength of perceptual grouping stems from the fact that most man-made objects obey certain symmetry rules. This allows for an important property in our bottom-up modeling - rotation and scale invariancy.

When developing rules for a box we emphasise the importance of parallelism, orthogonality and connectivity. From a general viewpoint three faces of a box are usually visible. Consequently, a characteristic signature would be a third order V3 junction where three lines intersect. We use this as the principle signature of a box. When a box is seen face on, or when much of the box is shadowed, then no V3 junction may be detected. So another, but less characteristic, signature of a box is a V2 junction where the two lines are nearly orthogonal to each other. A final important signature of a box is that the connected structure should contain three sets of 3 parallel lines of similar length. It is not necessary to find all of this signature to have a candidate box, but the more structure that is found to correspond with this signature, the more interesting the candidate becomes. We shall discuss how this is achieved in more detail in Section 4.1.

Our rules for grouping sets of features together as belonging to a single box can be expressed as follows: Locate a V3 junction and then expand from that to a connected node (junction). If any line which meets at this node is parallel to any lines in the grouping then we add this to the group. We continue this process until we exhaust the list of connected features and then proceed with the next V3 junction. Once we have considered all the V3 junctions, we then look at any remaining V2 junctions composed of two orthogonal lines to look at the faces on the box to see if any have high contrast. The list of these groups is then passed to the next stage of processing where interest is assigned to each grouping.

4.1 Feedback For Focus of Attention

The next strategy for the feedback modules is to focus attention on regions of the image most likely to contain the target object(s) and extract as much

of the structure of these objects as possible. For this we use the bottom-up model of the target object's signature to locate the region of the image where the target objects may be found. It is important to keep in mind that not all the signature may be available, but the more numerous the number of component features in the structure obtained from the perceptual grouping stage, the greater our interest in the structure as a hypothesis becomes. To reduce the number of hypotheses to be considered, the structures are assigned an interest value based upon an *interest operator.*

For each structure we form a region of interest. Some regions will cross-over heavily and they may have arisen due to the same box contributing to more than one structure. These regions are then merged and adjusted (see Figure 1). Another important aspect of the interest operator is that some structures from the low level processing stages will be false alarms, even though they may have high interest. As the feedback proceeds and attention is focused on regions of false alarms, the interest level will at best remain static and in most cases drop. In this way, feedback removes objects that show only vague similarity to the target object sought.

The interest operator [17] is designed to increase exponentially as more features are included within the structure. In this way, the many small structures that do not resemble at all closely the target object signature are all assigned comparatively small levels of interest and are not searched further. A similar approach to grouping features from among a large clutter has been proposed by Sha'ashua & Ullman [18] to produce saliency maps based upon rules of human perception.

4.2 Interest Operator for Boxes

The interest for a box is based upon the rules described above. We therefore assign interest for V3 junctions, V2 junctions formed by two orthogonal lines and sets of parallel lines of comparable length. We express the interest in the whole structure as:

$$I_{\text{box}} = I_{\text{V}} I_{\text{par}} \tag{1}$$

where I_{V} is the interest due to the existence of either a V3 junction within the connected region or a V2 junction which was formed by two perpendicular lines. Since the existence of such junctions is a binary operation, then there is no exponential decay term:

$$I_{\text{V}} = \prod_{m=1}^{N} K_{\text{V}} \tag{2}$$

where K_{V} is a constant. N is the number of such junctions found in the structure. We combine this interest with an interest associated with sets of parallel lines of comparable length, I_{par}, given by:

$$I_{\text{par}} = \prod_{m=1}^{M} K_L \mathrm{e}^{-\Delta\alpha_m} \tag{3}$$

where K_L is a constant. The exponential indicates the rate at which the interest decreases as the lines deviate from being parallel. $\Delta\alpha_m$ is the difference in the Hough angle between the two lines. Since we check with all lines in the structure for parallelism, then if a long line should appear segmented from the low level procedures, for whatever reason, then all parts of the segmented line will still be held in the structure. With feedback such problems will disappear as attention becomes more focused.

