Towards a control model of object recognition

R.Young and J. Illingworth
Centre for Vision, Speech and Signal Processing,
School of Electronic Engineering,
Information Technology and Mathematics,
University of Surrey, Guildford GU2 5XH, United Kingdom.
e-mail: R.Young@surrey.ac.uk

Abstract

In this paper we present some preliminary investigations into the development of an active vision system with the aim of developing a real-world model of simple visual behaviour, based upon a control theory [4] view of purposive behaviour. The goal of the system is to control its fixation with respect to objects of a relatively complex nature.

Keywords: Fovea, PCT, fixation, segmentation, clustering

1 Introduction

Previous work in the area of object recognition has concentrated mainly on matching geometric models with information derived from single, mainly, images. Research that has involved multiple images, whether from stereo or from sequences obtained from mobile sensors follows a similar rationale with additional information from the extra images enhancing the construction of the observed model. The main influence on such computer vision research has been David Marr [2] who proposed a computational, reconstructive approach to visual processing that has little to do with the vision of natural living systems. The success of other object recognition research in dynamic scenes has been limited to the tracking of simple outlines, motion and objects of a single colour [3, 1, 6].

The work presented in this paper represents a shift away from traditional approaches of computer vision towards a more natural *control* model based on a hierarchy of signals exemplified by Powers' Perceptual Control Theory (PCT) [4]. The methods employed in this system follow the standard design of PCT controllers along with conventional computer vision techniques of segmentation.

2 Scene representation

For the purposes of biological plausibility and computational (efficacy) the scene as viewed by the tracking system is represented by a distribution of visual elements similar to that of the human retina. The centre of the field of view is sampled at a high resolution decreasing logarithmically to to the periphery.

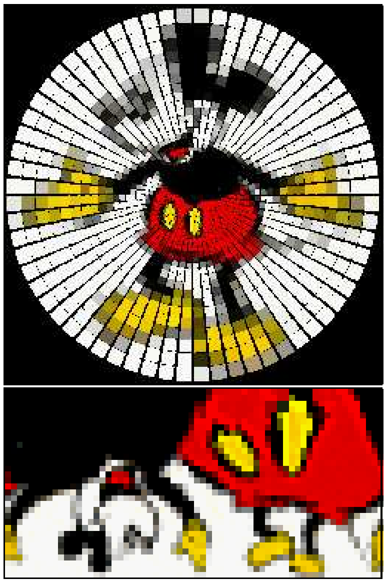

Figure 1: The foveal representation of a well-known cartoon character field of view.

The top half of figure 1 shows what such a view would look like. Each square area represents one colour input signal and is taken as the most prominent colour which falls on that area. There are 32 rings of square regions each with 64 elementrs. These can easily be mapped into a rectangular array which is more suited to processing within a computer program. The array of the same scene is shown in the bottom half of figure 1, where each row represents one ring. The rings from the fovea to the periphery map to the top to bottomn rows, respectively. This foveal representation of 2,000 pixels signifies a substantial reduction in the amount of the information that needs to be processed, compared with the standard uniform image of 60,000 pixels covering the same

3 Fixation input signal

In succeding sections we describe how particular regions of interest in a scene are segmented from the background. Since each row and column of the pixels, in the foveal distribution, that make up the region of interest represent the direction and magnitude from the centre of the field of view we are able to derive a fixation signal which can be used as the input to a standard PCT control system. Sparks [5] reports that animal visual fixation works in a similar manner. Populations of cells in a neural map in the *superior colliculus* define the direction and magnitude of eye movements.

The fixation signal is derived by simply taking the mean of all the position vectors within the region of interest. Figure 2c shows the foveal representation of figure 2a where the small white blob corresponds to the white circle in 2a. The dark line in 2a from the central cross hair is a visual representation of the fixation input signal derived from the mean of the position vectors of the blob. Figure 2b shows the end result of control of the fixation signal. The cross hair tracker is now centred on the target circle. Notice in 2d that the circle now corresponds to a white band in the foveal view. What has happened is that the tracker has moved until all the position vectors are in equilibrium (their average is zero) resulting in the fixation on the centroid of the region.

