Complex Feedback Strategies for Hypothesis Generation and Verification

M. Mirmehdi, P. L. Palmer, J. Kittler & H. Dabis
Department of Electronic & Electrical Engineering,
University of Surrey,
Guildford, Surrey GU2 5XH, U.K.
M.Mirmehdi@ee.surrey.ac.uk

Abstract

We extend the idea of the single-pass feedback framework by employing complex feedback strategies for both more robust hypothesis generation and hypothesis verification. These strategies are developed at every level of our object recognition application; from low-level parameter optimisation, through the low level processing chain, to higher level recognition stages. The strategies are independent of the techniques used at each level. We introduce various control mechanisms to achieve such complex feedback strategies. Within our implementation, we minimise the amount of feedback to false alarms by using an interest operator which directs the search through the hypotheses in an optimal manner. Furthermore, we obtain detailed information about a complex object and not just its location. Thus, following top-down recognition of the object our feedback control directs the search for missing information. We illustrate our approach using noisy infra-red images of bridges in real outdoor scenes and demonstrate the insensitivity of the system to noisy data and partial occlusion.

1 Introduction

The more constraints there are in the analysis domain of a vision system the more it can enjoy the luxury of real-time processing. Widening the domain of constraints, and wandering into the realm of outdoor scenes where there is noise, clutter, and occlusion amongst other image degradation issues, the problem becomes difficult manifold for a real-time vision system to handle. In this paper we describe feedback control strategies that can be employed within a real-time vision system towards the reduction of computational requirements by:

- enhancing the performance of lower levels of processing for feature extraction
- generating hypotheses for focusing attention on regions of interest
- generating hypotheses to seek out extra information on object of interest

The general *hypothesise-verify* paradigm can be viewed as a feedback process. The error signal between the image data and the hypothesised model and its pose

BMVC 1996 doi:10.5244/C.10.32

is used to accept or reject the hypothesis. If the current hypothesis is rejected, a new one is generated and the corresponding error signal computed. This illustrates a one-pass hypothesis generate, test, and accept/discard approach within a feedback framework. For example, Brooks [3] with the ACRONYM system or Grimson and Huttenlocher [7] use this approach at the same time as using a statistical occupancy method to achieve model recognition. Also, Lai and de Figueiredo [8] have implemented an iterative contextual feedback system based on optimal iterative neural networks. They feedback only the high-level description of their scene labeling system to the lowest level to increase the constraints and tune to more likely candidates using specific context-based rules. However, they have applied this process to only a very simple, simulated scene. Also, the extent of the feedback is only at one stage and there is no hypothesis generation and verification, just iteratively increasing recognition.

The feedback control strategies introduced here are multi-pass and more strictly in control. Furthermore, although we use a bounded error recognition model in our application described later and elsewhere [14], our feedback strategies are not tied to any particular matching or recognition technique. For example, we could employ Breuel's [2] grouping and matching approach which is based on high-order statistics and that target parts could be searched for assuming that their occurrence must be mutually dependent.

A serious issue within the one-pass framework is 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 single-pass hypothesis generation and verification processes. In this paper we show that the low level feature extraction stages are optimised as a complete chain rather than as individual stages [15, 13]. We discuss the feedback control scheme used to achieve an optimal set of features.

We also apply feedback control for focusing attention on interesting regions of the scene for both feature extraction and as a necessary component of hypothesis verification in the top-down stage. There will not be a single unique sequence of hypothesis generation and testing. Instead, we consider multiple feedback with decision making procedures based upon different sources of information intended to minimise the search time and focus on regions of the image where the target object may be located. To be successful, the feedback strategies must quickly identify false alarms and merge partial hypotheses together. Control mechanisms for performing these tasks are explored. To identify the regions of interest we employ an interest operator to give a quantitative measure based on perceptual grouping of low level object features [4].

There is a vast body of work based on focus of attention, e.g. [18, 9] to name but a very few, describing the benefits and implementations of attentive processing most usually modeled on the psychophysical and attentive behaviour of human vision. Most of such implementations are based on attentional spotlight/beam mechanisms and/or multi-resolution pyramidal structures. Lack of space does not permit extensive comparisons but it suffices to say, that we also base our approach on the principle of selective tuning and that our approach differs through the use of complex feedback strategies (and that it is not pyramidal). Our intention is to introduce novel control strategies rather than direct recognition algorithms.

An alternative schema for deriving processing and recognition strategies is presented by Draper and Hanson [5] in which knowledge-directed recognition strategies are learnt under supervision from training images and a library of generic visual procedures; recognition is then applied as a sequence of hypothesis generation and verification tasks at multiple levels of representational transformation. Thus, their Schema Learning System learns and controls which sequence of image processing procedures should be used to accomplish object recognition. However, whereas their system requires to learn about both the structure of object at hand and possibly several associated control strategies, we present our investigations in a feedback control strategy which is potentially applicable to most object classes.

We also use focus of attention on small regions to extract missing features using a top-down model. Thus, after locating an object, we will be able to obtain detailed information about it. For example, this could be used to control the zoom of the camera or the flight path of an aircraft. Once a candidate object is identified, a general class-wide model will be used as a basis for searching for missing information, fusing partial hypotheses and deriving a final measure of interest for the outputs. We shall consider the feedback strategies required for the recognition of a generic class of bridges to be found in IR images of noisy and cluttered scenes.

2 The Vision System

Figure 1 represents a flow-diagram and a schematic of the vision system developed. The system is divided into three sections: the low level processing modules, the intermediate levels which will derive the object features and group them into structures that resemble, more or less, the target object being sought; and the high level modules for model matching. These are all controlled by our feedback control modules. This is a basis for a plug-in modular vision system so each stage can be swapped appropriately for different applications, including the corresponding feedback control modules.

Worrall et. al. [1] discuss a model based vision system for classifying and tracking moving vehicles. In their work motion is used for segmentation and subsequently model invocation. Furthermore the models used are 3D geometrical representations together with calibrated camera and scene models. Our feedback control strategies are designed for situations where there is not such a rich pool of information and we have available only an intermediate representation of an object class. Furthermore, we need to generate hypotheses in still images. Thus, unlike normal model-based object recognition systems in which the knowledge of an object's appearance is provided by an explicit model of its shape, we adopt a *signature* and a *stereotype* of the object [13] as model and image representations respectively, with both being very loose and generic descriptions of the object based on their functionality after Stark and Bowyer [17]. The signature provides the feedback control system with a bottom-up description of the target and the stereotype provides it with matching criteria and the ability to decide when to terminate the recognition feedback control, as well as the initiation of the feedback stage for missing information search.