The exponential in Equation 3 decays to zero as the deviation from parallelism increases. On the other hand as more parallel lines are found which belong to a particular structure, then the interest increases multiplicatively. The constants K_V and K_L control the rate of rise of the interest as more features are detected. In effect, these values reflect the theory of non-accidentalness.

5 Feedback In Top-Down Analysis

The process of top-down object recognition starts by considering each region of high interest and consists of the confirmation of the type of box in the region and associated missing information analysis on a per structure basis to strengthen our hypothesis.

Initially we invoke a model-matching stage (see Figure 1) which uses the stereotype of the target model to analyse the structure and determine if a match can be made. The stereotype of the box is based on similar definitions to the signature. It states that a box-shaped object should consist of four V3 junctions and three parallel lines which in association will form faces of the box. This loose definition allows us to capture boxes independent of rotation and scale. The next stage is to search for missing information which is carried out by imposing regional windows upon the faces of the box and searching for more line segments. This stage is currently under further improvement but a similar approach which is fully implemented for our other application is reported in [10]. Any extra and valid line segments found after the missing information

search will add to the interest level and we presently use this as our confidence measure in the hypothesis.

6 Feedback Strategies

We now consider and summarise the strategies for dealing with the fusion of information and decisions on how the feedback should proceed. The general approach we use is a hypothesise/verify approach to identify as many features of the structure as possible. In the bottom-up phase of the control, following the early optimisation stages, we initially perform our low level image processing tasks on the whole image. To reduce the amount of data we focus our attention on regions which are most likely to contain the target object. This is based upon the interest level generated using our interest operator. The control system loops back to regions of interest where the interest level computed earlier is high. Most false alarms have insignificant interest and are quickly discarded. This looping back continues (see Figure 1) until the size of the region of interest does not shrink significantly, i.e we can not focus our attention any further, or the interest level has dropped or remained static. At this stage any remaining false alarms are usually discarded with one loop back, due to falling interest. During this loop back stage, the control module checks to see whether, with the focus of attention, the structure extends to the edge of the region of interest. In such cases, this side of the region of interest will expand for the next feedback loop. If more of the significant structure is found, it would lead to a significantly larger interest level, otherwise that side of the region will no longer grow and the interest would not be further affected by that structure.

We compare the level of interest generated in the region of interest with that found on the previous pass through for hypothesis validation. When the conditions stated above are satisfied the control module terminates the bottom-up procedure and all remaining regions will have high interest values.

In the top-down stage, for each box structure in each region of interest we perform model matching; this entails validation of the components of each structure by resorting to the stereotype of the target object.

7 Results and Conclusions

We present here the results of our system applied to some box shaped objects. Figure 2(a) shows a set of building blocks with the initial regions of interest superimposed on top. The system does not merge sub-windows for box hypotheses as box-shaped objects may overlap in the image. Also marked are the line segments that the system believed to be part of box-shaped objects. These are the results of the feedback on optimisation for the best low-level parameters. We note that since the interest operator is designed to look for V3 junctions, the hexagon is pulled out with the highest interest. Figure 2(b) displays the final effects of feedback after the focus of attention approach fed back to each of the initial regions. The process continued until there was no further improvement in a region. The block on top of the hexagon starts with a sub-window which is too small to contain the whole block. With the repetition of the feedback, this sub-window expands downwards because line segments are found which are associated with the structure of interest, but reach the bottom edge of the window. The control module recognises this and expands the region accordingly. Overall we see a significant improvement in the features extracted as we feed back.

To show the effect of the top-down approach we selected one region and generated windows for each face of the cube. Figure 2(c) shows the features extracted as we feed back to each of the face windows. We see that missing data which could not be identified using the focus of attention have now been extracted. Finally, Figure 2(d) shows the overall result after applying the polyhedral recognition technique using pre-defined models introduced in [4] to the data from Figure 2(c). However, we are currently investigating perceptual grouping techniques to derive the polyhedron by further analysis of the box junctions and without resort to a strict model; just the stereotype. Example Figure 3 shows the same sequence of events as the previous figure but for a more realistic scene where there is extra noise and clutter from the labels and the writings on the box packet.

