Iconic Object Matching Using Primal Sketch-like Features

R. B. Fisher, T. D. Grove, A. Gionis, A. MacKirdy
Department of Artificial Intelligence, Edinburgh University

Keywords: Feature detection, Appearance based recognition, Log-polar transforms, Attention

Abstract

In this paper we present an iconic object matching approach. The question that we were investigating was how to integrate and evaluate a number of functions known to exist in the primate vision system in the context of a machine vision system. The testbed was a system that attempts to recognize objects using a form of image matching, rather than well-explored symbolic-model based matching paradigm. The system consists of two parts: an attention system that locates objects of interest (described in [5]), and a multi-scale foveated matching algorithm that uses primal sketch-like features to identify objects. The matching algorithm uses a stack of features obtained from intensity data through low-level feature detection operators. We show how this approach to object recognition can achieve scale, translation and rotation invariant recognition of iconically represented objects with some generalization ability.

1. What is the original contribution of this work?

   There is only a small amount of recent iconic-based recognition research. None use Marr's primal-sketch features in a log-polar context. This paper shows how one can integrate feature extraction with an appearance-based recognition approach.

2. Why should this contribution be considered important?

   The symbolic approach to object recognition has been successful with CAD objects, but has not had much success beyond that. The appearance approach is not yet well explored, and has much promise. The paper presented here extends the limited work published previously, so as to improve the solution to several open problems.

3. What is the most closely related work by others and how does this differ?

   Rao and Ballard [8] used an n-dimensional feature vector for classification, but did not exploit the spatial distribution of the image features. Schiele and Crowley [11] matched 2D intensity histograms, but this also ignores the spatial distribution of image features. Seibert and Waxman [12] used an ART network to match feature vectors extracted from log-polar processed images, but only investigated binary images. Siebert and Eising [13] used the log-polar architecture, only with a difference-of gaussians receptive field. Their matching scheme used templates applied directly on the log-polar image, rather than extract features.

4. How can other researchers make use of the results of this work?

   The paper does not present self-contained mathematics nor a new technique, rather a new way to investigate an old problem. The investigation is likely to take some years to complete, so this paper should be seen more as introduction of a new research area.

1 Introduction

It is hypothesized [1] that humans use two systems to process visual information. One system matches geometric feature descriptions extracted from an image to view-invariant symbolic models, typified by Marr's [7] geometric models. This paradigm has been much explored in the context of machine vision. The other hypothesized human system matches iconic models to the image data, or low-level extracted features, along the lines of Marr's primal sketch. Only a small amount of machine vision research has pursued this approach, in part because of problems with changing illumination and the need to perform image registration at the correct position, rotation and scale.

Iconic representations, being image oriented, are viewer-centered descriptions of objects and as such are highly sensitive to the viewpoint. Different views of an object must be represented by different models. Iconic models are also weak at capturing subtle global differences between objects (such as the difference between a many-sided polygon and a circle) and the descriptions are not compact, unlike symbolic descriptions. However, little processing has to be performed before the data can be matched against the models and the models are stable, as small changes in the image data cause only small, local, changes to the descriptions. It is also relatively easy to design a matching algorithm that is insensitive to small changes in view (*i.e.* is quasi-invariant) and better at capturing important differences between models.

Rao and Ballard [8] used a number of filters (derivatives of gaussians at several different scales) to build an n-dimensional feature vector. The feature vectors are then fed into a simple neural net which associates each vector with one of a number of objects. Their system is able to distinguish between a large number of objects under varying pose by learning a set of poses. How-

ever, their system suffers from the fact that, because of the filtering, it will be unable to distinguish between objects with similar frequency responses, nor does the global filtering approach represent the spatial distribution of features necessary for distinguishing subtle appearance differences. Schiele and Crowley [11] matched 2D histograms of pairs of image properties (mainly gradient-based), and achieved good matching results (using a χ^2 metric) but their approach also ignores the global organization of the image features. Seibert and Waxman [12] used an ART network to match feature vectors extracted from log-polar processed images. Their features were interest points extracted from binary images of single isolated objects. 2D image-based recognition was linked into a 3D aspect and multiple competing identity object recognition network. Siebert and Eising [13] used the log-polar architecture, only with a difference-of gaussians receptive field. Their matching scheme used templates applied directly on the log-polar image.