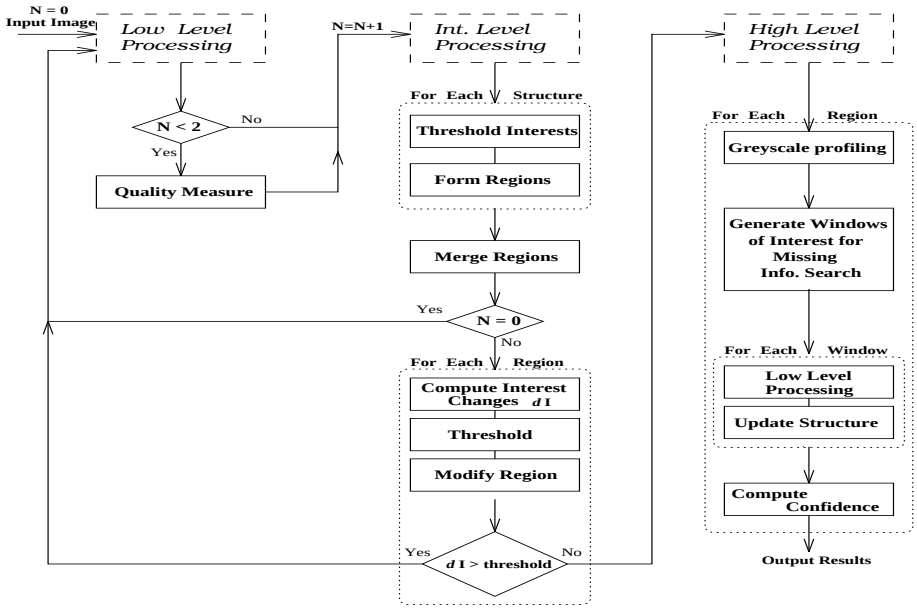


Figure 1: Overview of the feedback control mechanisms.

3 Feedback at Low Levels of Processing

A major role of the control structure, as shown in Figure 1, will be to make quantitative measures of the performance of the system. At the low level stages this will be required for the adjustment and optimisation of the various parameters in the algorithms. Lindeberg [10] approaches the issue of optimisation by using scale space theory to detect features at coarse scales followed by feature localisation at finer scales. He applies this to the detection of features such as junctions and edges. In our system, the various parameters are controlled by the low level feedback modules. We ran the low level modules at four different parameter settings. A cubic was then fit to the data and from its turning point the location of the optimal parameter setting was obtained. If no maximum turning point could be found, we returned to initial state and extended the parameter range automatically. In total we optimised five parameters [12]; these were the mask width at edge detection stage, the lower and upper hysteresis thresholds at the linking stage and finally the kernel widths in ρ and θ at the Hough transform stage. This approach is similar to that of Peak Holding in Control System theory [6].

The measure we use to determine the performance of the low level procedures is called the *performance quality measure* [15] which is the outcome of the whole low level stage - edge detection, linking and Hough transform - to which we fit the cubic equation for varying parameter values. This is a different concept to measuring performance of the individual procedures within the processing chain [16].

In Figure 1 we see that the quality measure is firstly applied to the whole image and then again, on each iteration, to each set of regions of interest. In our present implementation the optimistaion process is a computationally expensive process

but we further justify the need for optimisation in [12].

4 Feedback and Intermediate levels of processing

After low level optimisation, the output features are grouped into structures which resemble, more or less, the signature of the target object using junction finding and perceptual grouping. The junction finder groups together sets of lines to form second order junctions, third order junctions and occluded junctions. We then use perceptual grouping to group lines and junctions which are likely to belong to the target object based upon rules for symmetry, parallelism and orthogonality. We 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 [11].

Each component of the target object and its association with other components, through the perceptual organisation rules, contributes to an interest measure for the object as the grouping happens. This interest measure [4, 13] is designed to increase exponentially and the more numerous the number of component features in the structure, the greater our interest in the structure as a hypothesis becomes. For each structure we form a region of interest. Some regions will cross-over heavily and they may have arisen due to the same object contributing to more than one structure. These regions are then merged (Figure 1). Also, some regions may need to be grown outwards if any important features are touching the region-borders. Thus, the feedback proceeds and attention is further focused on those regions most likely to contain the target object. The interest level will remain static or increase for interesting regions and in most cases drop for false alarm regions. In this way, feedback removes objects that show only vague similarity to the target object sought and instead, further focus on real targets.

5 Feedback In Top-Down Analysis

Let's assume now that there is one region of interest containing a single structure of very high interest. We invoke a model-matching stage (hidden in the High Level processing box in Figure 1) which is essentially an error bound matching model although we could have also employed a more elaborate technique, e.g. [7, 2]. Our matcher uses the stereotype of the target model to analyse the structure and determine if a match can be made. At this stage some lines will be discarded (as clutter) which will reduce the interest measure for the structure.

Once we verify our hypothesis through successful model matching, we analyse the region of interest further to identify possible missing information. Our feedback control produces new hypotheses to focus attention on regions in and around the object of interest and seeks further evidence. In the ACRONYM system [3] this is achieved by referring back to a precise 3D model. In our application described later, we do not hold such information; only some loose information via the stereotype. But we can emphasise the independence of the feedback control process by stating that an alternative module carrying out such a task could be

plugged in at this stage. In our case if any missing information is found, it increases our confidence in the final result. If not, we accept our earlier match and presume occlusion or an occurrence of a generic form of the target object.

6 Summary and Results

We now summarise and consider the strategies for dealing with the fusion of information and decisions on how the feedback should proceed using a bridge detection application. Figure 2 illustrates how the hypotheses generated during the feedback procedures evolve. In Figure 2(a) we show an IR image of a scene with the set of line segments produced by the low level routines after optimising the parameter settings superimposed on it. This corresponds to the initial cycle of the low level steps. We note that the line segments associated with the bridge area are very poor representations of the bridge. In fact this is a generally difficult type of bridge as there are very few bridge-supports and therefore less overall evidence. Furthermore, it is evident that initially in the whole image, a number of line groupings can be regarded as a signature of a bridge. This is verified in Figure 2(b) after the first iteration when four hypotheses are generated as candidate bridge regions. These hypotheses are formed upon grouping line features and computing interest levels for the various groupings obtained. Of these four regions, the interest level of the region surrounding the bridge is many orders of magnitude larger than the interest in the next best region. The actual interest values themselves are not significant, but this emphasises the way that the interest operator can distinguish regions associated with the target object from false alarms.

As shown in Figure 1, we feedback to significant regions and Figures 2(c) and 2(d) show the results of the top-down phase of the feedback control, after hypotheses verification, on the third and fifth iterations. This entails verification of the components of each structure by resorting to the stereotype of a bridge. The stereotype of the bridge-type we are considering here is a long line running along the top of the bridge, and a set of near-orthogonal lines representing the supporting structure. This stereotype is only a loose model for the target and it is possible, as with the signature of the target, that other objects may be found that match it. We believe, however, that real bridges will result in higher levels of interest and hence higher confidence levels on output. This will therefore not detract from the system's performance. At this stage of processing all extraneous line segments not corresponding to the rules for the stereotype of the bridge are rejected. At each feedback stage hypothesis validation is achieved by comparing the level of interest generated with that found on the previous pass through. For regions which are false alarms, this interest level usually decreases significantly, and never rises at the same rate as regions which do contain the target object and the control module terminates the bottom-up procedure. The remaining regions will all have high interest and so are very likely to match the target object (or part of the target object). At each step of focus of attention we can extract a new set of features either by re-using the parameters from the previous stage or by allowing the control strategies to re-optimise within that region.

Finally, a successful match in a region whose interest measure has become stabilised, initiates the search for missing information. For bridges, the profile

of the smoothed and differentiated grey level values along the columns below the bridge-span are computed. This would give a number of peak values for bridge supports assuming that they are of fair contrast to the region below the bridge which in turn is of a uniform nature. Having thresholded the absolute peak values, we generate small windows of interest for missing information search as displayed in Figure 2(e). Figure 2(f) shows the ultimate results when the new supports, detected after edge detection and Hough transformation in the windows of interest, are fused with the bridge structure from the model matching stage in Figure 2(d). Figures 3 shows more results obtained for three other bridge structures where there is even more clutter.

7 Conclusion

In this paper we have presented complex feedback strategies designed to allow a more flexible approach to hypothesis generation and verification. It loops back through the low levels of processing to produce the best possible sets of features. An interest operator is then used to focus attention on regions most likely to contain the target object, and reject false alarms generated by this stage. The system then feeds back to these regions to improve the set of features that can be extracted. A loose object model is used to determine the best structure resembling the object of interest and the region that it occupies. Feedback continues on successful regions until the interest measure stabilises. Using this region we again employ feedback and look for missing information and increase our confidence in the output hypothesis.

A further use of our feedback system is that it uses the low level information from the image to minimise the search through all possible hypotheses, and so can locate target objects in acceptable time-scales. An alternative schema based on learning multiple control strategies has been suggested [5]. However, this concentrates on the choice of recognition algorithms to achieve a visual goal, rather than on methods that attempt to maximise the exploitation of the information content of the visual data. Our current system is slow at the optimisation stage, as are most techniques, e.g. [10], but this does not underly the principle of the use of feedback strategies, and we can plug in a faster technique when necessary.

An interesting point to note is the division between protagonists of the principles of minimum commitment and maximum commitment in visual processing. The former insist that hard decisions should not be made at any level of processing; the latter contradict this by providing psycho-physical evidence of maximum commitment in human visual processing. Our work is effectively a half-way house between these two (the principle of evidential commitment) based on our focus of attention given a region of high interest. Initially, commitment is made to generate hypothesis and identify regions of interest. This commitment is then relaxed as a result of feedback when the interpretation process returns to the raw image data to refine the focus of attention and/or to verify the hypothesis.

Overall, the feedback control approach has a great deal of importance in the field of object recognition in terms of better recognition performance and speed of interpretation. In the future, we intend to use feedback strategies in motion

prediction in image sequences for mobile robot and aircraft tracking applications.

Acknowledgments: The authors wish to acknowledge the support of the DRA, Farnborough, UK, for this work.

References

[1] A.D.Worrall, G.D.Sullivan, and K.D.Baker. Advances in model-based traffic vision. In *BMVC93*, pages 559–568, 1993.

[2] T. M. Breuel. Higher order statistics in visual object recognition. Technical Report TR #93-02, Institut Dalle Molle d'Intelligence Artificielle Perceptive, Switzerland, 1993.

[3] R.A. Brooks. Model-based three-dimensional interpretations of two-dimensional images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):140–149,, 1983.

[4] 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.

[5] 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.

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

[7] 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.

[8] G.C. Lai and R.J.P. de Figueiredo. Image interpretation using contextual feedback. In *Proc. of IEEE Int. Conf. on Image Processing*, pages 623–626, 1995.

[9] V.F. Leavers. Preattentive computer vision: towards a two-stage computer vision system for the extraction of qualitative descriptors and the cues for the focus of attention. *Image and Vision Computing*, 12(9):583–599, 1994.

[10] T. Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision*, 11(3):283–318, 1993.

[11] D. Lowe. *Perceptual Organisation & Visual Recognition.* Kluwer Academic, 1985.

[12] M. Mirmehdi, P.L. Palmer, J. Kittler, and H. Dabis. Framework for control of parameters in early vision. *Submitted to ECIS Workshop on Computer Vision, Rosenen, Sweden*, 1995.

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

[14] M. Mirmehdi, P.L. Palmer, J. Kittler, and H. Dabis. Multi-pass feedback control for object recognition. *Accepted for publication in VI 96, Toronto*, 1996.

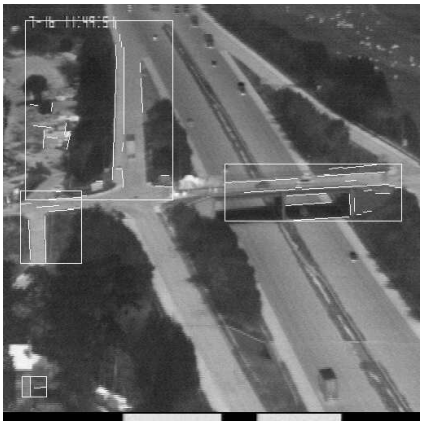
[15] P.L. Palmer, H. Dabis, and J. Kittler. A performance measure for boundary detection algorithms. *Accepted for CVGIP: Image Understanding*, 1996.

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

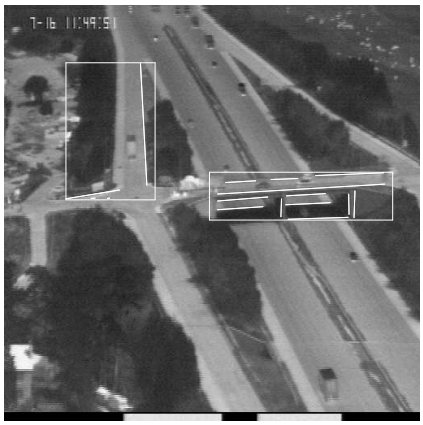
[17] L. Stark and K. Bowyer. Achieving generalized object recognition through reasoning about association of function to structure. *PAMI*, 13(10):1097–1104, 1991.

[18] J. Tsotsos, S. Culhane, W. Wai, Y. Lai, N. Davis, and F. Nuflo. Modelling visual attention via selective tuning. *AI*, 78:507–545, 1995.

(a) Lines in IR bridge image



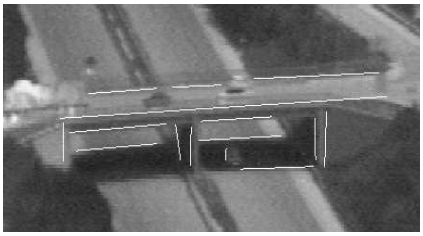
(b) 4 best ROIs - 1st iteration



(c) 2 best ROIs - 3rd iteration

(d) Best ROI - 5th iteration

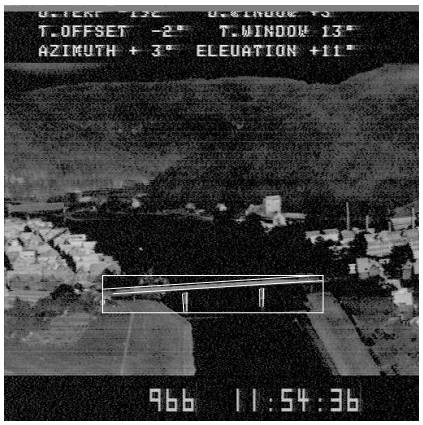
(e) Windows of interest



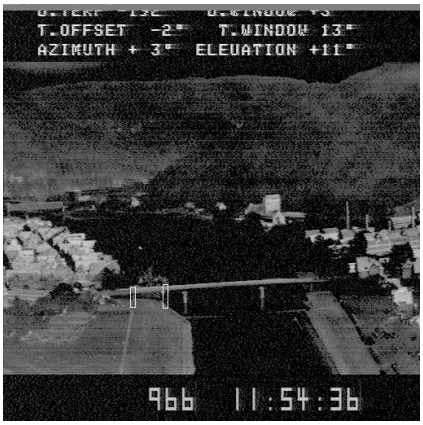
(f) Final results

Figure 2: (a) Significant lines in IR image, (b) Four ROIs after first feedback (c) Best two ROIs after 3rd feedback, (d) After 5th feedback, (e) Windows of interest for missing info. search, (f) Final results

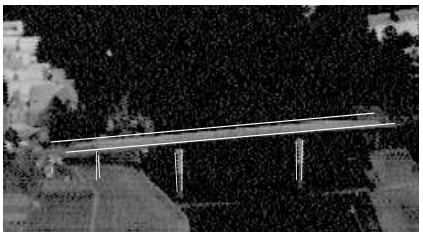
(a) Original IR bridge image



(b) Best ROI - 4th iteration



(c) Windows of interest



(d) Final result

(e) A Bridge in cluttered image

(f) Another bridge

Figure 3: (a) Original IR image: (b)-(d) are γ-corrected for enhanced viewing, (b) After 4th feedback, (c) Windows for missing info. search, (d) Final results, (e) and (f) two more bridges