In this paper we have introduced a complex feedback strategy for the generation and verification of hypotheses for object recognition. The feedback manipulates and controls the various processing levels and provides focus of attention for better extraction and recognition of objects. For one application, we intend to use our system in guiding a robot to manoeuver in indoor spaces and to avoid collisions. In another application, we have applied the system to the recognition of bridges in IR images [10]. In the future, we intend to further focus on aspects of incorporating feedback of knowledge from the higher level scene interpretation modules

to the low level image processing algorithms within the area of active vision.

Acknowledgments

The authors wish to acknowledge the support of the DRA, Farnborough, UK, for this work. Also thanks to Miss Ratna Rambaruth for the French abstract.

References

[1] R.T.Chin and C.R. Dyer. Model-based recognition in robot vision. *ACM Computing Survey*, 18:67–108, 1986.

[2] W.E.L. Grimson and D.P.Huttenlocher. On the verification of hypothesized matches in model-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13:1201–1121, 1991.

[3] I. Shimshoni and J. Ponce. Probabilistic 3d object recognition. In *International Conf. on Computer Vision*, pages 488–493, 1995.

[4] K.C. Wong. *Representation, Feature Extraction and Geometric Constraints for Recognising 3D Objects from a Single Perspective View.* PhD thesis, Surrey University, Guildford, UK, 1992.

[5] T. Lindeberg. Scale selection for differential operators. Technical Report ISRN KTH NA/P–94/05–SE, KTH (Royal Institute of Technology), Sweden, 1994.

[6] B.A. Draper and A.R. Hanson. An example of learning in knowledge-directed vision. In *Scandinavian Conference on Image Analysis*, pages 189–201, 1991.

[7] D. G. Lowe. *Perceptual Organisation and Visual Recognition.* Kluwer Academic Publishers, 1985.

[8] D.C. Baker, S.S. Hwang, and J.K. Aggarwal. Detection and segmentation of man-made objects in outdoor scenes: Concrete bridges. *Journal of the Optical Society of America*, JOSA-6:938–950, 1989.

[9] W.E.L. Grimson. The combinatorics of object recognition in cluttered environments using constrained search. *Artificial Intelligence*, 44:121–165, 1990.

[10] M. Mirmehdi, P.L. Palmer, J. Kittler, and H. Dabis. Complex feedback strategies for object recognition. *Submitted to IEEE Transcations in Image Processing*, 1996.

[11] M. Petrou and J. Kittler. Optimal edge detectors for ramp edges. *IEEE Trans. in Pattern Analysis and Machine Intelligence*, PAMI-13:483–491, 1991.

[12] P.L. Palmer, M. Petrou, and J. Kittler. A hough transform algorithm with a 2d hypothesis testing kernel. *CVGIP: Image Understanding*, 58:221–234, 1993.

[13] J.Matas and J. Kittler. Junction detection using probabilistic relaxation. *Image and Vision Computing*, 11:197–202, 1993.

[14] V.W. Eveleigh. *Adaptive Control and Optimisation Techniques.* McGraw-Hill, 1967.

[15] P.L. Palmer, H. Dabis, and J. Kittler. A performance measure for boundary detection algorithms. *accepted for CVGIP: Image Understanding*, 1996.

[16] W. K. Pratt. *Digital Image Processing.* Wiley and Sons, 1978.

[17] H. Dabis, P.L. Palmer, and J. Kittler. An interest operator based on perceptual grouping. In *Scandinavian Conference on Image Analysis*, pages 315–322, 1994.

[18] A. Sha'ashua and S. Ullman. Structural saliency: The detection of globally salient structures using a locally connected network. In *International Conf. on Computer Vision*, pages 321–327, 1988.