Fixation on the centroid occurs not only regular geometric figures but also for irregular shapes as shown in figure 3. The image in figure 2a contains a number of irregular coloured shapes. The foveal view when fixated on the centroid of each object is shown in figure 3b.

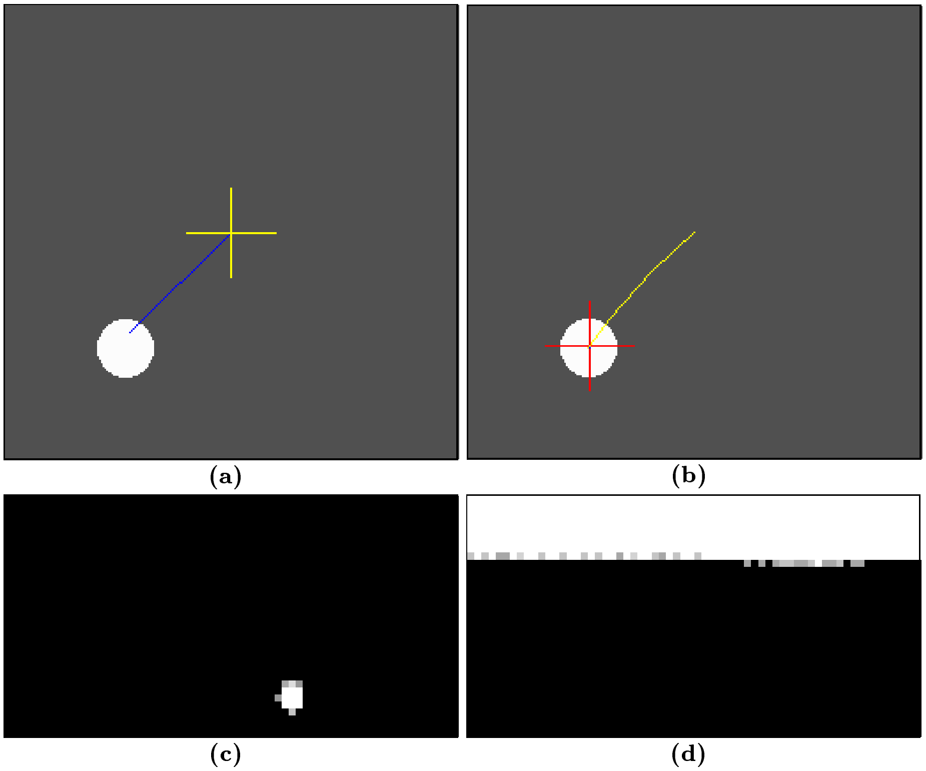

Figure 2: A simple, single-level fixation control simulation. Images (c) and (d) are the foveal representations of the initial (a) and final (b) uniform scenes, respectively.

4 Image and Robot output

Our experiments are performed within static images, offline, and in real-time with live static images as well as with a PUMA700 robot arm and camera system. In the static experiments the movement of the robot is represented by

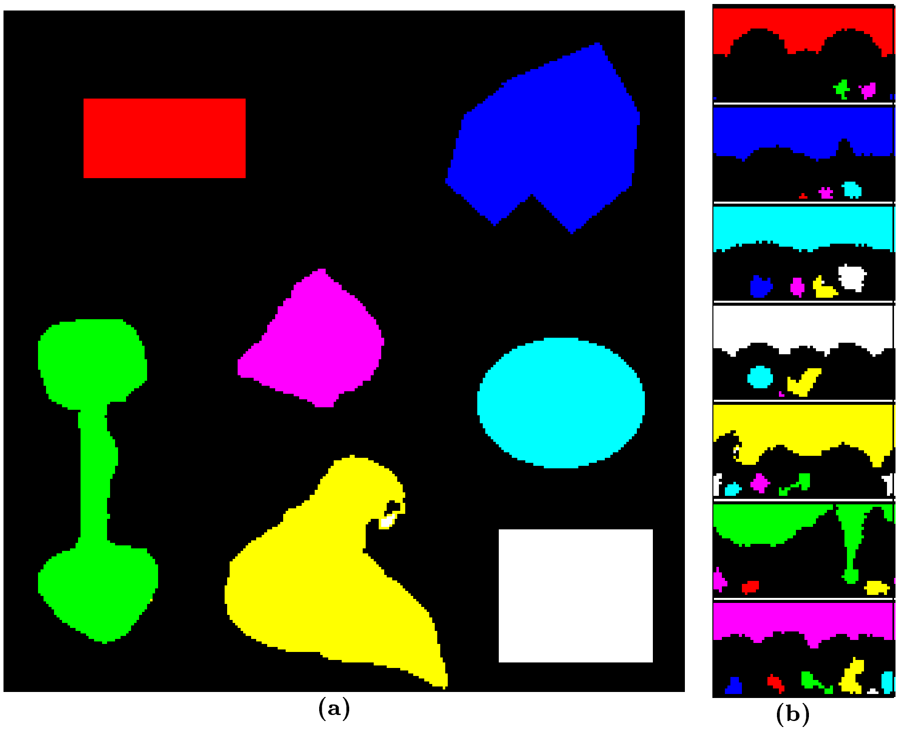

Figure 3: Simple colour fixation. a) The image of simple coloured figures, b) The foveal view when fixated on the figures, clockwise from top left.

a moving cross-hair. The input which is controlled is the size of the offset from the centre view to the target, with the reference signal being zero. The error signal, therefore, is the same as the input signal.

The output signal is the direction and *velocity* of the movement towards the target and the velocity is a linear function of the error signal. So, as the sensor centre gets closer to the target the velocity decreases until fixation, when the error will be zero and, therefore, the velocity. Relating the error signal to the velocity, in this way, avoids oscillations and jerky movements as fixation is reached.

With the real-time controller it is possible to execute commands defining the direction and velocity of movement required. The image processing is performed in parallel with the robot movements and so it is not necessary to wait for a movement to cease before updating the error signal. Also commands can be sent to the controller while the robot is in motion which override all previous commands. In this way we are able to continually monitor and control the fixation signal.

Control within static views is handled in the same way with the exception that the movement in each iteration is computed discretely from the current

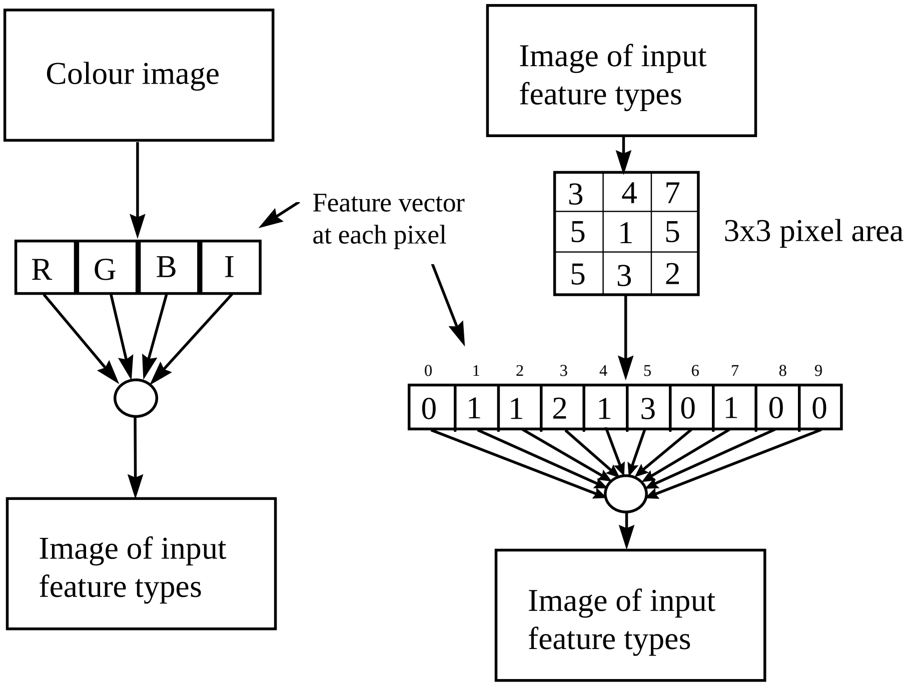
Colour image
R
G
B
I
Feature vector at each pixel
Image of input feature types
Image of input feature types
3 4 7
5 1 5
5 3 2
3x3 pixel area
0 1 2 3 4 5 6 7 8 9
0 1 1 2 1 3 0 1 0 0
Image of input feature types

Figure 4: On the left is the input function for the lowest level, the RGB image. On the right the higher levels derived from a 3x3 pixel region of its preceding level.

desired velocity and the length of time of each iteration.

5 Single-level Control

Single-level control is sufficeint for tracking simple lights or areas in grey-level or colour scenes. Areas within an image of a particular grey-level range (such as the brightest) can easily be segmented, from which the fixation signal of a blob can be derived. Similarly for colour regions, by defining the upper and lower thresholds for the red, green and blue values. Figures 2 and 3 show examples of tracking to simple regions in simulated images. Tracking experiments to simple lights and single-coloured objects have been performed successfully in real-time with the robot.

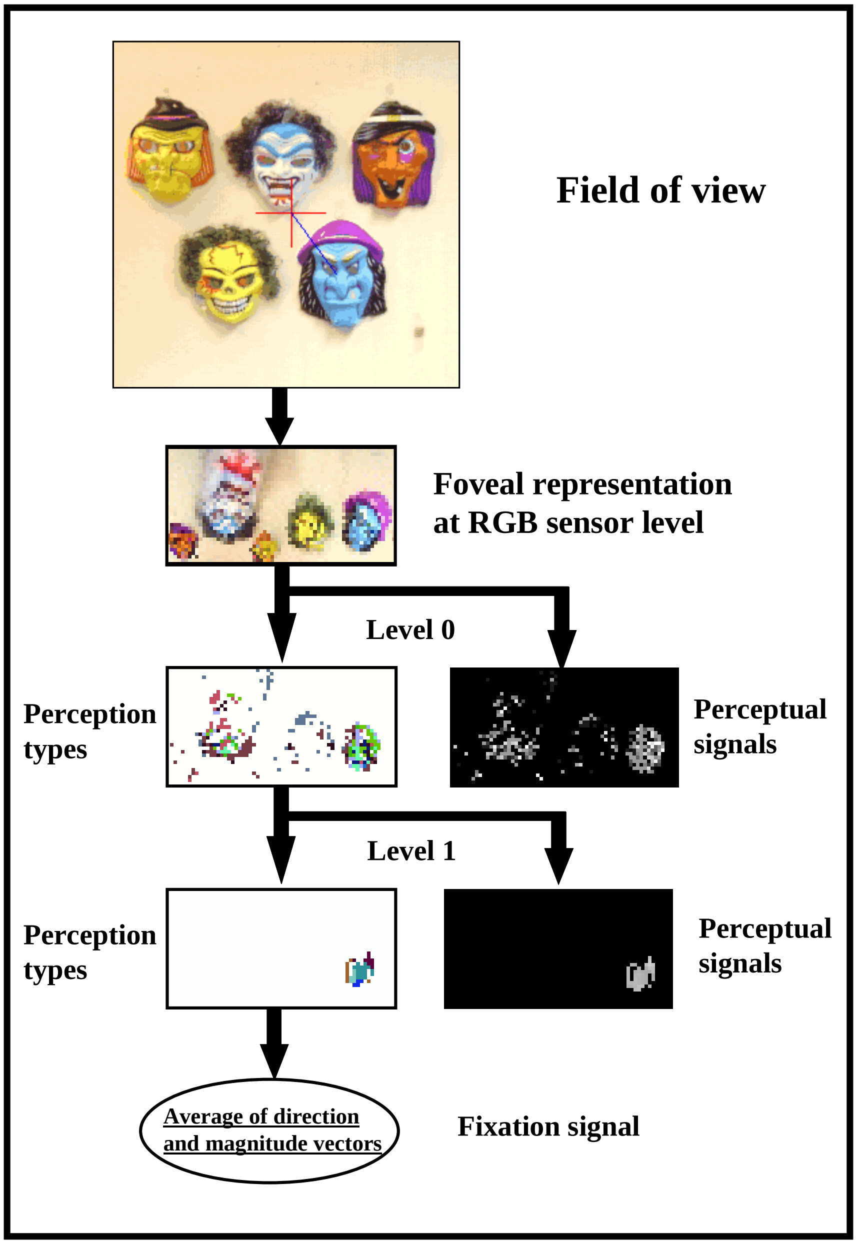

Figure 5: The two level colour processing control system used in our experiments. The outputs from these levels, of the magnitude and direction to the target, define the input to the highest level (fixation) control system.

6 Object model representation and acquisition

A couple of problems arise when extending tracking control to multi-coloured objects,

- determining the RGB values of the different colours which belong to a target object
- distinguishing between areas of the same colour which belong to different objects (or the background)

The first problem is partly addressed by the method of model acquisition employed. The target object is isolated from its surroundings and the RGB vectors at each pixel are recorded and clustered (for the purposes computational efficiency) into a small (10-20) number of *ideal* vectors which are said to represent the input vector weights for the object when it is assumed to be under perfect control. This enables single-level, multi-feature control.

Input vectors at higher, additional levels are derived by examing a 3x3 area of the preceding level. Within this area the feature *types* are counted giving an input vector which is the length of the *number* of possible features (see figure 4). Adding these higher levels partly solves the second problem as the input vectors will be more specific to the target object than to others.

7 Multi-level Control

Figure 5 shows a block diagram of a multi-level control system. Level 0 processes the basic RGB vectors and higher-levels (only one is shown) the vectors from the 3x3 pixel area. The input which is controlled at the highest level is the the perception of the direction and magnitude of movement to the target.

Some preliminary results of the multi-level control system are shown in figure 6. Each row of images show the results of fixation for each of the halloween mask targets, clockwise from top left. The columns, from left to right, show the results with levels 0, 1 and 2. In each case the starting position is the centre of the image and the cross-hair indicates the end (which should be the nose of each face) position with the dark line showing the course of fixation.

From the left column it can be seen that control, solely with level 0, is poor. Although fixation is made towards the correct targets interference from background and extraneous signals adversely affects the fixation signal. Control which includes level 1 (centre column) is greatly improved, with fixation (ending), correctly, at the centre of the target face each time. Including another level (level 2, right column) does not seem to improve control further and in fact seems slightly worse. However, this is probably more to do with the fact that much of the signal is lost at this level than with higher levels not being of benefit. Given the instability of the input signal at higher levels we limit the hierarchy to the lower two feature processing levels.

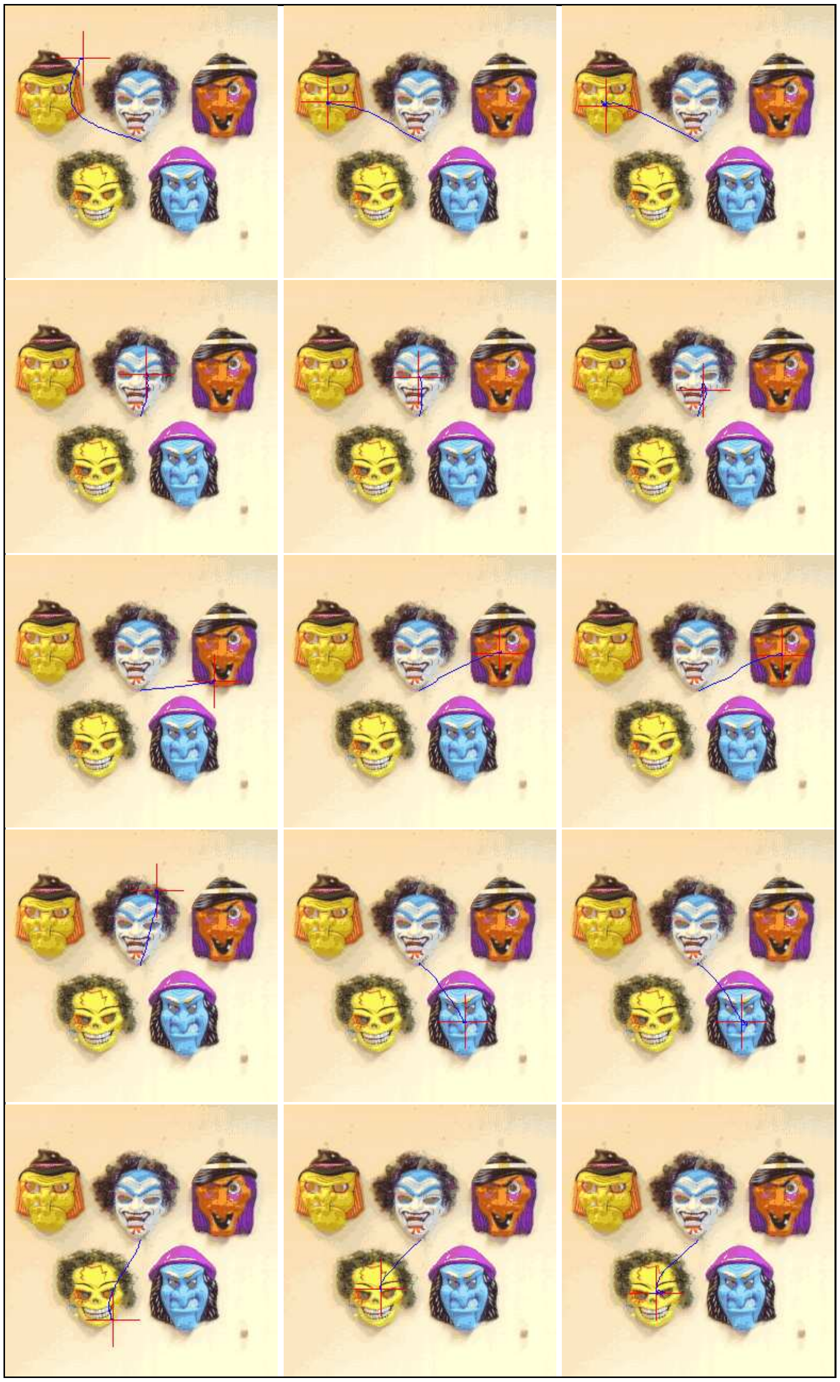

Figure 6: Multi-level control

8 Conclusions

The fixation system presented in this paper performs well in real-time on simple lights and single coloured figures in synthetic and real scenes. Results have also been presented of some preliminary work concerning fixation to more complex, multi-coloured objects. Control improves with added levels in a hierarchy. Each level embodies signals which are more specific to the target object enabling the target to be more easily distinguished from its surroundings. The main problem is deriving the input functions and their weights. In the present scheme the signals at the higher levels are rather improverished with much of the lower level inputs being lost resulting sometimes erratic control. Future work would benefit from further investigation into the reorganisation and development of the input functions.

We have presented some preliminary results in offline images which show that reasonable fixation control, to complex objects, can be achieved with signals based only upon colour. Control may be improved further by including feature dimensions such as edges and motion to add even greater discrimination.

References

[1] Paul Hoad and John Illingworth. Automatic control of camera pan, zoom and focus for improving object recognitiony a moving observer. In *Appendices to PPR-3 of VAP II*, chapter D-2. Esprit Basic research Project 7108, 1995.

[2] David Marr. *Vision: A Computational Investigation into Human Representation and Processing of Visual Information.* Freeman, San Francisco, 1982.

[3] Peter Nordlund and Tomas Uhlin. Closing the loop: Detection and pursuit of a moving object by a moving observer. In *Appendices to PPR-3 of VAP II*, chapter B-4. Esprit Basic research Project 7108, 1995.

[4] William T. Powers. *Behavior: The Control of Perception.* Aldine DeGruyter, Hawthorne, NY, 1973.

[5] David L. Sparks. Translation of sensory signals into commands for control of saccadic eye movements: Role of primate superior colliculus. *Physiological Reviews*, 66(1):118–171, January 1986.

[6] M Spratling and R Cipolla. Uncalibrated visual servoing. In *BMVC*, pages 545–554, 1996.