

Fusion Descriptors for Content-Based Sketch Retrieval - A Comparative Evaluation of Retrieval Performance

Daniel Heesch and Stefan Ruger

Department of Computing, Imperial College, London, UK

Abstract— The paper introduces three transformation-invariant shape descriptors and provides an evaluation of their relative strengths in the context of content-based retrieval of graphical sketches. We show that the use of a combination of different shape representations may lead to a significant improvement of retrieval performance. We provide evidence that there exist an optimal combination that proves robust across different data sets and queries.

I. INTRODUCTION

Over the previous ten years there has been a growing interest in content-based image retrieval, a trend that coincides and appears to be causally linked with the rise of the world-wide web and the spread of digital technologies for image creation. Research into CBIR has resulted in a number of commercial systems such as IBM’s QBIC, Excalibur Visual RetrievalWare and Virage VIR Image Engine to name but a few of the most prominent. The practical applications of content-based image retrieval are multifarious and range from medical diagnosis over remote sensing to journalism and home entertainment. A special area of application is in architecture and engineering where information is often in the form of graphical sketches. Sketches constitute a special type of image with a few characteristics that impinge on the kind of techniques employed for successful retrieval. First, in contrast to a fully-fledged image that often requires segmentation to separate local objects from their context, sketch objects are already well-defined since spatial context and background are typically trivial. Secondly, the object shape captures most of the semantic

information present in a sketch. With little, if any, semantic contribution from colour and texture, the performance of a sketch retrieval system is essentially determined by its ability to capture the shapes detectable in the sketch.

The issue of shape-based object recognition has long been the central interest in the area of computer vision. Traditionally, most research has aimed at finding ever more powerful shape descriptors [1]. Little work has been done, however, on how to combine different representations in an attempt to boost retrieval performance.

The number of shape representations developed over the past fifty years is bewildering. Each representation differs in applicability and complexity. The simplest representations are based on so called "shape factors", which are single values that capture but the minimal information of the object. These factors include, *inter alia*, the perimeter, the area, the radius, the elongation, the compactness, the number of corners, the circularity and the convexity of the object. The first three of these measures suffer from the fact that they vary with the object size and lose important information. Although the remaining measures provide scale invariance, they still do not provide sufficient information for reliable object recognition and retrieval when used in isolation [2]. An early attempt to achieve a boundary representation that can be used for recognition purposes is the chain coding method proposed by Freeman ([3], [4]). A chain code approximates a curve with a sequence of directional vectors lying on a square grid. The technique is computationally efficient but suffers from digitisation noise. A

well-established technique to describe closed planar curves is the use of Fourier descriptors (see for example [5] and [6]) which describe a curve with coefficients derived from a Fourier analysis of some parametric representation such as the curvature or spatial coordinates of the curve. By normalising the Fourier descriptors one can achieve representational invariance with respect to a variety of affine transformations. Normalized Fourier descriptors have successfully been used for recognition tasks (e.g. [7]). Among area-based representations, moments are the most notable. They can be made invariant and have variously been used for recognition and retrieval tasks ([8], [9] and [10] for aircraft recognition, [2] for trademark retrieval).

All of the above-mentioned representations have in common that they lend themselves very well for comparison as they can be stored as a simple vector. More elaborate shape representations have been introduced some of which are the curvature scale space representation or the spline curve approximation which require sophisticated shape matching techniques [11]. Even though they can provide very powerful shape representations, their computational costs make them ill-suited for the purpose of interactive retrieval of objects from large databases where performance becomes an issue.

For the present study we have selected three descriptors, namely Fourier descriptors, invariant moments and difference chain codes. The choice was based on three criteria: (i) suitability of the descriptor for large databases, (ii) efficiency of extraction from image and (iii) retrieval efficiency as documented in the literature. We describe how the three descriptors can be combined with one another in a retrieval model to produce retrieval results superior to any of the single-descriptor models.

With the increasing number of shape descriptors and the growing need for sketch retrieval, formal evaluation of retrieval performance becomes a major desideratum. Information retrieval has already developed a set of elaborate evaluation measures, among which *precision-against-recall graphs* and the derived measure of *average precision*

have found widespread use [12]. We use these measures to evaluate the retrieval performance under different feature regimes for a database of 238 black-and-white sketches (see Appendix I for example sketches).

