Attention based similarity

Fred Stentiford[a]

[a]UCL Adastral Park Campus, Martlesham Heath, Ipswich, Suffolk, IP5 3RE, UK
E-mail address: f.stentiford@adastral.ucl.ac.uk

[a]*Corresponding author*. Tel.: +44 1473 663702; fax: +44 1473 635199

Abstract
A similarity measure is described that does not require the prior specification of features or the need for training sets of representative data. Instead large numbers of features are generated as part of the similarity calculation and the extent to which features can be found to be common to pairs of patterns determines the measure of their similarity. Emphasis is given to salient image regions in this process and it is shown that the parameters of invariant transforms may be extracted from the statistics of matching features and used to focus the similarity calculation. Some results are shown on MPEG-7 shape data and discussed in the paper.

Keywords: Visual attention; similarity; shape; Image Retrieval; CBIR

1. Introduction

Similarity measures are central to most pattern recognition problems not least in computer vision and the need to access huge volumes of multimedia content now being broadcast and offered on the Internet. These problems have motivated much research into content based image retrieval [1-4] and many commercial and laboratory systems are described in the literature [5-10].

The notion of similarity is elusive. Quite often having selected a number of appealing features that we believe characterises similarity, we soon discover that new structure in unseen patterns does not contain the desired features despite still apparently possessing a high degree of similarity. By the same token other patterns that do satisfy the similarity criteria seem totally dissimilar to the human eye.

This paper proposes a similarity measure that imposes only very weak assumptions on the nature of the features used in the recognition process. This approach does not make use of a pre-defined distance metric plus feature space in which feature values are extracted from a query image and used to match those from database images, but instead generates features on a trial and error basis during the calculation of the similarity measure. This has the significant advantage that features that determine similarity can match whatever image property is important in a particular region whether it be a shape, a texture, a colour or a combination of all three. It means that effort is expended searching for the best feature for the region rather than expecting that a fixed feature set will perform optimally over the whole area of an image and over every image in the database. By generating thousands of random features and applying them on a trial and error basis as an integral part of the calculation of the similarity value, it is shown that a consistent measure is obtained that is not dependent upon any one or group of specific pattern measurements or representative training sets. Such a system is emergent rather than cognitivist as it does not rely upon a priori specification or programming, but rather constructs its own representation as it interacts and explores the visual scene [11].

Section 2 outlines a selection of papers describing related research and Section 3 then describes the visual attention model used in this paper. The next section defines the similarity measure and Section 5 reports results obtained on shapes from the MPEG-7 test set. This is followed by a discussion section and some conclusions.

2. Background

A great deal of wide ranging research has been carried out on similarity and shape matching and much of this is covered in survey papers such as [12], [13]. Many approaches involve the use of pre-determined features such as edges, colour, location, texture and functions dependent on pixel values e.g. [14]. Mikolajczyk et al [15] use edge models to obtain correspondences with similar objects. Hidden Markov Models derived from shape boundary features are employed by Almageed et al [16] to classify silhouettes. Dao et al [17] obtain measures of similarity between shape boundaries through the use of potential functions borrowed from electrostatics. Thinning algorithms are applied to shapes by Sebastian et al [18] and Iyer et al [19] to obtain graphical features that can be more easily transformed between similar shapes.

The selection of features dependent upon the spatial arrangement of sets of points sampled from shapes is a strategy used by several authors to obtain some of the best

results to date such as [20]. However, all the approaches use pre-determined point selection rules and metrics that can limit performance on unseen data. Viola et al [21] restrict themselves to a specific type of rectangle feature which works well in their face recognition application, but may not perform as well on data that is not suited to this feature.

Increasingly research is turning to models of perception in order to reflect the behaviour of the human visual system in measures of similarity. Mojsilovic et al [22] use perceptually important colours to construct a feature vector for similarity measurement, and overcome the problem of close colours occupying different quantization bins. Super [23] defines critical points of high curvature on boundaries and normalises the shape for rotation and scale before calculating a distance measure. Law et al [24] introduce a measure of saliency in their development of a feature selection and clustering algorithm. A feature is deemed irrelevant if its distribution is independent of class labels. Visual attention models by Itti [25] are used by Frintrop et al [26] to focus computational resource and recognise 3D objects. Shape contours are detected by Grigorescu et al [27] using a model of human visual surround suppression that identifies perceptually significant edges.

3. Visual Attention

Studies in neurobiology [28] are suggesting that human visual attention is enhanced through a process of competing interactions among neurons representing all of the stimuli present in the visual field. The competition results in the selection of a few points of attention and the suppression of irrelevant material. It means that people and animals are able to spot anomalies in a scene no part of which they have seen before and attention is drawn in general to the anomalous object in a scene.

Treisman [29] describes experiments that reveal pre-attentive behaviour in human vision. She points out a “masking effect” that depends upon the presence elsewhere of other elements sharing the local distinctive property. A locally salient feature can be suppressed by more distant structures in the image. Single distinctive features such as colour or orientation promote immediate saliency, but if these properties are cojoined the search for a target is more difficult. Triesman describes several examples of images that exhibit pop out some of which behave asymmetrically. Treisman suggests that there are separate maps for a few properties such as colour, orientation, and brightness, but does not offer a mechanism for their implementation.

Tsotsos [30] presents a pyramidal processing attention model as a means of reducing the complexity arising from a large number of selected features. A winner takes all strategy is imposed on the processing layers in the pyramid so that the more salient objects are identified by location and features at the top of the pyramid. Provision is made to offset a boundary effect in the pyramidal structure that lays emphasis on central items even if they are less significant that peripheral items. Tsotsos highlights features such as size, luminance, edge contrast and orientation as possible features for defining saliency in static images, but there is little guidance on how these might be selected or combined.