This paper describes an investigation started by Grove *et al* [4, 5] into iconic vision, that is, performing visual tasks (*e.g.* object recognition) using pictorial data obtained directly from images. The question that we were investigating was how to integrate and evaluate a number of functions known to exist in the primate vision system in the context of a machine vision system. The approach has three main components:

1. A feature extraction mechanism that provides the input to the system, capturing important detail and suppressing noise.

2. A model matching mechanism that copes with variability in the relative rotation, scale, illumination, etc. between the models and data.

3. A visual attention mechanism (described in [5]) responsible for locating items of interest. This is important in an iconic vision system, since

we cannot afford to search the entire visible world for models.

The research described here concentrates on only 2D image recognition, but exploits the log-polar mapping for scale invariance, uses an attention mechanism to control model registration in the image and uses primal-sketch type features to exploit object shape in grey-level images. The system developed is not intended to model the primate vision system, but is inspired by some of its competencies and behavior.

2 Architectural Overview

We use a foveated (r, θ) log-polar coordinate system [14] for retino-centric coordinates, with 20 bands, each containing 48 sectors. The receptive fields (*i.e.* the area of the (i, j) image from which they take input) of each pixel in the (r, θ) representation increases (logarithmically by 1.2) as r grows larger, in order to cover the entire foveated area. The receptive fields in the innermost bands take their input from only one or a few pixels, averaging the value. This gives high resolution around the foveation point. Receptive fields in the outermost bands average over large numbers of pixels, giving lower resolution. Receptive fields overlap by about 33% to avoid gaps which leads to a certain amount of blurring. The polar representation is attractive because it maps rotation and scaling into translation, and this feature is used in the matching algorithm described below to deliver scale and rotation invariance.

The main representations are:

1. **The World** - a large static (r, g, b) image (here 512^2) within which the iconic matcher saccades and extracts smaller (here 128^2) foveated views.

2. **The Image Stack** - Foveating the world image maps part of the raw (r, g, b) image to (r, θ) space, to form the first part of the image stack. The remainder is extracted by operators described in Section 3.

3. **The Model Base** - a set of models that may be matched to the current image stack. Each model has the same format and contents as the image stack. Each feature plane has a weight associated with it indicating how useful the feature is in identifying this object. In addition, each model may have a list of associated models (*e.g.* an eye may link to a likely nearby nose position). This list is of the form { (*model_type, relative_position, importance_weight*) }. These links can be used to also form an iconic model hierarchy (to be described in another publication). Models are created by selecting pictures that are representative of the class the model will denote. A model is normally registered on a feature that will attract the attention system[5]. Models are learned at three scales (50%, 100% and 200%) because not all features will be visible at all scales.

4. **The Stable Feature Frame (SFF)**[2] - represents the system's visual memory. It is registered on the world rather than the gaze location and incrementally records a stable, non-retinocentric view of the world. It contains defoveated (r, g, b) data obtained during the system's visual exploration, plus a list of recognized image structures (*i.e.* model instances) and their image locations and matching scores.

   Defoveation is used to map from (r, θ) space back to (i, j) space for use in both the attention mechanisms and the SFF. This results in a circular image which is blurred in the periphery and detailed in the central foveal area.

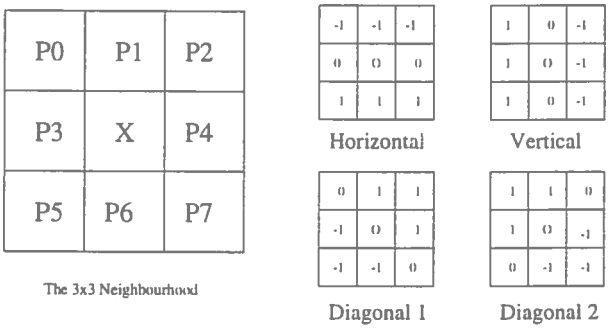

Figure 1: The 3x3 Neighborhood and Edge Templates

5. **The Interest Map** - The interest map[5] is an image structure registered with the world. Its contents record a value representing the interestingness of a given point in the scene. Interestingness values increase as center-surround and corner image features are found, and as models are identified (as these predict locations of likely associated models). Interestingness values decrease at parts of the image that have been explored.

   The scene is explored in a saccade-like process by selecting the current highest interest point as the next location to foveate.

3 Feature Description Extraction

The descriptions extracted here are edge, bar, center-surround and corner/end features [7] and are motivated by the standard interpretation of neurophysiological investigations of the retina and primary visual cortex. All feature detection operators are based on a 3x3 operator (see Figure 1), and are applied at each point in the (r, θ) image to yield a new image. Rather than use larger operators (*e.g.* as in [9] who used unfoveated, linear operators at a single

scale), we reapply the 3x3 operator to a scaled intensity image. Halving the scale of the source image is equivalent to doubling the size of the operator. Here, we use three scales : 1, 1/2 and 1/4.

Each image stack consists of 42 (r, θ) feature planes (14 features at 3 scales) From the (r, g, b) planes we create an intensity image (I) using the formula $I = 0.35R + 0.45G + 0.2B$ and apply the feature detectors to the derived (r, θ) intensity image. The following features are used (at each of 3 scales): 1) raw red, green and blue intensity images (color constancy issues are ignored here.), 2) edges at four orientations, 3) unoriented corners, 4) on and off-center surround features, and 5) on and off bars at two orientations. The feature detectors are applied everywhere in the (r, θ) intensity image, in the same manner as normal image convolution. However, as these feature detectors operate in (r, θ) space, they do not detect the same features as they would in conventional (i, j) space – "vertical" edges correspond to scene edges that lie along radii passing through the foveation point and "horizontal" edges correspond to scene edges perpendicular to lines running through the center of foveation. A shift in the θ direction corresponds to a rotation about the foveation point and a shift in the R direction corresponds to a change in scale (or radial distance) from the foveation point.

Edges: We convolve the (r, θ) intensity image with the four templates shown in Figure 1 and take the absolute value of each result.

Corners/Ends: Function (1) gives a strong response to a corner.

$$negativecorner = \min(p_1 - x, p_3 - x, p_4 - x) - \frac{|x - p_6|}{2} \qquad (1)$$

The first term in this function returns a large value if there is a large difference between the center receptive field (x) and receptive fields p_1, p_3 and p_4. This defines the "tip" of the corner. The second term will suppress the first if

this "tip" is not joined to a base (i.e, if there is a large absolute difference between x and p_6.) This term is scaled, since there is likely be some intensity difference between the "tip" and "base", due to the tip occupying less area of the receptive field. Functions similar to (1) are defined for detecting white-on-black and inverted (the 'corner' receptive fields are p_1 and x) corners, but only a limited range of corner/end orientations are implemented.

Center surround features: A center surround feature is a local maximum (an on-center surround feature) or local minimum (an off-center surround feature), *i.e.* a receptive field that is either lighter or darker than all of its neighbors. An on-center surround feature detector can therefore be defined as:

$$on\_center\_feature = \min(x - p_0, x - p_1, \ldots x - p_7) \tag{2}$$

The off-center surround feature is defined with the x and p_n terms reversed.

Bars: Four bar detectors are used; an on and and off radial bar detector, and an on and and off orthogonal (orthogonal to the radial lines) bar detector. The on and and off radial bar detector is defined as:

$$on\_radial\_bar = \min(p_1 - p_0, x - p_3, p_6 - p_5, p_1 - p_2, x - p_4, p_6 - p_7) - \max(|x - p_1|, |x - p_6|) \tag{3}$$

The first term finds the minimum difference along the sides of the bar. This is suppressed by the second term, which finds the absolute difference along the bar. The other bar detectors have analogous definitions.

Examples illustrating the feature detectors applied to the face image in Figure 5 are presented in Figure 2. From these examples, it can be seen that the feature detectors work, although there is some overlap between the responses. Here, the output of the feature detectors have been scaled for display purposes, so the absolute magnitudes of the responses is not meaningful,