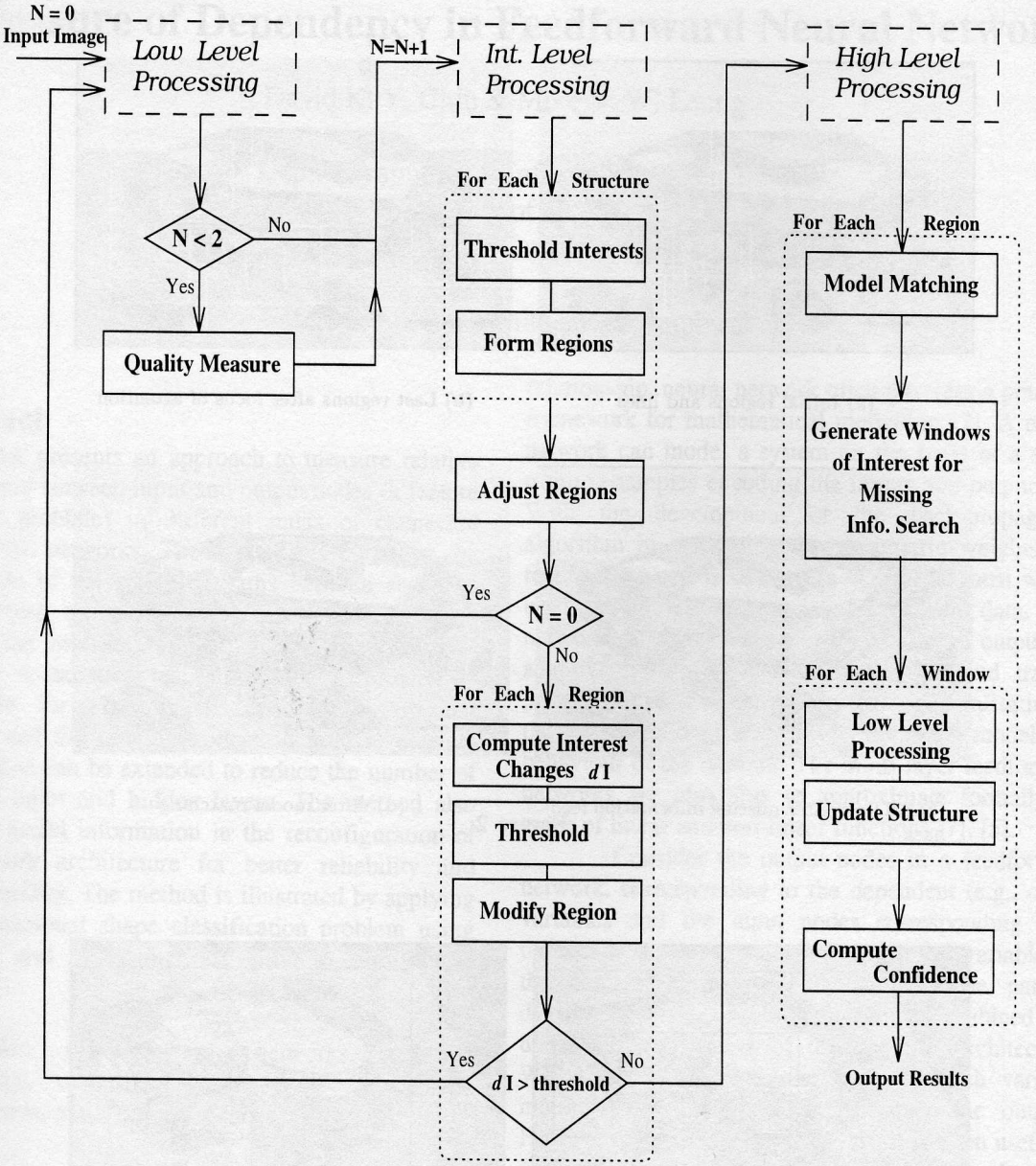
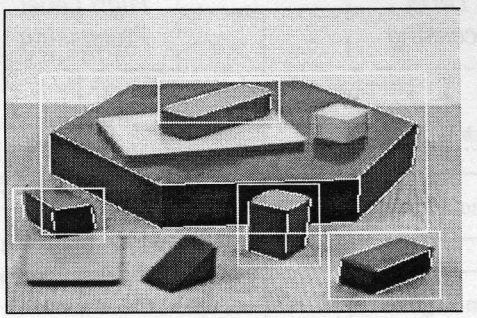
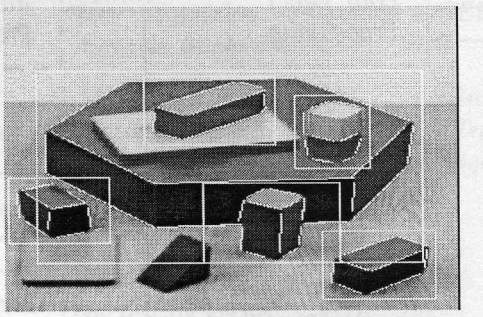


