

# A comparison of the effectiveness of alternative feature sets in shape retrieval of multi-component images

J P Eakins<sup>\*</sup>, J D Edwards<sup>\*</sup>, J Riley<sup>\*</sup> and P L Rosin<sup>†</sup>

<sup>\*</sup>Institute for Image Data Research, University of Northumbria at Newcastle, Newcastle upon Tyne NE1 8ST, United Kingdom

<sup>†</sup>Department of Computer Science, Cardiff University, Queen's Buildings, Newport Road, Cardiff CF24 3XF, U.K.

email: {john.eakins,jonathan.edwards,jonathan.riley}@unn.ac.uk;  
Paul.Rosin@cs.cf.ac.uk

## ABSTRACT

Many different kinds of feature have been used as the basis for shape retrieval from image databases. This paper investigates the relative effectiveness of several types of global shape feature, both singly and in combination. The features compared include well-established descriptors such as Fourier coefficients and moment invariants, as well as recently-proposed measures of triangularity, rectangularity and ellipticity. Experiments were conducted within the framework of the ARTISAN shape retrieval system, and retrieval effectiveness assessed on a database of over 10 000 images, using 24 queries and associated ground truth supplied by the UK Patent Office.

Our experiments revealed only minor differences in retrieval effectiveness between different measures, suggesting that a wide variety of shape feature combinations can provide adequate discriminating power for effective shape retrieval in multi-component image collections such as trademark registries. Marked differences between measures were observed for some individual queries, suggesting that there could be considerable scope for improving retrieval effectiveness by providing users with an improved framework (such as relevance feedback) for searching multi-dimensional feature space.

**Keywords:** trademark image retrieval, shape features, retrieval effectiveness, comparative evaluation

## 1. INTRODUCTION

Retrieval of images from a database on the basis of shape similarity is one of the most challenging problems currently facing researchers within the field of CBIR. Shape similarity matching techniques are of importance both in their own right (for applications such as fingerprint identification, trademark image registration, and engineering design retrieval), and as components of more general object identification and image retrieval systems. Even though shape similarity matching has been the subject of research for over two decades, no completely satisfactory technique has yet been developed. This is largely due to the difficulty of devising shape similarity measures that accurately model human visual perception, an issue discussed by Ren et al<sup>1</sup>, who highlight the gulf between experimental findings on human similarity judgements and current techniques for automatic image similarity matching, and by Latecki and Lakämper<sup>2</sup>, who postulate five requirements that any useful shape measure needs to meet:

- (a) It should permit recognition of perceptually similar objects
- (b) It should not be affected by noise or segmentation errors
- (c) It should preserve significant visual parts of objects
- (d) It should be independent of scale, orientation or position of objects
- (e) It should not be restricted to any particular class of shapes

Like Ren et al, they hold that since the first three requirements are of a cognitive nature, they should be tested by cognitive experiments - unlike the latter two, which can be demonstrated by mathematical arguments - though their paper does not specify the nature of these cognitive experiments. As discussed below, the design of such experiments is not a trivial task.

Our paper reports the results of one series of experiments aimed at testing the overall retrieval effectiveness of a number of different shape measures, using ground truth from previous evaluation experiments based on real-life shape queries and

similarity judgements. In Latecki and Lakämper's terms, it thus represents an attempt to assess whether the measures tested can satisfy their first three criteria sufficiently well to prove operationally useful in a restricted but important domain - trademark image registration.

## 2. SHAPE SIMILARITY MATCHING

### 2.1 Techniques for similarity matching

A wide variety of techniques for shape similarity matching has been proposed over the years, though few (if any) have been shown to meet the criteria set out above. Some are based on direct matching of complete (information-preserving) representations of object shape, such as chain-codes or splines. Examples of this class of technique include string-matching of chains of boundary pixels<sup>3</sup> comparison of turning angle<sup>4</sup> and elastic deformation of templates<sup>5</sup>. Methods of this kind can have high discriminating power, at least when matching highly similar shapes, but often involve a matching process which is computationally very expensive. Hence most frequently-used techniques for shape similarity matching are based on the comparison of features such as edge direction histograms or moment invariants, which can capture important aspects of an object's appearance, but which cannot be used to reconstitute its entire shape. Commonly-used types of feature include:

- **Simple global features.** Several computationally simple measures of a region's overall shape have been proposed over the years, such as aspect ratio, circularity, and convexity<sup>6</sup>. Though widely used in the past, such features have become less popular as the overhead of using the more computationally expensive features listed below has been rendered less significant by increases in computing power.
- **Local features.** Features representing shape characteristics of small regions of an image can often act as a useful complement to global measures. Examples include the line-angle-line triplet features devised by Eakins<sup>7</sup>, and the longer segment sequences used by Mehrotra and Gary<sup>8</sup>. Their use in operational image retrieval systems to date has been limited.
- **Edge direction histograms.** Another indirect measure of shape within an image is to compute a histogram of edge directions by identifying edge pixels, computing edge directions, and accumulating these into bins at appropriate intervals<sup>9</sup>. This can give an indication of directionality within the image, though not necessarily the shape of any object it depicts. Such measures are included in at least one commercially available CBIR system.
- **Fourier descriptors.** A very popular way of representing a region's overall shape is to represent the cumulative curvature around the boundary as a function of curve length, and expand this function as a Fourier series<sup>10</sup>:

$$\theta(t) = \mu_0 + \sum A_k \cos(kt - a_k)$$

The coefficients  $A_k$  and  $a_k$ , the  $k$ th harmonic amplitude and phase angle respectively, known as the *Fourier descriptors* of the curve, provide a description of the curve which appears to reflect its overall shape fairly consistently.