II. METHODS

A. Shape descriptors

A.1 Chain Code Histograms

Freeman chain codes are derived by reading the contour in a consistently clock-wise or counter-clockwise direction starting at position 0 and noting for each pixel its relative position with respect to its neighbour. Specifically, a 0,2,4 or 6 is noted for right, top, left, bottom, respectively. The chain-code thus obtained is invariant with respect to translation and scaling, but not with respect to rotation and starting point. To achieve the latter, the chain code is further processed to derive a code for the change in direction d at any one pixel, which is encoded as $d = (C_{i+1} - C_i) \bmod 8$ where C_k denotes the chain code at pixel k . The resulting representation is a sequence of difference chain codes. In the last step the sequence is reduced to a 4-bin histogram resulting in a four-number representation of the contour that is invariant with respect to translation and starting points as well as, though to a lesser extent owing to digitisation noise, to rotation and scaling.

A.2 Moment invariants

Moments capture distributional properties of a random variable. The first moment is equivalent to the expected value, while the second central moment yields the variance of that variable. For 2-d black-and-white images of size $N \times M$, an n th order moment is defined as:

$$M_{pq} = \sum_{i=1}^{N \times M} x_i^p y_i^q f(x_i, y_i),$$

where $p, q \in \mathcal{N}$ with $p + q = n$ and $f(x_i, y_i) = 1$ if the i th pixel at position (x_i, y_i) is set and 0 otherwise. Translation invariance can be achieved by using central moments

$$\bar{M}_{pq} = \sum_{i=1}^{N \times M} (x_i - \bar{x})^p (y_i - \bar{y})^q f(x, y),$$

where $\bar{x} = M_{10}/M_{00}$ and $\bar{y} = M_{01}/M_{00}$. Further normalization renders the moments scale-invariant thus:

$$\mu_{pq} = \frac{\bar{M}_{pq}}{(M_{00})^\lambda},$$

where $\lambda = (p + q)/2 + 1$. Using only the second and third moments, one can finally derive the following seven moments that are invariant with respect to all three affine transformations [8]:

$$\begin{aligned} m_1 &= \mu_{20} + \mu_{02} \\ m_2 &= (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2 \\ m_3 &= (\mu_{30} - \mu_{02})^2 + (3\mu_{21} - \mu_{03})^2 \\ m_4 &= (\mu_{30} + \mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \\ m_5 &= (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - \\ &\quad 3(\mu_{21} + \mu_{03})^2] + (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) \\ &\quad [3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \\ m_6 &= (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{12} + \mu_{03})^2] + \\ &\quad 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \\ m_7 &= (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - \\ &\quad 3(\mu_{21} - \mu_{30})^2] + [(3\mu_{12} - \mu_{30})(\mu_{21} + \mu_{03})] \\ &\quad [3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] \end{aligned}$$

The above seven moments vary considerably as the last two moments are exceedingly small. To grant each moment invariant an equal weight, they need to be normalised such that they cover approximately the same range. In the approach taken in this study the moment values computed for the 238 sketches are first grouped around the origin by subtracting the mean. The resulting values are then divided by the standard deviation computed for the entire collection of images, i.e.,

$$m_i[v] = \frac{m_i[v] - \bar{m}_i}{SD_{m_i}}.$$

where \bar{m}_i and SD_{m_i} denote the mean and the standard deviation, respectively, of the i th moment invariant (taken over the sketch database). The majority of the resulting values will now range between -1 and 1. An alternative nor-

malization method would map for each moment its maximum value to 1 and its minimum to 0, with all other values coming to lie inbetween. This approach has the disadvantage of rendering the final distribution sensitive to individual outliers, whose effect can be reduced by dividing through a summary measure, such as the standard deviation.

A.3 Fourier Descriptors

For the computation of Fourier descriptors the contour pixels need to be represented as complex numbers $z = x + iy$ where x and y are the pixel coordinates. As the contour is closed we get a periodic function which can be expanded into a convergent Fourier series. Specifically, let Fourier Descriptor C_k be defined as the k th discrete Fourier transform coefficient with $-N/2 \leq k \leq N/2$ and we compute

$$C_k = \sum_{n=0}^{N-1} (z_n e^{\frac{-2\pi i kn}{N}})$$

from the sequence of complex numbers z_0, z_1, \dots, z_{N-1} where N is the number of contour points. To characterise boundary properties any constant number of these Fourier descriptors can be used. The most interesting descriptors are those of the lower frequencies as these tend to capture the general shape of the object.

To be of any use in object classification and object retrieval the descriptors thus obtained need to be made invariant with respect to rotation, the starting point of the contour, translation and scaling. The first two properties can be achieved by using only absolute values of the descriptors, while translational and rotational invariances are obtained by discarding the Fourier descriptor C_0 and dividing the other descriptors by $|C_1|$, respectively. The final feature vector v has the following form:

$$x = \left[\frac{|C_{-L}|}{|C_1|}, \dots, \frac{|C_{-1}|}{|C_1|}, \frac{|C_2|}{|C_1|}, \dots, \frac{|C_L|}{|C_1|} \right]^T$$

where L is an arbitrary constant between 2 and $\frac{N}{2} - 1$. In this study L is 10 resulting in 19 descriptors per object.

B. Evaluation

Evaluation of retrieval systems is a critical part in the process of continuously improving the existing techniques. While text information retrieval has long been using a sophisticated set of tools for user-based evaluation, this does not as yet apply to image retrieval. Only the minority of systems provide evaluation that goes beyond simple common-sense judgments. And yet, when applied with care, evaluation techniques from information retrieval can profitably be applied to image retrieval as well. For the present study we borrow the two well-known measures *recall* (the proportion of relevant images which are retrieved) and *precision* (the proportion of retrieved images which are relevant). Owing to the potential polysemy of images, the use of these measures in the context of image retrieval is not free from difficulties. The problem of polysemy can be resolved, however, if the image database relies on the strong semantics provided by a label or other textual descriptions [13]. In the present study all 238 images have been assigned to 34 categories. Each of these images corresponds to only one category (see Appendix I for some example categories). Removing semantic ambiguities, we obtain an image which is regarded as relevant with respect to a query if both share the same category.

The two measures precision and recall can be combined to form an integrated measure referred to as a *precision-against-recall* graph. This graph is obtained by, firstly, ranking all items of the database with respect to a query. Then, the precision and recall values are determined for the sequence of relevant items and the values interpolated as described in [12]. Each query results in one characteristic precision-against-recall graph and an average measure capturing the performance for the entire database can be obtained by averaging the precision-against-recall graphs over all queries. From the final graph, a more concise measure of performance can be derived in the form of *average precision* which can be thought of the area under the *precision-against-recall* curve. A high value indicates a good trade-off between high precision and high recall.

C. Similarity functions

Each shape descriptor reduces the image information to a vector of real numbers. For any one descriptor the similarity between two images is then computed as the similarity between the two vectors. Various similarity functions have been employed to map two vectors into a single-value measure of similarity. Of all, the most commonly used measures are the Euclidean distance measure, the cosine similarity measure and the intersection distance measure. Experiments performed on the sketch database (the results are omitted in this paper) have suggested the use of the cosine similarity measure for all three descriptors. It is defined as

$$\text{Sim}(u, v) = \frac{\sum_i (u_i \cdot v_i)}{|u| \cdot |v|}.$$

D. Fusion of shape descriptors

Descriptors are combined in an integrated retrieval model such that the overall similarity between two sketches is given by a convex combination of the similarity values calculated for each descriptor as below,

$$S(Q, T) = \sum_{f=1}^F w_f \text{Sim}_f(Q_f, T_f),$$

where $\text{Sim}_f(Q_f, T_f)$ denotes the similarity between two sketches Q and T with respect to the f th descriptor using a descriptor-specific similarity function and weighted by a factor w_f with $0 \leq w_f \leq 1$ and $\sum_f w_f = 1$.

E. Finding the optimum combination of weights

Using only three descriptors together with a medium-sized database, it was possible to find the optimal combination by raster scanning of the two-dimensional effective search space with a resolution of 0.01.

F. Comparison between shape descriptors

One of the aims of our investigation is empirically testing the hypothesis that some combination of the descriptors can outperform the most successful single-descriptor model in retrieval performance. In order to examine this hypothesis, we performed a paired t -test on the 238 aver-

age precision values computed for each of the models to be compared.

G. Robustness of the optimum

The optimal combination of descriptors is achieved when a particular combination of weights for which average precision averaged over all 238 queries reaches a maximum. It is natural to ask whether this optimum varies with the data set or whether it proves to be a generic feature that is largely unaffected by the type of objects present in the database. In the latter case the optimal combination of weights for the entire data set may be taken as an estimate of the optimal combination for an unknown dataset.

To evaluate the robustness of the optimal combination, the optimum is determined not only for the entire database but, in addition, for a sample of 12 smaller subsets. Each subset contains all the sketches from 20% of the categories selected randomly from the total of 34 categories. The inverse of the scatter of the optimal combination can be regarded as a reasonable measure of the robustness.

III. RESULTS

A. Single-descriptor models

For each of the three descriptors, retrieval performance was evaluated using the method described in II-B. Fourier descriptors prove most successful for the given retrieval task followed by moment invariants and difference chain codes. Despite the considerable standard deviation, the difference between the performance of Fourier descriptors and those of the other two descriptors is significant at $\alpha \leq 0.05$ ($p < 0.001$). The precision-against-recall graphs for the three models are depicted in Figure 1.

B. Multi-descriptor model and Robustness

When the three descriptors are combined as described in II-D, overall performance can be plotted in a three-dimensional space against the weights of two descriptors (note that the third is determined by the first two). Using an equilateral triangle *all* three weights can be represented

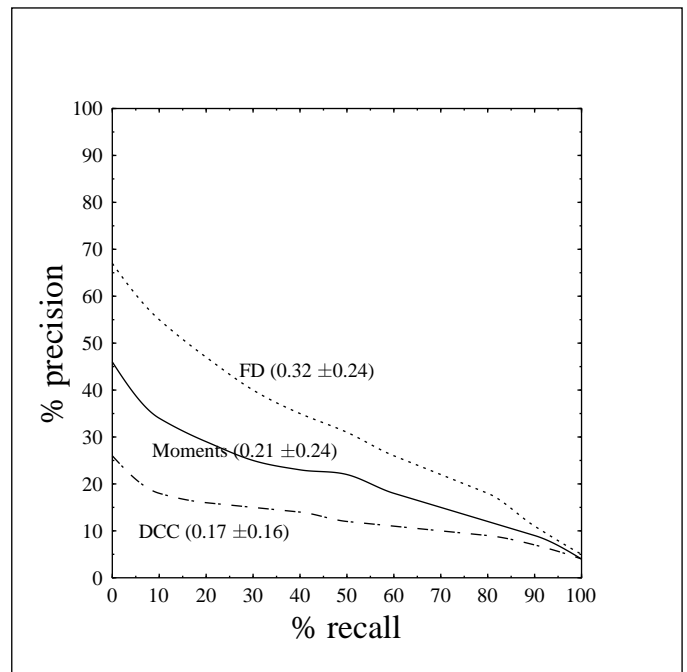


Fig. 1. Precision-against-recall graphs for the three single-descriptor models (FD = Fourier descriptors, DCC = difference chain codes). The differences in performance are significant at $\alpha \leq 0.05$ level ($p < 0.001$). The standard deviation is given in brackets.

as axes in two dimensions. For any given point within the triangle, the contribution of a particular descriptor is given as the distance between the point and the respective side of the triangle. Since the sum of the perpendiculars from any point onto the three sides is a constant, the condition $\sum_f w_f = 1$ is satisfied.

Figure 2 shows the average precision values for all possible combinations of the three weights. Note that the performance of the single-value models is found at the respective corners of the triangle (where two of the three weights are zero).

The large circle near the base of the moment invariants represents the combination of descriptor weights for which average precision reaches the global maximum. The optimal weights are 0.63 for Fourier descriptors, 0.35 for difference chain code histograms and 0.02 for moment invariants. The precision-against-recall graph corresponding to this optimal combination of weights is depicted in Figure 3.

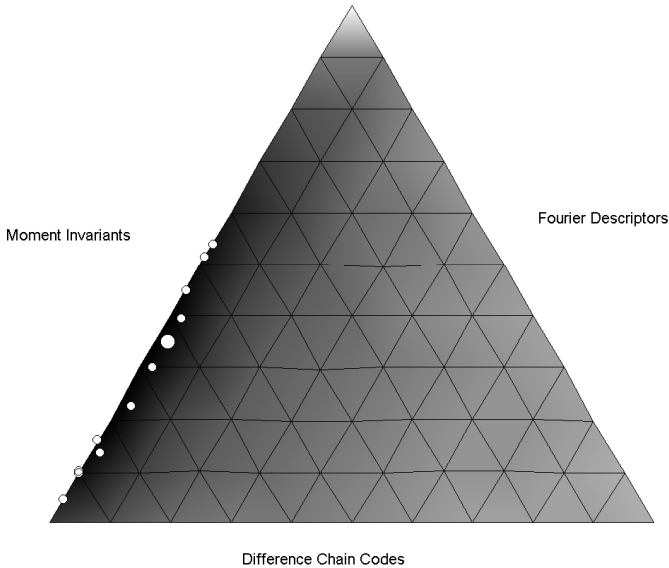


Fig. 2. Average precision values plotted for all possible combinations of the three descriptors. The values are derived by averaging the average precision values over all 238 queries.

A paired t-test was performed on the average precision data of the optimal model and the best single-descriptor model (Fourier descriptors). The difference in performance between the two models is highly significant ($p < 0.001$).

To answer the question to what extent the optimum is independent of the query set, we determined the optimum weight sets for a number of subsets. The optima are plotted in Figure 2 (smaller circles). As can be seen, the scatter defines an elongated region close to and along the base of the moment invariants. Fourier descriptors vary between 0.95 and 0.45 and hence there exists query sets for which difference chain codes outperform Fourier descriptors. The consistently low weights assigned to moment invariants suggests that their weakness in a multi-descriptor model is of a fairly generic nature.

These results are corroborated by additional experiments (not shown) in which a fourth descriptor (co-occurrence matrix [14]) has been added to the previous three. When used in isolation the descriptor has a marginal advantage over difference chain codes. When used in conjunction with the other three descriptors, it does not upset the optimum weight set, i.e., moment invariants remain negligible with

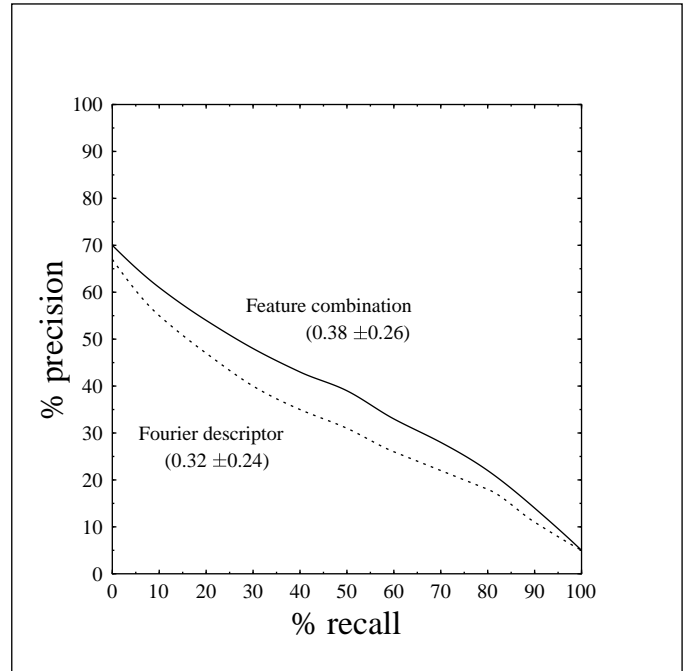


Fig. 3. The precision-against-recall graph for the multiple-descriptor model using the optimal combination of weights: 0.63 (Fourier descriptors), 0.35 (Difference chain codes) and 0.02 (Moments). Performance has improved significantly ($p < 0.001$) compared to the best single-descriptor model (Fourier descriptor)

the weights for Fourier descriptors and difference chain codes adding up to one.

IV. DISCUSSION

As can be seen in Figure 1 and 3 the standard deviation of the average precision values are considerable which implies that retrieval performance varies quite substantially with the query. This is to be expected since the image collection used in the present study shows a high degree of heterogeneity. In addition, categories range in size from as few as 3 to as many as 23 objects. As average precision values tend to increase with the size of the category the query belongs to, one must expect variation in category size to further inflate the standard deviation. It is interesting to note that, despite the large standard deviation, retrieval performances differ significantly between the three single-descriptor models as revealed in a paired t-test. This observation suggests that retrieval performance of different descriptors co-vary with the query, in other words, a "dif-

difficult" query will be difficult to handle by *any* descriptor. One reason for this co-variance is, once again, variation in category size: a query from a common category will tend to result in better performance, no matter which descriptor is used. In addition, some categories are quite "unique" while others are not. Members of the "seastars" category, for example, are very different from all other database sketches. By contrast, there are a number of categories that are very close in feature space to other categories (e.g. the categories "cars" and "fishes", "vans" and "trucks", "cats" and "dogs" etc.). Retrieval of these categories is generally more difficult.

One of the most notable findings of our experiments is arguably the non-additive, synergistic interaction between descriptors in multiple-descriptor models. Although moment invariants perform significantly better than difference chain codes when considered in isolation, it is only the latter that improves performance beyond the level of the best single-descriptor model when the three descriptors are allowed to combine with one another. One may want to argue that both Fourier descriptors and moment invariants capture similar aspects of the shape and will therefore offset with each other rather than complement each other. Likewise, difference chain codes and Fourier descriptors are probably sufficiently orthogonal to each other in order to allow for some degree of complementation. Also, as their name suggests, the strength of moment invariants lies in their invariance with respect to all three affine transformations. Because they are based on the area rather than the contour of an object, moment invariants are less affected by digitisation noise than, for example, difference chain codes. The sketches within each category are presently of similar scale and are aligned in a consistent manner, so that no reward is given to descriptors that are highly invariant. This may help to explain the performance of moment invariants in the single and multiple-descriptor models.

APPENDIX

I. EXAMPLE CATEGORIES

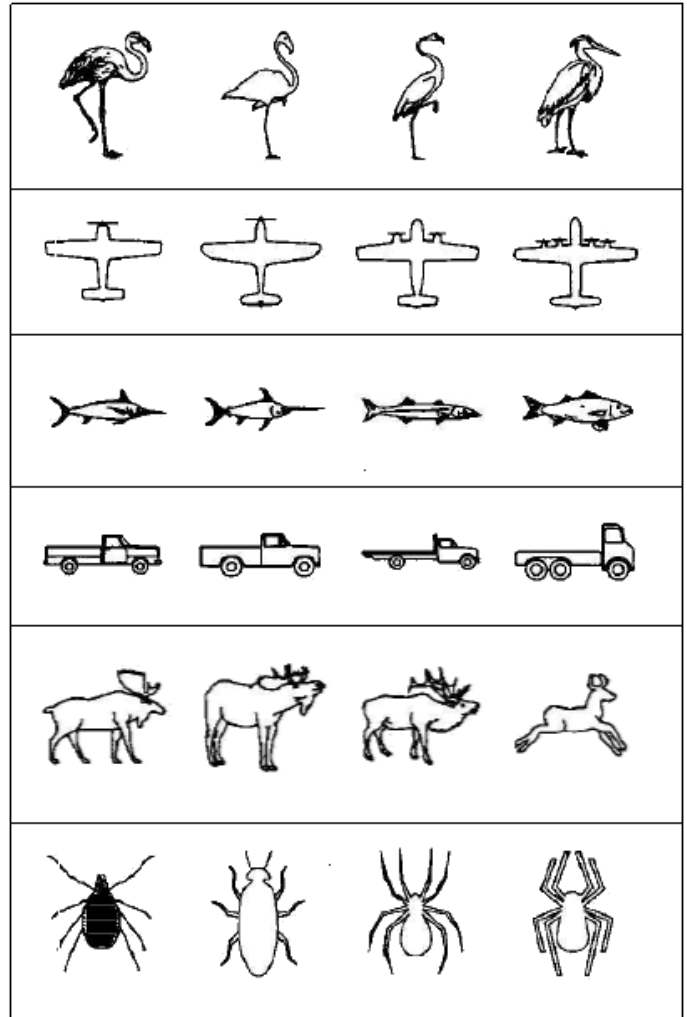


Fig. 4. Example sketches for six of the 38 categories present in the database. Other categories include, *inter alia*, rockets, airbusses, seastars, deers and various other types of animals.

II. EXAMPLE RUN OF THE SYSTEM

Below are the explicit results for a particular retrieval problem for each of the single-descriptor models and the multiple-descriptor model. The query is the sketch as shown in Figure 5.



Fig. 5. The query

Segmentation of the image and contour tracing of each of the objects found results in a collection of contours as shown in Figure 6.

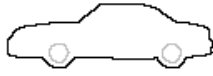


Fig. 6. Contours of the objects found by the segmentation algorithm. The black contour is the one that has been activated by the user.

The user can choose one or more contours by double-clicking on them. In this case, the user chose the outer contour of the car and ignores the hub-caps.

The system output for the four models is shown in Figures 7 to 10. The ten most similar images are arranged such that the distance to the center reflects the distance or dissimilarity between that image and the query.

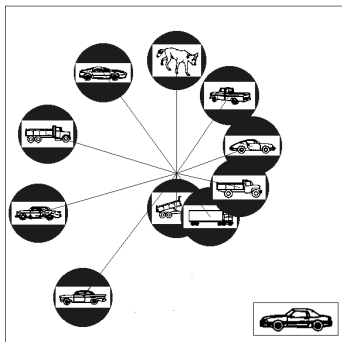


Fig. 7. Retrieval results for moment invariants.

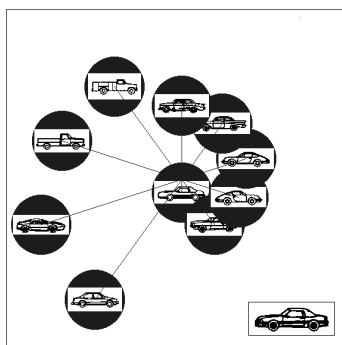


Fig. 8. Retrieval results for Fourier descriptors.

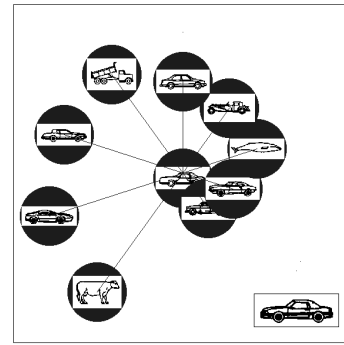


Fig. 9. Retrieval results for difference chain codes.

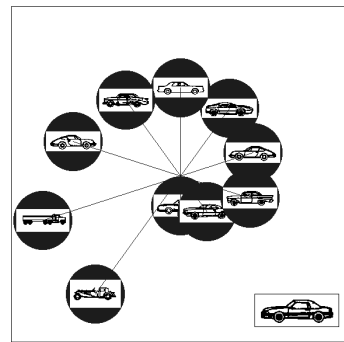


Fig. 10. Retrieval results for the optimal combination of Fourier descriptors and difference chain codes.

REFERENCES

- [1] T S Rui, T S Huang, M Ortega, and S Mehrota, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Trans. Circuits and Systems for Video Technology*, pp. 123–131, 1998.
- [2] B M Mehre, "Shape measures for content-based image retrieval: a comparison," *Information Processing and Management*, vol. 33, no. 3, pp. 319–337, 1997.
- [3] H Freeman, "On the encoding of arbitrary geometric configurations," *IEEE Trans. Electron. Comput.*, vol. 10, no. 2, pp. 260–268, 1961.
- [4] H Freeman and L S Davis, "A corner finding algorithm for chain coded curves," *IEEE Transactions on Computers*, vol. 26, pp. 297–303, 1977.
- [5] C T Zahn and R Z Roskies, "Fourier descriptors for plane closed curves," *IEEE Trans. Comput.*, vol. C-21, no. 3, pp. 269–281, 1972.
- [6] E Peerson and K S Fu, "Shape discrimination using fourier descriptor," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-7, no. 3, pp. 170–179, 1977.
- [7] T P Wallace and P A Wintz, "Fourier descriptors for plane closed curves," *Computer Graphics and Image Processing*, vol. 13, pp. 99–126, 1980.
- [8] M K Hu, "Visual pattern recognition by moment invariants."

- IRE Transactions on Information Theory*, vol. 8, pp. 179–187, 1962.
- [9] S Dudani, “Aircraft identification by moment invariants,” *IEEE Transactions on Computers*, vol. 26, pp. 39–45, 1977.
- [10] J Flusser and T Suk, “Pattern recognition by affine moment invariants,” *Pattern Recognition*, vol. 26, no. 1, pp. 167–174, 1993.
- [11] A del Bimbo and P Pala, “Visual image retrieval by elastic matching of user sketches,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 121–132, 1997.
- [12] E M Voorhees and D Harman, “Overview of the eighth text retrieval conference (trec-8),” in *Proc. TREC*, 1999, pp. 1–33 and A.17 – A.18.
- [13] J R Smith and C S Li, “Image retrieval evaluation,” in *Proc. Workshop Content-Based Access of Image and Video Libraries*, 1998, pp. 343–353.
- [14] A L Reno, *Object recognition by stochastic model adaptation and selection*, Ph.D. thesis, Imperial College, London, 1998.