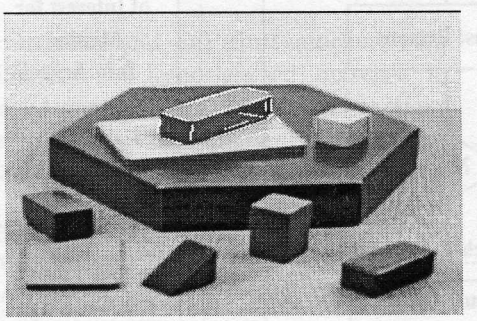
Figure 1: Overview of the vision system used, divided into three separate stages - low level processing; intermediate levels (object features) and feedback control.



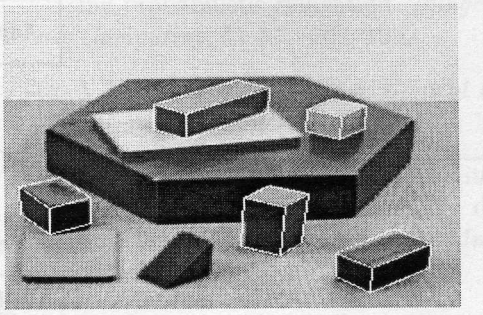
(a) Initial regions and lines



(b) Last regions after focus of attention

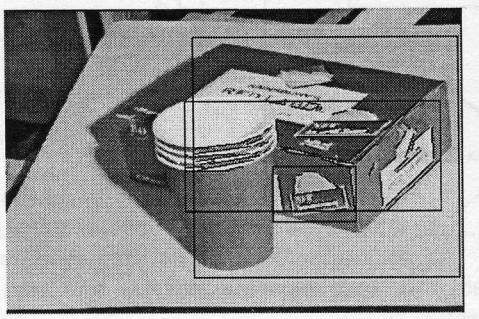


(c) All lines after missing information feedback

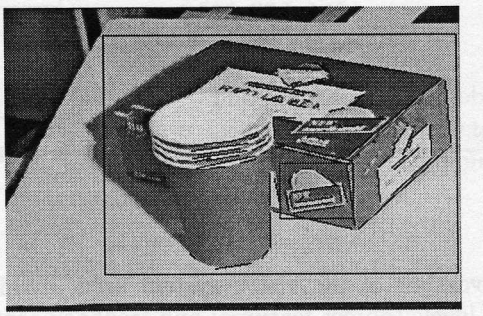


(d) After model matching

Figure 2:

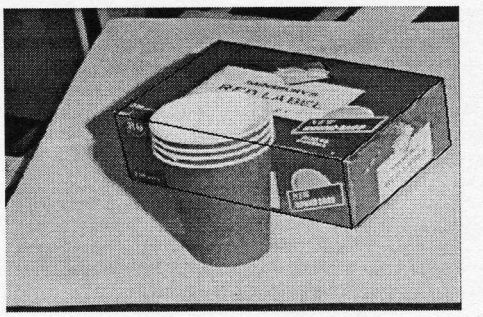


(a) Initial regions and lines



(b) Last regions after focus of attention

(c) All lines after missing information feedback



(d) After model matching

Figure 3: