Feature Aggregation in Iconic Model Evaluation

Kay Brisdon, G.D. Sullivan & K.D. Baker

Intelligent Systems Group, Department of Computer Science,
University of Reading, RG6 2AX, UK.
Kay.Brisdon@reading.ac.uk

This paper presents recent work on iconic model-matching. The idea of iconic feature evaluation is reviewed, and methods for setting adaptive noise thresholds for use in feature combination are described. Extensions to the adaptive thresholding technique are explained and illustrated, and the relevance of this technique to feature combination is discussed. Finally demonstrations of the performance of the system are shown, with particular reference to the discrimination ability of the method with multiple models.

This paper describes a method of model-matching, applicable as a verification procedure within a knowledge-based vision systems containing three-dimensional geometric models. Most approaches to object verification in model-based vision merely extend the initial model instantiation process which uses a symbolic edge description, and thereby refine the initial hypothesis until a solution is reached. Symbolic edge descriptions are always inaccurate since it is difficult to translate real world scenes in to a set of discrete entities[1]. The iconic approach returns to the original image and can thus make use of information missed by the low-level processes.

If a three-dimensional object is to be represented as a geometric model the model description should be analogical, to reflect the spatial isomorphism between the two entities. Geometric models used in vision systems are usually converted into an entirely symbolic form, for example graph structures[2], or bit strings[3], to facilitate matching with a symbolic image segmentation. A preferable approach is to use the spatial isomorphism present in the model and match it to an iconic representation of the image. Thus instead of simply applying global operations to the image to produce a fixed set of data structures, which can only be used uniformly, computational procedures of arbitrary complexity may be devised to manipulate the information in the image. Reliance for the final classification on the output of region or edge segmentations then becomes unnecessary.

ICONIC MODEL-MATCHING

There are a number of methods available with which iconic model-matching to an image can be performed. One example is image rendering, which was discussed in particular by Besl and Jain[4]. They believed that a common fault of vision systems is that "high-level results are not projected back into a low-level form for final error checking" and said that "more research is needed in this area". Their proposed solution to the problem was to use computer graphics techniques in conjunction with the object models to predict the sensor data. This prediction could then be tested for its correspondence with the pixels in the image. On a sequential machine this is a very slow process and excessively sensitive to minor detail. In any real image, and particularly in natural scene analysis, it is impossible to predict the conditions in the image.

Another similar iconic/iconic process, namely normalised correlation, suffers from the same problems. In this technique template functions are produced which specify the expected binary or grey-level distribution of the image. The image is convolved with these masks and a metric used to measure the "best" or "sufficiently good" matches in the image. Correlation is more flexible than image rendering, as it is less dependent on local properties, and in addition deformations of the image are allowed. It still however relies on specifying what a portion of the image will look like and performing a quantative match. This is very difficult, and requires a much deeper knowledge of the conditions in the image than a qualitative matching process.

Knowledge-based iconic matching procedures used by model-based vision systems are uncommon. Bolles and Cain[5] used a very simple procedure as the final verification check in their vision system. They aligned a two-dimensional object's outline with a hypothesised instance in a binary image and matched the black-to-white transitions in the binary image. Since they were predicting from a model they could adapt the procedure only to take account of directly confirming evidence from the model. A transition across the boundary of the object from the object colour (black) to the background colour (white) was taken as positive evidence. A black object colour with no transition to white background

AVC 1988 doi:10.5244/C.2.4

was taken as neutral evidence. Anything else was negative evidence.

A more sophisticated iconic matching procedure has been described previously[6]. Viewpoint dependent predictions were made about the two-dimensional features in the image, hypothesised from a three-dimensional model. Additional information was associated with each feature to make stronger predictions, thus allowing a more selective test process. Individual feature evaluations were performed by adapting a typical data-driven edge detection process to work in a predictive manner. This gives similar information to unthresholded, unsegmented edgelet output produced in the first step of most edge detectors. Since the effort is focussed on the predicted areas rather than being applied globally it is much faster, as well as not merely being confined to step edges. The important task is then to combine these pieces of information meaningfully.

A very simple method of combining individual feature evaluations into a model matching procedure has already been described. The process was used for object verification in the Exemplar task of the MMI-007 Alvey consortium[7]. This paper expands on the method of thresholding outlined previously, and describes further work on feature evaluation combination.

DERIVING ADAPTIVE NOISE THRESHOLDS

The Initial Approach

All classification processes employ some kind of threshold. The decision about the level at which this should be set is very difficult, and is often rather arbitrary. As Marr says "It is a matter of unhappy experience that whenever we have to set a threshold in an image-processing task, we usually have problems .."[8]. This problem arises very commonly in edge detection, where a threshold normally decides which "edge measures", for example gradient slopes, are sufficiently significant to indicate the presence of an edge, rather than the presence of noise.

The feature evaluation process which we are considering takes the predicted end-points of a vector in a Gaussian blurred image and differentiates parallel to the feature at constant intervals along the vector. The average magnitude of the maximum/minimum gives a measure of the edge-ness of the feature, the gradient of the zero-crossings the bar-ness. The problem is to find at what level these measures are significant, and how they can be related across different types of features, different frequency filters and separate images, so an individual feature evaluation can be put in context with any other.

The first step in finding a solution was to find the expected response of any "type" of feature evaluation to noise. The result obtained from evaluating a predicted line on a template could then be weighted using this value. So, after such a normalisation factor has been applied, a value of one would signify that the feature is equal to the expected noise value. A value of some degree greater than one (the degree depending on how superior to the noise value the magnitude of the feature is) would imply that a feature had been found. A fractional value would be obtained if the feature strength is poorer than the noise response.

To achieve this it was necessary to discover which parameters of the evaluation affected the expected noise response value, and what simple measurements characterised these parameters. Suitable inexpensive measurements must be found, since each new image requires its own noise thresholds. This is of course desirable, because the thresholds are then automatically adapted to suit that particular image.

Initially an experimental investigation was conducted. A Monte Carlo simulation approach was taken. Randomly placed feature evaluations were performed on a sample set of natural scene images. Some plots of the expected scores obtained are shown below in figures 1 & 2 for the edge operator.

For an edge feature the mean of the noise samples is independent of length and, to a close approximation, independent of the image search area perpendicular to the feature. Bars are also independent of length, but the search area parameter, since it controls what sized bars can be discovered, cannot be deemed constant. The other parameters are the filter size (the sigma of the Gaussian) and the image itself.

Since the measure denoting a feature is an average of the evidence found along it, it is unsurprising that the expected noise value is approximately constant with length, although it is to be expected that the

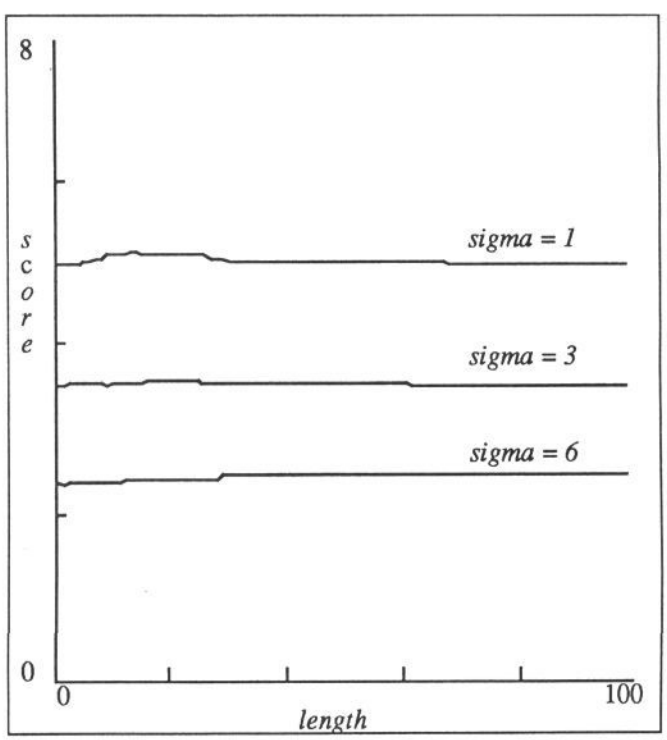


Figure 1. Expected scores for edges in a sample image at constant width.

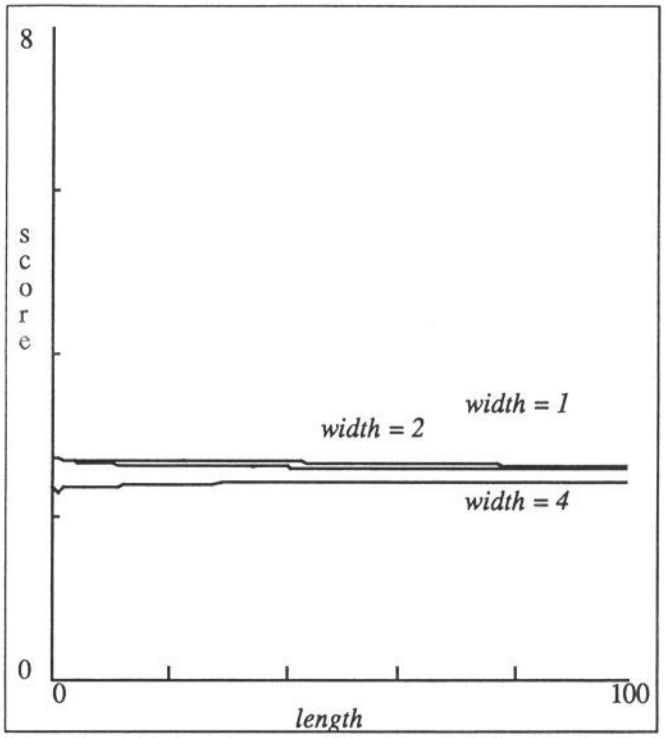


Figure 2. Expected scores for edges in a sample image at constant sigma

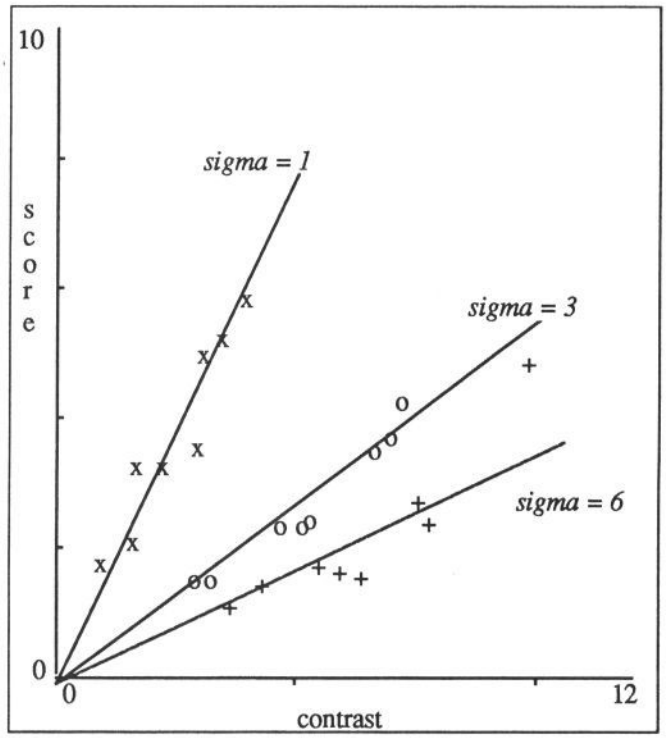


Figure 3. A graph of expected scores/contrast for a variety of outdoor scene images.

distribution of scores of randomly sampled short features would vary from that of long. For a long feature more values are used to obtain a single "averaged" score, so the range of scores is likely to be smaller. Since assuming that the expected noise response is constant with length is a gross simplification of the problem, this avenue was explored initially.

First it was necessary to determine a relationship between the changes in threshold level obtained in the various images. It was assumed that this level was related to some measure of the "contrast" in the image. Since the evaluator is in reality a second difference operation (the difference is taken between the maximum positive and negative responses in the first difference) contrast measures in second difference functions were explored. The best results were obtained by using the standard deviation in a difference of Gaussian (DOG) filtered image. Different frequency DOGs were used to match the different frequency Gaussians. As can be seen from figure 3 there was a simple linear relationship showing that the noise response was directly proportional to the contrast measure.

Extending the Approach

The model-matching results described previously[6,7] used the above method of thresholding feature scores. However the simplification which assumed that one threshold was applicable to features from a few pixels long to several tens of pixels proved to be too restricting. It is possible to get round this problem by the usual methods employed in model-matching schemes, namely to introduce a notion of saliency and weight the scores twice, firstly by their strength (our noise threshold) and then secondly by some metric which attaches more importance to larger features. However such a scheme is not in accord with the approach we are pursuing. We would like the proportional weighting for different length features to be determined automatically by the image, and the evaluation type. For this reason we started exploring the distributions produced by the Monte Carlo random noise trials, instead of looking merely at the mean values.

Figure 4 shows an example of the change in distributions with feature length. From log plots it can be seen that the functions consist of two exponentials. From the manner in which these two distributions vary between images of different composition it seems reasonable to hypothesise that the two exponential functions arise separately, one from the amount of random noise in the image, the other from the amount of structure.

At present we have no way of generating these two components from simple measurements in the image, in the manner described previously. Instead we are investigating generating the distributions of arbitrary length feature tests from that for vectors of length 1 pixel, since sampling vectors of length 1 in the image is a sufficiently simple process.

Once these distributions have been produced for a sample set of feature lengths and feature scores for an image, the tails of the functions can be used to provide probabilities of a feature of a certain length producing a score at least as good as the one recorded. This gives us a number of probabilities, p_1 .... p_k, one from each feature tested. Given that our initial hypothesis is that each piece of evidence occurred at a random point, at an insignificant level, then each test is independent. A standard method of combining probabilities arising from

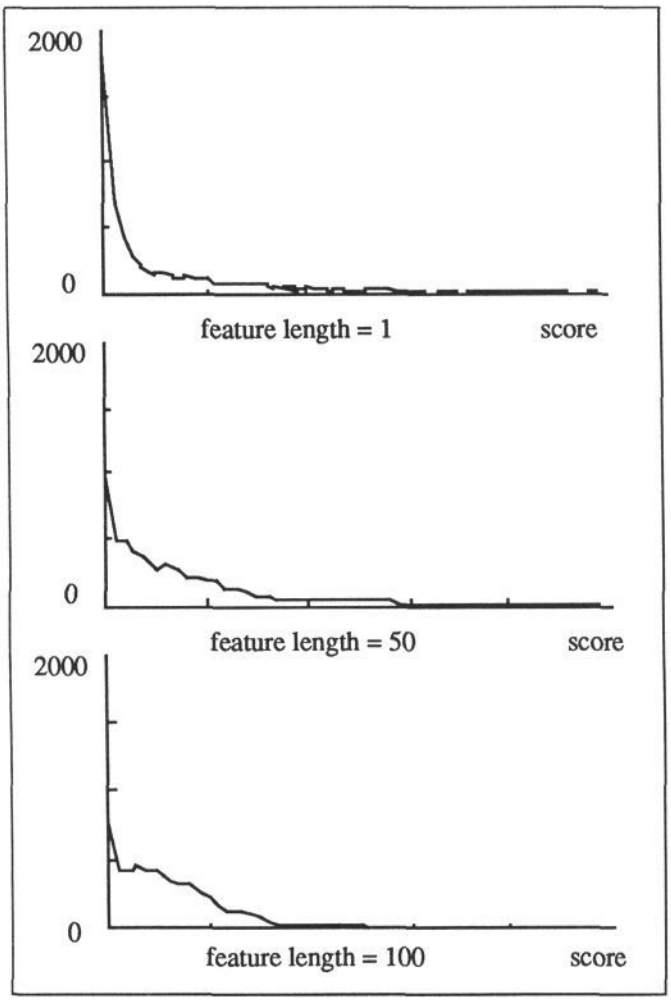


Figure 4. Three graphs showing the distributions of evaluation scores when edge features of different lengths are sampled randomly in an image.

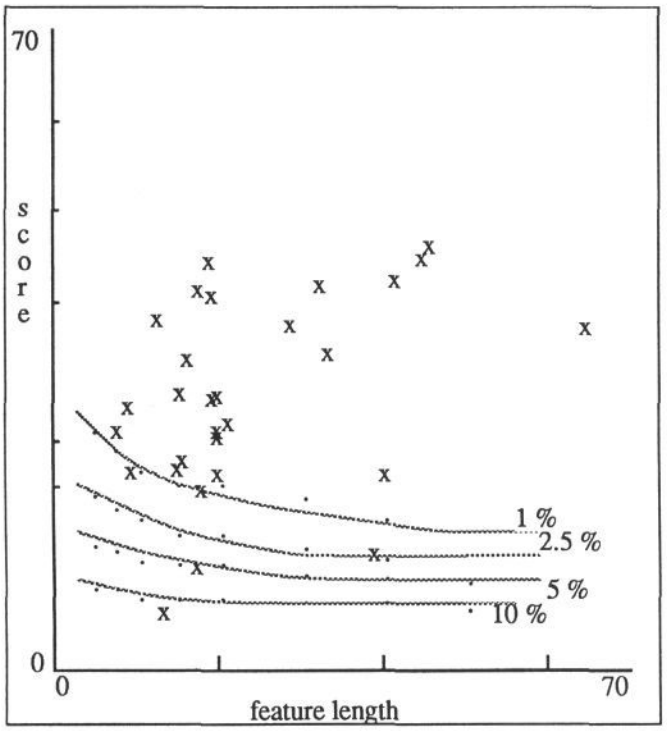


Figure 5. A graph showing score/length for each feature evaluation on a correctly positioned car template. Significant chi-squared distribution points are shown.

independent tests on different groups of data, so as to assess the probability of the overall test, is to use the statistic :

$$-2 \sum \ln p_i, \quad \text{for } i=1..k,$$

which has a chi-squared distribution with 2k degrees of freedom. Standard tables exist to calculate the significance of a value resulting from a chi-squared distribution. See figure 5 for an example of a set of feature evaluation scores for a template positioned on a car, shown with a series of significance levels of the chi-squared distribution.

Our threshold measure is similar to the "reliability" statistic that Goad uses. His measure is of the probability that the edges he has matched to could have been discovered to that degree, given that they arose from noise. However his overall statistics need to be more complicated than ours because he also has to take in to account the "plausability" of the edge detector output[3].

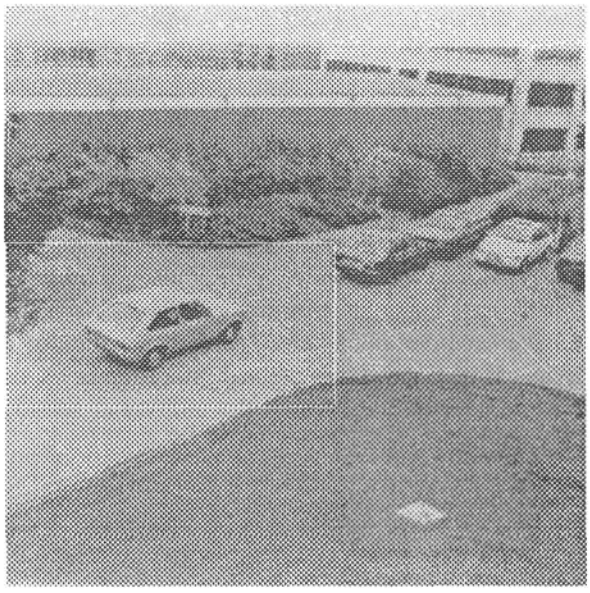
The kind of thresholding we have described above is image specific. Some of the implications of using this type of thresholding will be discussed below.

ISSUES OF FEATURE AGGREGATION

All the noise measurements and thresholds discussed above have been global. An interesting extension to the method is to adapt the measures to be more specific; to be either orientation dependent, or sensitive to only the immediate surroundings of the car.

To set orientation specific noise thresholds is particularly useful in our example of matching car templates. On any model different sorts of features occur with varying frequencies. On a car most of the features project to either horizontal or vertical lines (assuming that the car is upright on a horizontal ground plane), but there are just a few obliquely angled features. Obviously if the image background consists largely of horizontal and vertical lines, as is quite common in built-up scenes, then finding these sloping features is of particular significance, and they should be weighted as more salient. If isotropic expectation measures are taken this would adapt the score of a feature in a particular direction according to the number of features in that direction in the image. It is similar to the effect that humans experience when, after being exposed to high contrast gratings, the eye has reduced sensitivity to features in that same orientation at the same spatial frequency[9].

A second area of investigation is local-area threshold adaption. The contrast measures discussed above are taken across the whole image. This has the effect that the evaluator finds it more difficult to "see" a car in a cluttered background. Intuitively this sounds right, if the cluttered area is around the car. An alternative is to determine the noise thresholds from contrast measures taken from only a restricted portion of the image, centred on the template. Then if the car is parked in a



score for template when threshold is adapted to local area is 1.6 times score in full image.

Figure 6. When thresholds are determined from the less cluttered area surrounding the car (indicated on image) then the matching score of the same template is increased.

densely populate car park, taking the local area contrast reduces the likelihood of discovering the car, however the reverse is true if the car is reasonably isolated. See the example in figure 6.

In model-matching schemes it is common to use some notion of hierarchy. In our case we would wish to allow initial "gross" matches to give an indication of whether a more detailed study of that template is appropriate. A "salient features" evaluation should obviously be quicker than the full match, but should have the same general properties, ie degradation with noise, and should therefore allow any search process based on the iconic matcher to be more efficient.

The issue of saliency is closely related to the noise thresholding and feature aggregation method. If the features on the model are divided into sub-classes according to their direction in the image than the directional noise statistics automatically decide which features are more salient, due to the composition of the background in the image. The noise thresholds also determine which feature types are more important. The expectation values for bar features, for example, is much lower than the expectation values for edges. This is because a bar is a second order feature, and more difficult to find in the image by chance. Therefore the "score" for a correctly found bar of similar intensity to an edge is higher. Combining the features according to length-related probabilities also automatically sets the saliency of longer features higher than shorter features. Thus certain feature types and long features are pre-judged to be more salient and can always be use in a "salient features" evaluation. Others depend on the particular image, and can only be determined after the image has been assessed.

Normally the total set of features discovered from an edge detector have been discovered at a variety of scales. Combining features across separate scales is a very difficult and largely unsolved problem. Since our model-matching process is predictive the problem we have is not so great. Every feature has associated with it, either a feature "size", or "extent", as appropriate. For example a bar knows its size in pixels in the image. An edge knows the length of step, ie its "extent" before another feature is reached. From this information it is possible to calculate either the optimum frequency, or any allowable frequencies in which to evaluate a feature. So a feature can be constrained to produce a score only once, and our thresholding measures allow us to combine uniformly across different-sized Gaussians. Alternatively, a single frequency Gaussian can be used, and although some features will not score optimally, they will only be evaluated if they are visible at that frequency.

The issues above are all image dependent. Obviously there are also object dependent matters. For example one question that has not been addressed so far is the strength with which different features on a model occur in the image.

USE OF THE MODEL-MATCHING PROCESS WITHIN A MODEL-BASED VISION SYSTEM

An issue, not previously addressed, is that of setting a cut-off value for the acceptance or rejection of a model instance. Earlier work was with only one modelled object. It was assumed that the cueing process which triggered the model verification procedure had correctly picked an area in which there existed a target object. This assumption obviously will not be correct so we must have a criterion with which to reject the "best" match if it is not sufficiently good.

A number of initial model positions will be presented to the iconic matching technique. Small fragments of structure can generate very distant car templates. These can be ruled out immediately since the majority of features on such a template are only a few pixels long, making any detailed analysis inappropriate.

Although the feature combination technique is independent of the number of features tested, on more distant cars, many features on the car are either too short, or have too small an "extent" to be evaluated. In these cases an unreliable score can result since there is much evidence present, but few feature tests. It is therefore necessary to consider a model-match with a very small proportion of its features evaluated as unreliable.

In general templates thrown on the image at random locations have middle-order, insignificant probabilities. These probabilities increase as the template is thrown on

an area overlapping a car or on another highly structured object, so a higher probability suggests that an interesting area has been discovered, and the best instance in that area can then be explored further. If a model instance is generated, based on features of the car having been found and labelled, then these features are ignored by the model-matching process, thus the matcher is looking for further evidence for a car, other than that already found.

The final probaility from the chi-squared distribution for a correct match tends to be very high. For this reason we do not look-up these probability scores, but use the chi-squared distribution value, weighted by the degrees of freedom.

DISCRIMINATING BETWEEN MULTIPLE MODELS

An additional problem we are now faced with is multiple models. It is probable that the initial analysis will only cue "a target object" instead of a particular model, especially if the differences between the models is fairly small, for example a hatchback car versus an estate car. In this case the verification process is applied twice, using both models. The results should clearly disambiguate the two models, by indicating a much stronger fit for the correct instance. See the example below, in figure 7, which demonstrates the "cross-over" of the scores for the evaluation of a Chevette model and a Cavalier estate model when applied to a series of vehicles. Both models were initially fit to the same data, then separate search processes were applied using the different models. The relative scores are the best-fit solutions to each model.

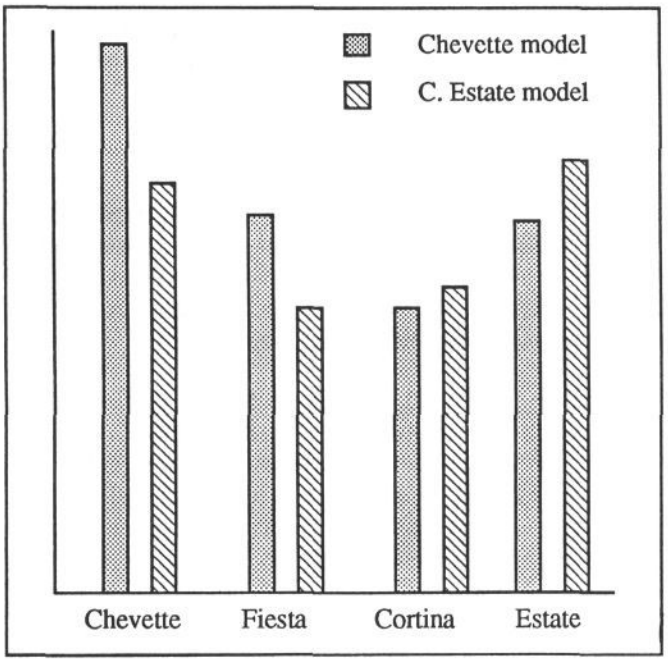


Figure 7. Shows the relative scores of the best-fit Chevette hatchback model and Cavalier estate model to a series of different vehicles.

CONCLUSIONS

The technique of iconic feature evaluation has proved both fast and effective. Using a model-based approach has allowed many specific tests to be made that improve greatly on the performance possible by data driven techniques. One of the most important issues for the technique is the combination of the individual iconic feature evaluations in to a model-based iconic matching process. The idea of adaptive thresholding works very well, and has many interesting extensions, some of which have been described here. The demonstration of the selectivity of the matching process to different models in the same general class proves the utility and sensitivity of this approach.

REFERENCES

1. **Fischler M.A.** 1978, "On the Representation of Natural Scenes" in *Computer Vision Systems* Eds A.R. Hanson & E.M. Riseman, Academic Press.

2. **Brooks R.A.** 1984, "*Model-Based Computer Vision*" UMI Research Press, Comp Sci : AI. No.14.

3. **Goad C.** 1987, "Special-Purpose Automatic Programming for 3-D Model-Based Vision" in *Readings in Computer Vision* Eds M.A. Fischler & O. Firschein, Morgan Kaufmann, 1987.

4. **Besl P.J. & Jain R.C.** 1985, "Three-Dimensional Object Recognition", *ACM Computing Surveys*, Vol 17, No.1, pp75-145.

5. **Bolles R.C. & Cain R.A.** 1982, "Recognising and Locating Partially Visible Objects : The Local Feature Focus Method", *Int Journal of Robotics Research*, Vol 1, No.3, pp57-82.

6. **Brisdon K.** 1987, "Alvey MMI-007 Vehicle Exemplar : Evaluation and Verification of Model Instances", *Proc. of Alvey Vision Club Conference*, Cambridge, England.

7. **Sullivan G.D.** 1987, "Alvey MMI-007 Vehicle Exemplar : Performance & Limitations", *Proc of Alvey Vision Club Conference*, Cambridge, England.

8. **Marr D.** 1982, "*Vision*" W.H. Freeman & Co

9. **Campbell F.W.C. & Robson J.** 1968, "Application of Fourier Analysis to the Visibility of Gratings" *Journal of Physiol* (Lond) No. 197, pp551-566