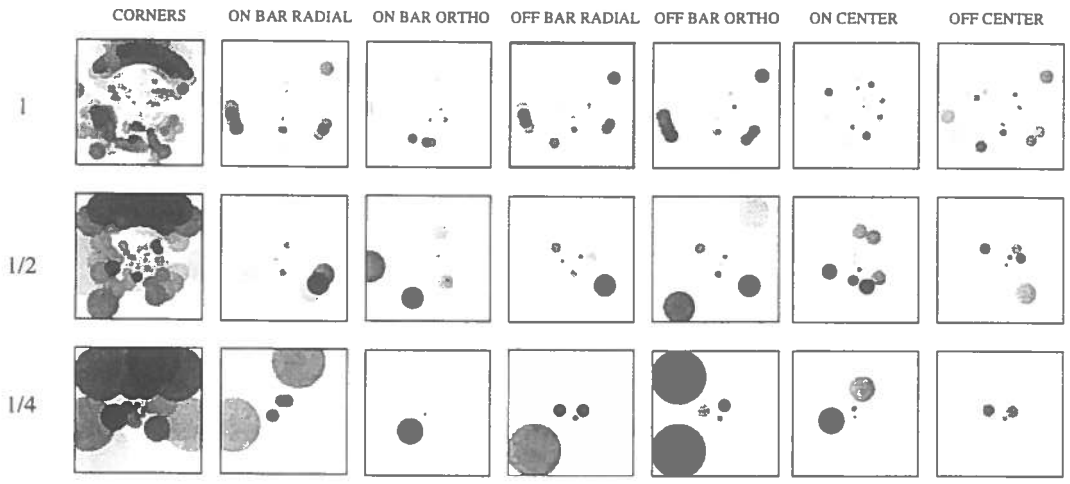

Figure 2: Feature detectors applied to face image (Fig 5).

only the relative colors inside each window. The center surround features are easier to understand, having picked up features registered with the eyes and eyebrows, amongst others. The corner detector is largely responding to texture regions.

4 Matching

The properties that we would like our matching algorithm to have are 1) invariance to small changes in rotation and scale, 2) invariance to differences in the illumination and frequency of incident light, 3) use of context to disambiguate potentially ambiguous local intensity patterns and 4) ability to recognize patterns based on their defining features while ignoring irrelevant details. In general, iconic matching involves measuring the distance between two vectors - one representing the model and one being the data - announcing a match if this distance is below a certain threshold. Ideally, a metric should

be chosen that incorporates some knowledge of the imaging process in order to go some way to achieving 1-4.

Fisher and Oliver [3] compared several standard distance metrics and concluded that a multi-variate cross-correlation function is appropriate in situations where images have a linear relationship. Their analysis showed that a good multi-variate correlation function is the average of the single channel correlations. Multi-variate correlation is simple enough to implement in specialized hardware and could also easily be implemented in parallel across a set of neurons. Our approach builds upon multi-variate correlation (like [10], except that we use grey-level and extracted features instead of binary images). The main extensions are:

- Improved generalization by variable weighting of each channel, as opposed to simply using the mean. This helps to achieve 4 above.
- Inclusion of context information, by including evidence from previously recognized models as well as feature images (to achieve 3).
- Allowing limited rotation and scale variations (to achieve 1).

The matching score for a given model has two components: the image component ρ_m, which is composed of the intensity and feature images (see Section 3), and the associated model component ρ_a. The evidence value is a function of the displacement from the predicted position and the match score of the previously recognized model. The two evidence types are combined with weighting

$$0.7\rho_m + 0.3\rho_a$$

and if the combined score exceeds a threshold (0.8 in all our experiments), then a match occurs.

The image evidence match score is obtained by correlating the 42 feature planes of the image stack with the corresponding 42 image planes of a model, using a modified form of multi-variate correlation:

$$\rho_m = f(\sum_{k=1}^{42} \rho_k w_k + w_0) \tag{4}$$

where ρ_k is the single channel match score and w_k are weights, which are learned according to a procedure detailed in [4]. These weights reflect the relative importance of the correlation scores in determining object identity. The bias w_0 reflects the *a priori* probability of data belonging to this class (defaults to 0). $f(x)$ is the sigmoid function $1/(1+e^{-x})$.

The associated model component ρ_a is calculated by predicting where associated models might appear in the scene relative to the current foveation point (given the current model's rotation and scale) and then accepting any models of the correct types found near the predicted positions as evidence (where the list of models in the SFF is searched for matches).

Thus, context is implemented in the matching algorithm by using the associated model evidence. Generalization occurs through weighting the correlation scores such that those features that are most useful in constraining object identity have large positive (or negative) weights, while those that are irrelevant have small weights.

Invariance to small fronto-parallel plane rotations around the optical axis is obtained by rotating the data to several nearby rotations during matching. Here, we assume that distinct views are represented as different models (as in *e.g.* [6], [8]). Rotating an image stack simply involves shifting all the receptive field columns along the θ axis. The amount of rotation yielded by a shift of one receptive field depends on the foveation parameters, and in this case is equivalent to about 7.5 degrees (0.13 radians). Scaling is handled

similarly by shifting the rows of the (r, θ) image vertically along the r-axis. Moving one row up or down results in a 20% change in scale.

Because the attention mechanism (similar to [10], except using a activation score rather than corner features) identifies interesting features in the low spatial frequency channels and in the periphery, the location of an interesting feature is not normally known with any accuracy. Thus, the system performs the matching once at the actual point selected by the attention system, and 8 times at locations offset from this (by 4 pixels) in each of the eight compass directions. This microsaccade process varies the foveation point about the focus of attention to improve image registration.

The matching algorithm is:

```
FOREACH model, microsaccade, scale and rotation
  FOREACH feature/intensity plane in the image stack
     1. Correlate data plane with model plane
     2. Weight and add to total Score
  ImageScore = f( Score + constant)
  FOREACH associated model in the model base
     1. Match any previously found model stored in SFF
        having correct type and near predicted position
     2. Weight and add to total LabelScore
  Modelscore = 0.7 ImageScore + 0.3 max(LabelScore,0)
```

5 Matching Experiments

Rotation Invariance: This experiment examines the approach's ability to match a model obtained from an image at one orientation to the object at different orientations and a changing background. The approach should also

Model	1	2	3	4	5	6	7	8
Cat face	0.01	0.11	0.01	0.00	0.00	0.02	0.01	0.01
Horiz. Wedge	0.00	0.02	0.00	0.00	0.04	0.00	0.00	0.00
Human Face	0.75	0.82	0.44	0.76	0.87	0.78	0.94	0.97

Table 1: Match scores for three models tested against the faces in Figure 3

estimate how much the data should be rotated in order for it to match the model. In Figure 4, a model was obtained from image 2 and matched to the face in all 3 images. The match scores, graphed against rotation needed to align the data with the model, are shown in the graph shown in Figure 4. The images are labeled with the rotations of the best match, which are quite close to the real rotation. From the graph, it can be seen that a data must be rotated quite considerably before it fails to match the model. This is because cross-correlation matching is reasonably insensitive to small rotations even without modification.

Scale Invariance: The image of the cat in Figure 5 was rescaled to twice and half of its original size and was matched to a model at normal size. After matching, the model was then rescaled using the estimated matching scale and defoveated back into (i, j) space. The original images and those reconstructed by defoveating the matched model had feature sizes that were quite close to the input image size, which suggests the scale was estimated correctly.

Generalization: A face taken from a 9th person was matched against the other 8 faces in Figure 3. For comparison, two other models were also matched: a cat face model and a horizontal wedge model. The results of the matching is given in Table 1. The face scores are high, which reflects the

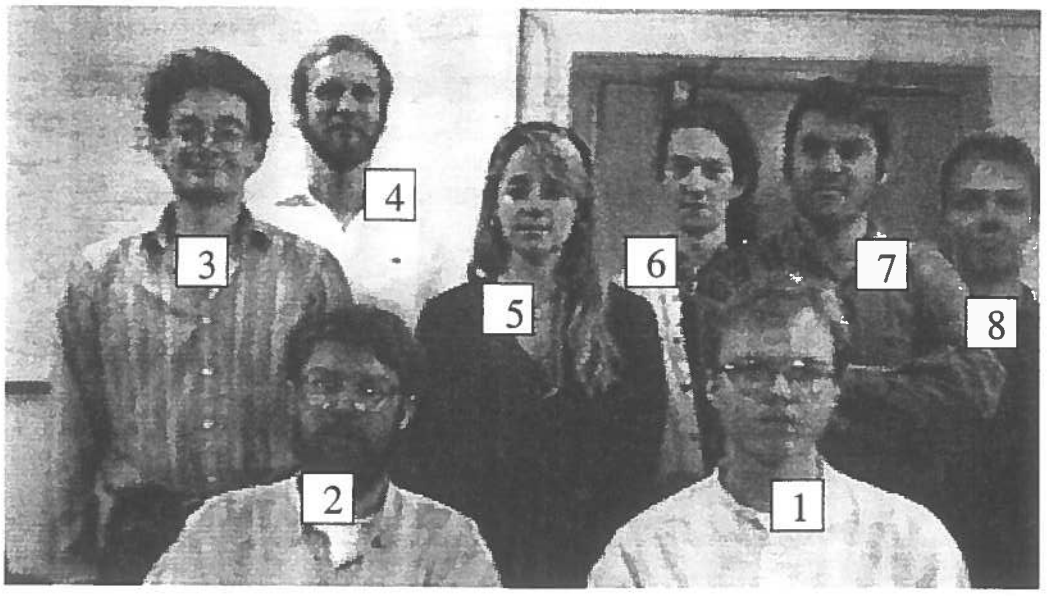

Figure 3: Generalization test image, whrein a model created from a 9^{th} face (see Figure 4) was matched against the faces seen here. Other models were also tried.

similarity between the face model and those in the image. The scores for the other two models are low, although the cat model does score more highly than the wedge model, as a cat face does share some of the same features with a human face.

Face: Figure 5 shows the path taken over 9 saccades on a human face using models for the face, eye, nose and mouth. The nose model was also found, but it was suppressed as it was found at the same time as the full face, only with a lower match score. As can be seen, the saccade path (which starts near the center of the image and then goes to the left nostril) is quite reasonable in going to features likely to be of interest. This matching run took 95 seconds on a SUN 200 Mhz Ultra (but we have also experimented

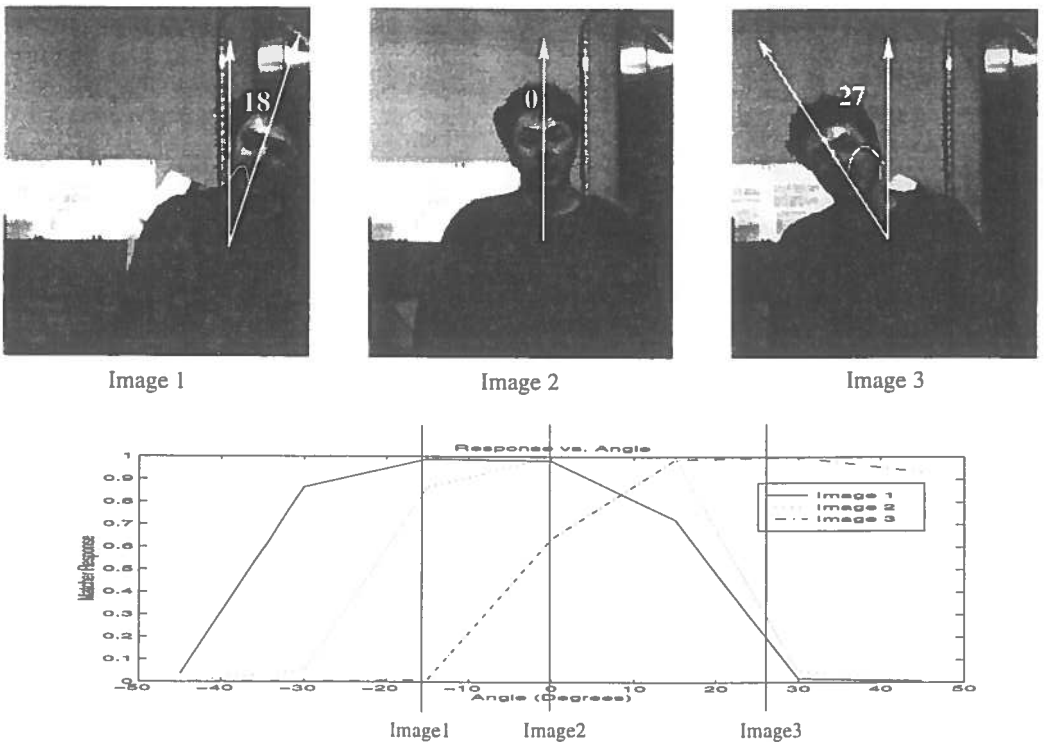

Figure 4: Rotation Invariance Test

with distributing the microsaccades to other workstations using MPI and achieved nearly linear speedup for up to 9 processors).

Cat: The matches found (using an eye, nose and face model) after 20 saccades are shown in Figure 5. Note that the system also matches the nose to part of the cat's fur. The left eye is not immediately found because finding the face suppresses the interest within a large enough area to suppress the region containing the eye until later in the run. In making 6 matches, the system foveates the right eye and face twice. One false match was found.

Ugaritic cuneiform script clay tablet: The tablet contains a large number of potential interest points. 90 saccades were made (before the

a. Cat: matching results

b. Tablet: matching results

C: Face: matching results

Figure 5: Matching results

matcher was manually stopped) with a model base containing two models - a vertical and a horizontal wedge. The matches are shown in Figure 5.b. The horizontal wedges are labeled 'H' and the vertical wedges 'V'. Of the 22 matches, 6 are mismatches. Half of these are due to the definition of the models, rather than any fault in the matching algorithm: when a horizontal wedge is next to a previously seen vertical wedge, the context alone is frequently enough to cause a match to the vertical wedge model, because the vertical wedge model is quite small.

6 Conclusions

This project has presented an investigation of an iconic vision system. In doing so, biologically inspired algorithms have been advanced for the mechanisms of feature extraction, visual attention and matching. Together these systems build up a stable representation of the world in which figure has

been separated from irrelevant background.

Although this project has many further areas of research, the overall approach (multi-variate iconic matching) works. To achieve the results, we use simple, easily extractable image representations, with large numbers of independent features in the input representation, coupled with a simple matching process that takes account of the features spatial distribution. The matching algorithm takes advantage of the fact (noted in [8]) that as the number of dimensions in an input space increases, the chance of two vectors being orthogonal (and therefore trivially separable) improves.

Most of the algorithms used here (local neighborhood feature extraction or shift correlation) have a simple, easily parallelizable, matching algorithm implementable using local fan-out network connectivity.

References

[1] MJ Farah. *Visual agnosia: disorders of object recognition and what they tell us about normal vision.* MIT Press, 1990.

[2] JA Feldman. Four frames suffice: a provisional model of vision and space. Behavioral Brain Sciences Vol 8, 265-313, 1985.

[3] RB Fisher and P Oliver. Multi-variate cross-correlation and image matching. In *Proc. British Machine Vison Conf.*, Birmingham, September 1995.

[4] TD Grove. Attention directed iconic object matching. M.Sc. thesis, Dept. of Artificial Intelligence, University of Edinburgh, 1995.

[5] TD Grove, RB Fisher. Attention in Iconic Object Matching. Proc. British Machine Vision Conference BMVC96, Edinburgh, pp 293–302, September 1996.

[6] J Howell and H Buxton. Invariance in radial basis function neural networks in human face classification. Technical Report CSRP 365, University of Sussex, 1995.

[7] D Marr. *Vision.* W.H. Freeman and Company, 1980.

[8] RPN Rao and DH Ballard. Object indexing using an iconic sparse distributed memory. Technical Report TR 559, Computer Science Dept., U. Rochester, 1995.

[9] J Malik and P Perona. Finding Boundaries in Images. in (ed. H. Wechsler), Neural Networks for Perception, Vol 1: Human and Machine Perception, Academic Press, Ch II.7, pp 315–344, 1992.

[10] G Sandini and M Tistarelli. Vision and Space-Variant Sensing. in (ed. H. Wechsler), Neural Networks for Perception, Vol 1: Human and Machine Perception, Academic Press, Ch II.9, pp 398–425, 1992.

[11] B Schiele and JL Crowley. Object Recognition Using Multidimensional Receptive Field Histograms. in Proc. 1996 Eur. Conf. on Comp. Vision, Vol 2, pp 610–619, 1996.

[12] M Seibert and AM Waxman. Learning and Recognizing 3D Objects from Multiple Views in a Neural System. in (ed. H. Wechsler), Neural Networks for Perception, Vol 1: Human and Machine Perception, Academic Press, Ch II.12, pp 426–444, 1992.

[13] JP Siebert and I Eising. Scale-space recognition based on the retino-cortical transform. In Proc. !EE Conf on Image Processing and its Applications, Edinburgh, 1995.

[14] J van der Spiegel, G Kreider, C Claeys, I Debusschere, G Sandini, P Dario, F Fantini, P Belluti, and G Soncini. A foveated retina-like sensor using CCD technology. In C. Mead and M. Ismail, editors, *Ana-*

log VLSI implementation of neural systems, chapter 8, pages 189–212. Kluwer Academic Publishers, Boston, 1989.