- **Moment invariants.** For any digital image  $I(x,y)$ , it is possible to compute a series of central moments  $\mu_{pq}$ , defined as

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q I(x, y)$$

from which a series of *moment invariants*  $\phi_n$  can be derived which characterize shape in a manner which is invariant to scaling, rotation and translation<sup>11</sup>. Moment invariants have been widely used in image analysis for many years.

- **Affine moment invariants.** A set of four moment invariants, invariant under the general affine transformation

$$\begin{aligned} u &= a_0 + a_1x + a_2y \\ v &= b_0 + b_1x + b_2y \end{aligned}$$

was proposed by Flusser and Suk<sup>12</sup>, who demonstrated the value of such invariants in matching objects deformed by affine transforms.

- **Zernike moments.** The Zernike moment of order  $n$  with repetition  $m$  for image  $I(r,\theta)$  is defined as:

$$A_{nm} = \frac{n+1}{\pi} \sum_{\rho} \sum_{\theta} (R_{nm}(r)e^{im\theta})^* I(r, \theta) \quad |r| < 1$$

where  $R_{nm}(r)$  are the set of radial polynomials originally defined by Zernike<sup>13</sup>. Zernike moments have the useful property of orthogonality, and have been applied to a number of image analysis problems, including trademark recognition<sup>14</sup>.

## 2.2 Effectiveness of different techniques

Little hard information on the relative effectiveness of different shape representation and matching techniques is currently available, since few cognitive studies of the kind called for by Latecki and Lakämper<sup>2</sup> have yet been carried out. Even fewer have yielded results which can be relied on. Although most reports of new image retrieval techniques do now include quantitative data on retrieval effectiveness, judgements on what constitutes a match are all too often made by the development team themselves. Since such judgements are inherently subjective, this inevitably introduces an element of subconscious bias - a problem well recognized in the wider information retrieval field<sup>15</sup>. A recent study by Squire and Pun<sup>16</sup>, comparing the results of human and machine classification of images by similarity of appearance, confirms that this problem does indeed exist in the field of image classification and retrieval. One of the most striking results of their trial was that the judgements made by the paper's author (who had also developed much of the software) correlated far better with the machine's results than any of the independent observers - highlighting the dangers of relying on in-house judgements.

Such considerations limit the validity of the results reported by Mehtre et al<sup>17</sup>, who compared the effectiveness of several boundary- and region-based shape methods (including Fourier descriptors, moment invariants, and Zernike moments) over 15 queries put to a database of 500 trademarks. All shape measures seemed to perform well, with moment invariants the best single technique, and moment invariants plus modified Fourier descriptors giving the highest overall scores. Unfortunately, the authors failed to obtain independent judgements of the relevance of each retrieved image to each query.

Rather different results were reported from a study conducted by Scassellati et al<sup>18</sup>, which compared machine estimations of shape similarity with those of independent human observers. Their experiments were designed to assess the relative effectiveness of a number of different methods for automated shape similarity assessment, including algebraic moments, parametric curve distance, turning angle, and modified Hausdorff distance. 40 human subjects were asked to scan a database of 1415 simple shapes to identify all shapes they considered similar to each of 20 query shapes. Machine rankings of similarity were then obtained for each of the 20 queries, with each similarity assessment technique. The effectiveness of each technique was calculated from the total number of times each of the top 20 retrieved images had been selected by human judges. In contrast to Mehtre et al, Scassellati et al found that no technique performed particularly well (scores were typically only 20-30% of perfect performance), though turning angle appeared to give better overall performance than any other method. It seems premature to draw any firm conclusions from either study about the effectiveness of different shape retrieval techniques.

## 2.3 "Natural" shape features

The concept of shape similarity is clearly a problematic one. More understanding of human visual perception is needed to tackle the shape matching and retrieval problem successfully. Unfortunately, few detailed investigations of human shape similarity judgement have been reported in the literature. Studies such as those of Goldmeier<sup>19</sup> provide fascinating insights into the process of shape similarity estimation. But they do not in themselves directly lead to definitions of shape features capable of modelling human similarity judgements.

One source of features which has received surprisingly little attention in this respect is the set of generic shapes which are so well known that they are distinguished by name, such as square, triangle or circle. Such features play an important part in the development of shape recognition capabilities in children<sup>20</sup>, and it can be hypothesized that the ability to recognize such basic features is a fundamental part of the mechanism of human shape characterization. Support for this hypothesis comes from the fact that some of the principal shape classes in well-known manual classification systems like the Vienna classification of trademark images<sup>21</sup> are triangles, circles, and squares. Similarly, Dyson and Box's investigation of elements that human observers found useful in distinguishing between pairs of images in order to derive a set of features of potential use in image retrieval showed that some of the most frequently used features were triangles, rectangles, circles and ellipses<sup>22</sup>. It therefore seems reasonable to postulate that the degree to which a shape can be regarded as (say) a triangle or a rectangle could be a useful feature for retrieval. While circularity has been extensively used as a measure in computer vision, few other types of shape have received attention. Recently, however, Rosin has proposed and evaluated a number of possible measures of triangularity, rectangularity and ellipticity, showing that they have considerable promise as perceptually significant shape discriminators<sup>23</sup>. Their robustness to noise and variation in aspect ratio could be particularly valuable in discrimination of artificial images designed to have a strong visual impact, such as product designs and trademarks.

### 3. TRADEMARK IMAGE RETRIEVAL

#### 3.1 Techniques for trademark matching

Trademark registration is a process of considerable commercial significance. It is the task of trademark registries around the world to ensure that when a new trademark is submitted for registration, it is sufficiently distinct from all existing marks that there is no danger of confusion. Trademarks may consist of words, images or a combination of both: images may be abstract geometric shapes, illustrations of real or mythical beings, or any combination. The process of analysing trademark image similarity is clearly complex. Wu et al<sup>24</sup> have identified three components of similarity - shape, structure and semantics. Traditionally, trade mark registries have relied on manually-assigned codes from schemes such as the Vienna classification<sup>21</sup>, which mixes all three elements to some extent.

The earliest report of a CBIR system designed specifically for trademark image retrieval is that of Kato<sup>25</sup>, who describes a system known as TRADEMARK. This used the relatively simple approach of mapping normalized trademark images to an  $8 \times 8$  pixel grid, and calculating a *GF-vector* for each image from various pixel frequency distributions. Query and stored images could then be matched by comparing GF-vectors. Kato's system matched trademarks purely as complete images, though most subsequent researchers have regarded trademark images as multi-component objects, capable of being matched at more than one level. For example, the STAR\* system developed by Wu et al<sup>24</sup> allows human indexers to segment trademark images into perceptually meaningful components, from which shape features such as Fourier descriptors and moment invariants are extracted. Overall similarity between trademarks is expressed as a distance measure computed from the weighted sum of component distances. Peng and Chen<sup>26</sup> take the principle of component matching one stage further. Their technique involves approximating each image component as a set of (possibly overlapping) closed contours, and representing each contour as a list of angle descriptors. Images are then matched in hierarchical fashion: contour-contour, component-component and finally image-image, using appropriate similarity functions to propagate similarity values between levels.

Some researchers prefer to match trademarks as complete images. Jain and Vailaya<sup>27</sup>, for example, describe a technique for shape retrieval based on comparing normalized edge histograms from whole images. They demonstrate its usefulness in retrieving scaled, rotated and noisy versions of a given image. Kim and Kim<sup>14</sup> show how Zernike moments can be used to capture rotational symmetry in an image, and illustrate their potential usefulness in retrieval with a number of examples. The authors suggest that these descriptors need to be combined with other types of feature to achieve fully effective retrieval. An alternative approach to multi-level trademark image representation and matching is that of Ravela and Manmatha<sup>28</sup>. They compute Gaussian derivatives for every point on the image at several scales, and then derive histograms of local curvature and phase. Image matching is then performed by calculating normalized cross-covariance between curvature and phase histograms. Their system has been used as the basis for a combined prototype text and image search engine which is currently being tested on a database of 63 000 US trademarks.

It is currently not possible to compare the effectiveness of these alternative techniques, since no comparative evaluation studies have been performed. The fact that no CBIR system is currently in routine use at any national trademark registry suggests that no system yet meets the exacting standards required by trademark examiners.

#### 3.2 The ARTISAN project

The ARTISAN<sup>†</sup> system<sup>29</sup>, developed at the University of Northumbria in collaboration with the UK Patent Office, is based on an underlying philosophy similar to that of STAR, though with a rather more restricted scope. It relies for shape matching on a combination of simple global features calculated both from individual image components and from perceptually significant families of components. Unlike STAR, ARTISAN identifies these perceptually significant regions automatically. Images are automatically segmented into closed-boundary regions, which are then aggregated into perceptually significant groupings known as *families*, using principles derived from Gestalt psychology<sup>30</sup>. Features are extracted and stored at the level of both the family and the individual component, allowing matching to be performed at the level of the entire image, the component family, or the individual component. The system also allows a choice of matching paradigms, to cope with the fact that in general, query and stored images have different numbers of components.

The retrieval effectiveness of version 1 of ARTISAN has been evaluated using a selection of real queries put to a database of 10 745 abstract geometric images from the UK Trade Marks Registry<sup>31</sup>. Overall retrieval effectiveness scores were encouraging, but not good enough for operational use. Work is now under way on the development of an improved version

---

\* *System for Trademark Archival and Retrieval*

† *Automatic Retrieval of Trademark Images by Shape ANalysis*

of the ARTISAN prototype, reflecting some - though not all - of the lessons learnt from failure analysis of version 1<sup>29</sup>. Important changes being incorporated into version 2 include the use of multiresolution analysis to remove texture and improve the system's ability to cope with noisy images, new ways of grouping low-level components into higher-level regions, and a wider range of shape and structural features.

Boundary creation within ARTISAN version 2 is performed by segmenting a multi-resolution representation of the trademark. The trademark is first processed into a multi-resolution pyramid using Burt and Adelson's algorithm<sup>32</sup>, in which a level  $L$  Gaussian image  $g_L(i,j)$  is computed from the corresponding level  $L-1$  image  $g_{L-1}(i,j)$  by the formula:

$$g_L(i,j) = \sum_{m=-2}^2 \sum_{n=-2}^2 w(m,n) g_{L-1}(i+m*2^{L-1}, j+n*2^{L-1})$$

where  $w(m,n)$  is a  $5 \times 5$  Gaussian convolution kernel.

As illustrated in Figs 1 and 2, four pyramid levels are created (zero being the original trademark). Although it is possible to construct higher levels of this pyramid, no advantage is gained in detecting the image's overall shape. The energy of each level of the pyramid (calculated using the square of the Laplacian pyramid at level  $L$ , the difference between levels  $L$  and  $L+1$  of the Gaussian pyramids) is then used to classify the original image as simple, intermediate, or complex. Examples of simple and complex images are shown above. For simple images such as that shown in Fig 1, it has been found advantageous to select levels 0 and 3 for further processing, as this allows both the overall shape of the image and the



Fig 1. The four levels of Gaussian smoothing 0-3 derived from a simple image. Level 0 (left) is the original image.



Fig 2. The four levels of Gaussian smoothing applied to a complex image

shapes of its components to be represented. For complex images such as the one in Fig 2, levels 1 and 3 are segmented, allowing us to discard unnecessary fine detail. For intermediate images (mainly images with large amounts of coarse texture), levels 2 and 3 are selected, again allowing us to avoid identifying spurious shapes in textured regions (a major problem in the initial version of ARTISAN). As with previous versions of ARTISAN, all image preprocessing and level selection is done automatically. No manual intervention is required at any stage. Finally, region boundaries are identified in

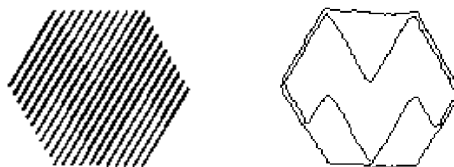


Fig 3. An example of image segmentation, showing region boundaries extracted from a complex image.

each of the selected image levels using a smoothed histogram trough detection algorithm similar to the method described by Pauwels et al<sup>33</sup>. An example of region boundary detection is shown in Fig 3.

Applying multi-resolution techniques to the segmentation process provide three benefits. Firstly, it improves segmentation of textured regions by merging texture into greyscale, allowing simple histogram thresholding techniques to extract the

region boundary. Secondly, the band pass filtering removes high frequency noise (often caused by fax scanning) increasing the quality of the segmented boundaries. Finally, multi-resolution techniques also simulate some of the Gestalt processes of the previous ARTISAN v1 system; in particular proximally associated boundaries are robustly coupled.

## 4. THE PRESENT STUDY

### 4.1 Shape description in ARTISAN

The main motivation for the present study is to compare the effectiveness for trademark image retrieval of alternative sets of shape descriptors, including some of the "simple" features used in version 1 of ARTISAN, three of the "natural" shape descriptors proposed by Rosin<sup>23</sup>, and some frequently-used measures such as Fourier descriptors and moment invariants. Our hypothesis is that if the "natural" descriptors do indeed model human shape perception more accurately than Fourier descriptors or moment invariants, this should be reflected in significantly improved retrieval performance.

Our current version of ARTISAN computes and stores the following features for each individual component of every image added to the database:

- (a) 3 "simple" shape features: aspect ratio  $w/l$ , circularity  $4\pi A/p^2$ , and convexity  $A/H$ , where  $A$  is the region area,  $p$  its perimeter,  $l$  and  $w$  the length and width respectively of its minimum bounding rectangle, and  $H$  the area of the region's convex hull.
- (b) 3 "natural" shape features proposed by Rosin<sup>23</sup>, using measures which appeared from his experiments to be most robust to noise and changes in aspect ratio:

- **Triangularity**, defined as

$$108I_1 \text{ where } I_1 \leq 1/108, \\ 1/108I_1 \text{ otherwise.}$$

where  $I_1$  is the first of the set of affine moment invariants proposed by Flusser & Suk<sup>12</sup> (see (e) below).

- **Rectangularity**, defined as

$$R_D = 1 - \frac{R+D}{B}$$

where  $B$  is the area of the region's bounding rectangle,  $D$  the area of the difference of the region and the rectangle, and  $R$  the difference of the rectangle and the region.

- **Ellipticity**, defined as

$$16\pi^2 I_1 \text{ where } I_1 \leq 1/16\pi^2, \\ 1/16\pi^2 I_1 \text{ otherwise.}$$

- (c) 8 normalized Fourier descriptors  $A'_k$  ( $1 \leq k \leq 8$ ), where  $A'_k = A_{k+1}/A_1$ , and  $A_k$  is the  $k$ th harmonic amplitude of the Fourier expansion of the cumulative curvature around the region boundary as a function of curve length:

$$\theta(t) = \mu_0 + \sum_{k=1}^{\infty} A_k \cos(kt - a_k)$$

- (d) The 7 normal moment invariants  $\varphi_n$  defined by Hu<sup>11</sup>:

$$\begin{aligned} \varphi_1 &= (\eta_{20} + \eta_{02}) \\ \varphi_2 &= (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \\ \varphi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\ \varphi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\ \varphi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\{(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\} + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\{3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\} \\ \varphi_6 &= (\eta_{20} - \eta_{02})\{(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\} + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\ \varphi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\{(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\} - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})\{3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\} \end{aligned}$$

$$\text{where } \eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma}, \text{ and } \gamma = \frac{p+q}{2} + 1$$

- (e) The 4 affine moment invariants defined by Flusser and Suk<sup>12</sup>:

$$I_1 = \frac{\mu_{20}\mu_{02} - \mu_{11}^2}{\mu_{00}^4}$$

$$I_2 = \frac{\mu_{30}^2\mu_{03}^2 - 6\mu_{30}\mu_{21}\mu_{12}\mu_{03} + 4\mu_{30}\mu_{12}^2 + 4\mu_{03}\mu_{21}^2 - 3\mu_{21}^2\mu_{12}^2}{\mu_{00}^{10}}$$

$$I_3 = \frac{\mu_{20}(\mu_{21}\mu_{03} - \mu_{12}^2) - \mu_{11}(\mu_{30}\mu_{03} - \mu_{21}\mu_{12}) + \mu_{02}(\mu_{30}\mu_{12} - \mu_{21}^2)}{\mu_{00}^7}$$

$$I_4 = \frac{(\mu_{20}^3\mu_{03}^2 - 6\mu_{20}^2\mu_{11}\mu_{12}\mu_{03} - 6\mu_{20}^2\mu_{02}\mu_{21}\mu_{03} + 9\mu_{20}^2\mu_{02}\mu_{12}^2 + 12\mu_{11}^2\mu_{20}\mu_{21}\mu_{03} + 6\mu_{20}\mu_{11}\mu_{02}\mu_{30}\mu_{03} - 18\mu_{20}\mu_{11}\mu_{02}\mu_{21}\mu_{12} - 8\mu_{11}^3\mu_{30}\mu_{03} - 6\mu_{02}^2\mu_{20}\mu_{12}\mu_{30} + 9\mu_{02}^2\mu_{20}\mu_{21}^2 + 12\mu_{11}^2\mu_{02}\mu_{12}\mu_{30} - 6\mu_{02}^2\mu_{11}\mu_{21}\mu_{30} + \mu_{02}^3\mu_{30}^2)}{\mu_{00}^{11}}$$

In addition, a *relative size* parameter is computed for each image component. This is the ratio of the area of each component to the area of the largest component in the image (normally the envelope of the entire image). This allows matching to be based either on shape alone, or on shape and size. Any combination of the above feature types can be selected at query time for similarity matching.

The system provides several alternative means of computing overall similarity scores between query and stored images from component similarities, to take account of the fact that in general, query and stored images will have different numbers of components. The two methods of comparison used in this study were the two which had been shown to perform most effectively in previous trials of ARTISAN. These were *symmetric strict*, which averages similarity scores for the  $\min(q,s)$  closest component matches, where  $q$  and  $s$  are the numbers of components in query and stored images respectively, and *asymmetric simple*, which averages similarity scores for the  $q$  closest component matches<sup>31</sup>. The results presented in tables 1-3 are all based on asymmetric simple matching: scores based on symmetric strict matching were slightly but consistently lower in every case. In contrast to earlier versions of ARTISAN, component similarities were computed as city-block distances between vectors representing the unweighted values of currently-selected shape features. Several other run-time options (such as whole-image matching) are provided in the new ARTISAN prototype, but were not tested in this study.

## 4.2 Evaluation experiments

As indicated above, the process of evaluating retrieval effectiveness is far from straightforward. The approach adopted in this paper follows the methodology developed for our evaluation of the original version of ARTISAN<sup>31</sup>. A set of 24 query trademarks selected by staff at the UK Trade Marks Registry for evaluation of the first version of ARTISAN (illustrated in Table 4) was run against the 10 745 image database of abstract geometric trademarks also supplied by the Registry. Results were compared with relevance judgements already generated by human trademark examiners, which formed our "ground truth". This approach ensured that the performance of each set of features was compared against the same benchmark - though it is still open to the criticism that the human judgements forming our ground truth were based on overall assessments of image similarity (potentially based in shape, structure and semantics), not necessarily judgements based purely on shape.

Several combinations of shape feature were tested against this database, as follows:

- The three "simple" measures defined in (a) above;
- The three "natural" measures proposed by Rosin, and defined in (b) above;
- Fourier descriptors (using either the 3 lowest-frequency or all 8 coefficients);
- Moment invariants (using either the first 4 or all 7 normal measures, or all 4 affine measures);
- Various combinations of simple features, Rosin measures, Fourier descriptors and moment invariants.

In each case, all 24 queries were run against all 10 745 images, using both the *asymmetric simple* and *symmetric strict* matching paradigms defined above. Retrieval effectiveness was measured using the same set of measures as in the original ARTISAN evaluation: normalized precision  $P_n$ , normalized recall  $R_n$ , and last-place ranking  $L_n$ :

$$P_n = 1 - \frac{\sum_{i=1}^n (\log R_i) - \sum_{i=1}^n (\log i)}{\log\left(\frac{N!}{(N-n)!n!}\right)}$$

$$R_n = 1 - \frac{\sum_{i=1}^n R_i - \sum_{i=1}^n i}{n(N-n)}$$

$$L_n = 1 - \frac{R_{last} - n}{N-n}$$

Each of these measures gives an estimate of retrieval effectiveness in the range 0-1, but emphasizes different aspects of system performance. Broadly,  $P_n$  gives an overall measure of system performance at all retrieval ranks,  $R_n$  emphasizes good performance at high retrieval ranks, and  $L_n$  indicates how effective the system is at retrieving *all* relevant images. While a less robust measure than the other two, it is particularly important in the context of trademark image registration, as the penalties for missing even a single relevant image can be substantial.

### 4.3 Results

The first set of experiments were designed to measure the relative effectiveness of the three "natural" descriptors proposed by Rosin and three of the "simple" shape features used in the original version of ARTISAN.

**Table 1: Retrieval results using simple measures and Rosin descriptors**

Query parameters	Size?	Rn	Pn	Ln
Rectangularity	+	0.79±0.03	0.48±0.05**	0.46±0.05
	-	0.80±0.03	0.34±0.03**	0.45±0.06
Triangularity	+	0.83±0.02	0.52±0.04**	0.46±0.05
	-	0.84±0.02	0.42±0.03**	0.44±0.06
Ellipticity	+	0.84±0.02	0.53±0.04**	0.45±0.05
	-	0.84±0.02	0.45±0.04**	0.48±0.06
3 Rosin descriptors (3RD)	+	0.88±0.02	0.59±0.04*	0.55±0.06
	-	0.88±0.02	0.54±0.04*	0.54±0.06
3 simple descriptors (3SD)	+	0.89±0.02	0.60±0.05	0.62±0.06
	-	0.89±0.02	0.60±0.05	0.61±0.06

*All figures represent mean and standard error effectiveness scores from the same set of 24 queries. Asymmetric simple matching was used in all cases. Key - \*\*: size difference significant over query set at  $P < 0.01$  level, \*: size difference significant,  $P < 0.05$  (Wilcoxon matched-pairs, signed-rank test)*

Overall retrieval results for each Rosin descriptor taken separately were only modest compared with some of the combined measures tried. For single descriptors their discriminating power was quite encouraging. When combined, their effectiveness increased markedly - their effectiveness was not significantly different from that of the three simple descriptors in combination. Perhaps disappointingly, there is no evidence from these experiments that the newly-proposed measures of triangularity, ellipticity and rectangularity can on their own capture human shape similarity judgments more effectively than more traditional features. All three Rosin descriptors appeared to perform more effectively when relative size was taken into account.

The second set of experiments examined the retrieval effectiveness of some widely-used but more computationally intensive shape features, Fourier descriptors and moment invariants.



**Table 2: Retrieval results using Fourier descriptors and moment invariants**

Query parameters	Size?	Rn	Pn	Ln
4 affine moment Invariants (AMI)	+	0.70±0.03**	0.40±0.05**	0.31±0.05*
	-	0.85±0.02**	0.49±0.04**	0.48±0.06*
4 normal moment Invariants (4MI)	+	0.79±0.03*	0.50±0.05	0.45±0.05
	-	0.85±0.03*	0.57±0.05	0.52±0.06
7 moment invariants (7MI)	+	0.79±0.03*	0.51±0.05	0.45±0.05
	-	0.85±0.03*	0.57±0.05	0.52±0.06
3 Fourier descriptors (3FD)	+	0.88±0.03	0.61±0.04	0.60±0.05
	-	0.89±0.02	0.62±0.04	0.64±0.06
8 Fourier descriptors (8FD)	+	0.90±0.02	0.65±0.04	0.66±0.06
	-	0.91±0.02	0.66±0.04	0.65±0.06

Key - \*\*: size difference significant at  $P<0.01$  level, \*: size difference significant at  $P<0.05$  level (Wilcoxon matched-pairs signed-rank test)

The more computationally intensive features performed no better - and in some cases worse - than the simple features above. Fourier descriptors gave overall results almost identical to those achieved using simple features; moment invariants significantly worse ( $P<0.05$ , Wilcoxon matched-pairs signed-rank test) for all three performance measures used. Use of just the three lowest-frequency Fourier descriptors gave performance that was almost as good as the use of the first eight - suggesting that the higher-frequency components may be matching noise as much as genuine shape variations. Neither normal nor affine moment invariants produced especially impressive figures. In contrast to the Rosin descriptors, they appeared to be less effective when relative size was taken into account in shape matching. The poor overall performance of the affine moment invariants seemed particularly surprising, as they form the basis for two of the "natural" measures proposed by Rosin, both of which appear markedly more successful in discriminating between the shapes used in this study than the affine moment invariants themselves.

























**Table 3: Retrieval results using combined measures**

Query parameters	Size?	Rn	Pn	Ln
3FD+4MI	+	0.89±0.03	0.62±0.04	0.62±0.06
	-	0.89±0.03	0.63±0.04	0.64±0.06
3SD+3FD	+	0.91±0.02	0.65±0.04	0.67±0.06
	-	0.90±0.02	0.65±0.05	0.63±0.06
3SD+3RD	+	0.92±0.02	0.66±0.04	0.67±0.06
	-	0.92±0.02	0.66±0.04	0.64±0.06
3RD+3FD	+	0.93±0.02	0.67±0.04	0.69±0.06
	-	0.93±0.02	0.68±0.04	0.67±0.06
3SD+3RD+3FD	+	0.93±0.02	0.69±0.04	0.71±0.06
	-	0.92±0.02	0.69±0.04	0.65±0.07
3SD+3RD+8FD	+	0.94±0.02	0.70±0.04	0.71±0.06
	-	0.93±0.02	0.69±0.04	0.66±0.07

Combining different measures produced a further slight but significant increase in system effectiveness, with the best performance among the combinations tested shown by the combination of simple, Rosin and Fourier descriptors. All combinations of two or more types of descriptor gave significantly higher  $P_n$  and  $R_n$  scores over the 24 queries used in our tests than simple, Rosin or Fourier descriptors on their own ( $P<0.01$ , Wilcoxon matched-pairs signed-rank test).  $L_n$  scores for these combinations were significantly higher than those for Rosin or Fourier descriptors on their own, but not for simple descriptors. The least effective combination tested was Fourier descriptors plus moment invariants, which proved significantly less effective ( $P<0.05$  or better) than any of the other combinations listed in table 3, at least when  $P_n$  or  $R_n$  measures were used. This was one of the combinations that proved most effective in the experiments performed by Mehtre et al<sup>17</sup>. Our study provides no support for their findings.

The average performance figures presented here tell only part of the story. For individual queries, different techniques often produced markedly different results. As an indication of this, Table 4 summarizes the performances of the best individual combination of features for each query. The parameter combinations appearing to give optimum query performance most frequently were Rosin, simple and Fourier descriptors (8 instances), Rosin and simple descriptors (4 instances), and Fourier descriptors alone (4 instances). The average effectiveness scores of these "best" parameter combinations were  $R_n$ :  $0.96 \pm 0.01$ ,  $P_n$ :  $0.74 \pm 0.04$ , and  $L_n$ :  $0.80 \pm 0.05$ , significantly better ( $P < 0.01$  whichever performance measure was used) than either the overall effectiveness scores achieved with any single parameter combination, or the best scores achieved with the first version of ARTISAN ( $R_n$ :  $0.90 \pm 0.02$ ,  $P_n$ :  $0.63 \pm 0.05$ , and  $L_n$ :  $0.56 \pm 0.06$ ).

**Table 4: Best retrieval results for each individual query**

Query Image	Best results			Best combination*			Query Image	Best results			Best combination		
	Rn	Pn	Ln	Shape	Match	Size		Rn	Pn	Ln	Shape	Match	Size
	1.00	1.00	1.00	Nearly all combinations	AS or SS	+ or -		0.98	0.71	0.93	3RD+3SD+8FD	SS	+
	1.00	1.00	1.00	3SD+3RD+8FD	AS	+ or -		0.98	0.59	0.95	3RD+3FD	AS	+
	1.00	0.93	0.98	3RD+3SD+3FD	AS	+		0.97	0.84	0.82	3RD	AS	-
	1.00	0.90	1.00	3RD+3SD	SS	+		0.97	0.80	0.91	8FD	AS	+
	1.00	0.84	0.98	3RD+3SD+8FD	SS	-		0.97	0.75	0.90	3SD+3RD	AS	+
	0.99	0.91	0.95	3SD+3RD	AS	+		0.97	0.57	0.85	3RD+3SD	AS	+
	0.99	0.88	0.95	3FD+4MI	AS	-		0.94	0.69	0.75	3FD+4MI	SS	-
	0.99	0.87	0.91	3RD+3SD+8FD	AS or SS	-		0.92	0.81	0.60	3SD	AS	+
	0.99	0.81	0.97	3 RD	SS	+		0.92	0.64	0.45	8FD	AS	-
	0.99	0.77	0.89	4MI or 7MI	AS or SS	-		0.81	0.43	0.28	3FD	AS	+
	0.98	0.77	0.87	3SD+3FD+8FD	AS	+		0.81	0.33	0.15	3SD+3RD+3FD	AS	+
	0.98	0.71	0.90	8FD	AS	-		0.79	0.28	0.30	3SD	AS	-

\* Key: 3RD - three Rosin descriptors; 3SD - three simple descriptors; nFD - n lowest-frequency Fourier descriptors; nMI - first n moment invariants; AS - asymmetric simple; SS - symmetric strict. Entries are ordered by  $P_n$  scores within  $R_n$ .

#### 4.4 Discussion

Overall, our findings suggest that a variety of global shape measures can prove useful in supporting retrieval of multi-component images such as trademarks. However, our initial hypothesis that "natural" measures of the type proposed by Rosin<sup>23</sup> would prove more effective than other feature types could not be confirmed. The differences in overall system effectiveness between different combinations of shape measure were generally small and not statistically significant. This suggests that there may be little to be gained from further research into new global shape measures. Improvements in retrieval effectiveness are more likely to come from the development of different paradigms for shape representation (e.g. Latecki and Lakamper<sup>2</sup>) or matching (e.g. Santini and Jain<sup>34</sup>).

The results in Table 4 are more an indication of system potential than actual effectiveness. They indicate that to get the best out of any system offering a variety of shape features, considerable effort needs to be devoted to selection of the most

appropriate features for any given query. At present the only realistic way in which an end-user can realistically exploit the flexibility offered by such systems is by using techniques such as relevance feedback<sup>35</sup>, where searchers are able to indicate the relevance of each item retrieved, and the system can adjust its search strategy accordingly – e.g. by increasing the weighting of parameters featuring more strongly in user-selected items. Our study suggests that techniques such as relevance feedback could indeed prove useful in improving the effectiveness of retrieval systems designed for multi-component images such as trademarks.

It should be remembered that human similarity judgements among images of this sort are based on a mixture of shape, structure and semantic cues<sup>24</sup>. Two of the query images providing the ground truth for our experiments (the last two in the right-hand column of images in Fig 4) demonstrate this point. Examination of the set of desired images for each of these queries suggests strongly that searchers are looking in one case for repeated patterns, in the other for groups of lines implying a particular type of crossover. In these queries, shape similarity does not appear to be the prime consideration. The current version of ARTISAN (which matches on shape but not on structural similarity) performs poorly on both these queries whatever set of shape features is chosen. Major improvements in system effectiveness thus appear to depend on developing better techniques for matching on image structure – and if possible on implied features, too. The development of better techniques for image segmentation, or preferably matching techniques which avoid the need for image segmentation altogether (e.g. Ravela and Manmatha<sup>28</sup>), represents another possible way forward.

The strength of our evaluation experiments lies in our use of ground truth derived from a set of real image queries and independent relevance judgements made by experts in the field. However, this is also a limitation in a number of respects: the total number of queries is not large, not all of them are pure shape queries, and human relevance judgements are inevitably subjective. There is clearly a danger in tuning our system exclusively to perform well on these 24 queries. We are actively seeking additional sources of ground truth for this purpose, though the difficulty of obtaining independent human similarity judgements with collections of over 10 000 images should not be underestimated.

## 5. CONCLUSIONS

Our experiments have shown that a number of different shape measures can, in the majority of cases, provide adequate discriminating power for effective retrieval by shape similarity in multi-component image collections such as trademark registries. Overall retrieval effectiveness scores appear remarkably similar for nearly all shape measures used. However, marked differences can often be observed between measures for individual queries, suggesting that there is indeed value in providing a wide variety of different measures within any one system, provided a suitable framework can be devised to allow users to search such a multi-dimensional feature space effectively. Future research effort will be devoted both to identifying improved methods of computing image similarity on the basis of shape and structural features, and to developing an appropriate relevance feedback framework to allow users to exploit the flexibility of multi-dimensional feature searching to the full.

## ACKNOWLEDGEMENTS

The financial assistance of Resource (formerly the UK Library and Information Commission) and the UK Patent Office for part of this project is gratefully acknowledged. The views expressed, however, are purely those of the authors.

## REFERENCES

- 1 M Ren, J P Eakins and P Briggs "Human perception of trademark images: implications for retrieval system design" *Journal of Electronic Imaging*, in press
- 2 L J Latecki and R Lakämper "Application of planar shape comparison to object retrieval in image databases" *Pattern Recognition*, in press
- 3 G Cortelazzo et al "Trademark shape description by string-matching techniques" *Pattern Recognition* **27**(8), 1005-1018 (1994)
- 4 E M Arkin et al "An efficiently computable metric for comparing polygonal shapes" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**(3), 209-216 (1991)
- 5 A K Jain et al "Object matching using deformable templates" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**(3), 267-277 (1996)
- 6 M D Levine *Vision in man and machine*, ch 10. McGraw-Hill, N Y, 1985
- 7 J P Eakins "Design criteria for a shape retrieval system" *Computers in Industry* **21**, 167-184 (1993)
- 8 R Mehrotra and J E Gary "Similar-shape retrieval in shape data management" *IEEE Computer* **28**(9), 57-62 (1995)

- 9 A K Jain and A Vailaya "Image retrieval using color and shape" *Pattern Recognition* **29**(8), 1233-1244 (1996)
- 10 C T Zahn, and C Z Roskies "Fourier descriptor for plane closed curves" *IEEE Transactions on Computers* **C-21**, 269-281 (1972)
- 11 M K Hu "Visual pattern recognition by moment invariants" *IRE Transactions on Information Theory* **IT-8**, 179-187 (1962)
- 12 J Flusser and T Suk "Pattern recognition by affine moment invariants" *Pattern Recognition* **26**(1), 167-174 (1993)
- 13 C H Teh and R T Chin "Image analysis by methods of moments" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **10**(4), 496-513 (1988)
- 14 Y S Kim and W Y Kim "Content-based trademark retrieval system using a visually salient feature" *Image and Vision Computing* **16**, 931-939 (1998)
- 15 D Ellis "The dilemma of measurement in information retrieval research" *Journal of the American Society for Information Science* **47**, 23-36 (1996)
- 16 D McG Squire and T Pun "A comparison of human and machine assessments of image similarity for the organization of image databases" in *Proceedings of 10<sup>th</sup> Scandinavian Conference on Image Analysis, Lappeenranta, Finland*, 51-58 (1997)
- 17 B M Mehre et al "Shape measures for content-based image retrieval: a comparison" *Information Processing and Management* **33**(3), 319-337 (1997)
- 18 B Scassellati et al "Retrieving images by 2-D shape: a comparison of computation methods with human perceptual judgements" in *Storage and Retrieval for Image and Video Databases II* (Niblack, W R & Jain, R C, eds), Proc SPIE 2185, 2-14 (1994)
- 19 E Goldmeier "Similarity in visually perceived forms" *Psychological Issues* **8**(1), 1-135 (1972)
- 20 J Piaget and B Inhelder *The child's conception of space*. Routledge and Kegan Paul, 1956
- 21 *World Intellectual Property Organization International Classification of the Figurative Elements of Marks (Vienna Classification)*, Fourth Edition. ISBN 92-805-0728-1. World Intellectual Property Organization, Geneva, 1998
- 22 M C Dyson and H Box "Retrieving symbols from a database by their graphic characteristics: are users consistent?" *Journal of Visual Languages and Computing* **8**, 85-107 (1997)
- 23 P L Rosin "Measuring shape: ellipticity, rectangularity and triangularity" in *Proceedings of 15<sup>th</sup> International Conference on Pattern Recognition, Barcelona, Spain*, **1**, 952-955 (2000)
- 24 J K Wu et al "Content-based retrieval for trademark registration" *Multimedia Tools and Applications* **3**, 245-267 (1996)
- 25 T Kato "Database architecture for content-based image retrieval" in *Image Storage and Retrieval Systems* (A A Jambardino and W R Niblack, eds), Proc SPIE 2185, 112-123 (1992)
- 26 H L Peng and S Y Chen "Trademark shape recognition using closed contours" *Pattern Recognition Letters* **18**, 791-803 (1997)
- 27 A K Jain and A Vailaya "Shape-based retrieval: a case study with trademark image databases" *Pattern Recognition* **31**(9), 1369-1390 (1998)
- 28 S Ravela and R Manmatha "Multi-modal retrieval of trademark images using global similarity" Internal Report, University of Massachusetts at Amherst, 1999
- 29 J P Eakins et al: "Similarity retrieval of trademark images" *IEEE Multimedia*, **5**(2), 53-63 (1998)
- 30 J P Eakins et al "ARTISAN - a shape retrieval system based on boundary family indexing" in *Storage and Retrieval for Image and Video Databases IV*, (I K Sethi and R C Jain, eds), Proc SPIE 2670, 17-28 (1996)
- 31 J P Eakins et al "Evaluation of a trademark retrieval system", in 19th BCS IRSG Research Colloquium on Information Retrieval, Robert Gordon University, Aberdeen, 1997. Available in BCS *electronic Workshops in Computing* series at <http://www.ewic.org.uk/ewic/workshop/view.cfm/IRR-97>.
- 32 P J Burt P J and E H Adelson "The Laplacian pyramid as a compact image code" *IEEE Transactions on Computers* **31**(4), 532-540 (1983)
- 33 E J Pauwels and G Frederix "Content-based image retrieval as a tool for image understanding" in *Multimedia Storage and Archiving Systems IV*, Proc SPIE 3846, 316-327 (1999)
- 34 S Santini and R C Jain "Similarity measures" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**(9), 871-883 (1999)
- 35 C S Lee et al "Information embedding based on users' relevance feedback for image retrieval" in *Multimedia Storage and Archiving Systems IV* (S Panchanathan et al, eds), Proc SPIE 3846, 294-304 (1999)