Lindenberg [31] provides a framework for detecting salient blob-like objects without relying on a priori information. He stresses that not all significant image structures are

blobs. His research makes the assumption that structures that are significant in scale-space will also be perceptually significant. Although this may be true for some blob configurations it does not apply to all, as for example, no provision is made for the attention suppressing effect of surrounding similar configurations as demonstrated by Treisman [29].

The model of attention [32] used in this paper makes the assumption that an image region is salient if it possesses few features in common with other regions in the image. It follows Lindberg in not making use of *a priori* information regarding the images, but goes further in not making assumptions about the choice of feature measurements as suggested by Treisman [29], Tsotsos [30] and Itti [25]. Such assumptions can force the model to emphasise saliency as determined by colour and orientation, for example, but ignore texture if that did not happen to be a measured feature. As attention is concerned with surprise rather than the expected, this issue is of critical importance.

Let a set of measurements $\boldsymbol{a}_x = (a_{x1}, a_{x2}, a_{x3})$ correspond to a pixel $\boldsymbol{x} = (x_1, x_2)$ Consider a neighbourhood N of $\boldsymbol{x}$ where

$$\{\boldsymbol{x}' \in N \text{ iff } |x_i - x'_i| < \varepsilon_i \; \forall \; i\}$$

Select a set of m random pixels S_x in N (which we call a fork) where

$$S_x = \{\boldsymbol{x}'_1, \boldsymbol{x}'_2, \boldsymbol{x}'_3, ..., \boldsymbol{x}'_m\}.$$

Select another random pixel $\boldsymbol{y}$.
Define the set $S_y = \{\boldsymbol{y}'_1, \boldsymbol{y}'_2, \boldsymbol{y}'_3, ..., \boldsymbol{y}'_m\}$ where

$$\boldsymbol{x} - \boldsymbol{x}'_i = \boldsymbol{y} - \boldsymbol{y}'_i \, .$$

In other words $\boldsymbol{y}$ and $\boldsymbol{y}'_j$ are in the same spatial relationship to each other as $\boldsymbol{x}$ is to $\boldsymbol{x}'_j$. The fork centred on $\boldsymbol{x}$ is said to match that of $\boldsymbol{y}$ if

$$\left|a_{xj} - a_{yj}\right| \le \delta_j \text{ and } \left|a_{x'_i j} - a_{y'_i j}\right| \le \delta_j \quad \forall i, j$$

So if all the colour values of corresponding pixels in S_x and S_y are within a threshold δ of each other the forks will match.

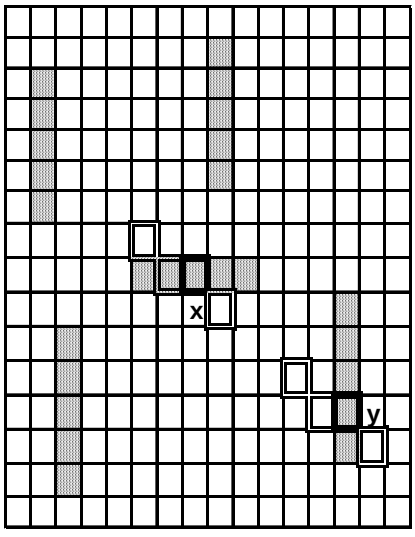

Fig. 1. Fork at $\boldsymbol{x}$ mismatching at $\boldsymbol{y}$.

A pixel $\boldsymbol{x}$ will be worthy of attention if a sequence of M forks matches only a few other neighbourhoods in the space as this will reflect the uniqueness of its properties and hence the saliency. To illustrate the algorithm in the case of a still image (Fig. 1), a fork of 3

pixels $\boldsymbol{x}'$ is shown in the neighbourhood of a pixel $\boldsymbol{x}$. Each of the pixels might possess three colour intensities, so $\boldsymbol{a} = (r,g,b)$. The neighbourhood of a second pixel $\boldsymbol{y}$ matches the first if the colour intensities of all m + 1 corresponding pixels have values within δ of each other. The attention score for each pixel $\boldsymbol{x}$ is incremented each time a mismatch occurs in the sequence of fork comparisons. Pixels $\boldsymbol{x}$ that obtain frequent mismatches over a range of M forks S_x and pixels $\boldsymbol{y}$ are assigned a high visual attention score. This means, for example, that regions possessing novel colour values that do not occur elsewhere in the image will be assigned high scores. It also means that neighbourhoods that span different coloured regions (e.g. edges) are naturally given high scores if those colour adjacencies only occur rarely in the scene.

The gain of the scoring mechanism is increased significantly by retaining the forks S_x if a mismatch is detected, and re-using S_x for comparison with the next of the t neighbourhoods. If however, S_x subsequently matches another neighbourhood, the score is not incremented, and an entirely new fork S_x is generated ready for the next comparison. In this way competing forks are selected against in a competitive fashion if they contain little novelty and turn out to represent structure that is common throughout the image. Indeed it is likely that if a mismatching fork is generated, it will mismatch again elsewhere in the image, and this fork once found, will accelerate the rise of the visual attention score.

If some of the distinguishing characteristics of anomalous objects are known it would be sensible to constrain or even fix the selection of forks S_x rather than generate them randomly For example, a horizontal bar against a background of vertical bars would be optimally identified as salient by a +shaped fork: $S_x = \{(-1,0),(0,-1),(0,1),(1,0)\}$ with $m = 4$ and $\varepsilon = 1$. A completely filled 3x3 patch with $m = 8$ would also function well. However, larger values of m lead to higher frequencies of mismatching on more complex images that obscure the most salient objects. This may be offset by introducing a measure of correlation and a threshold to obtain a match, but again there is no guarantee that a particular correlation measure will place the correct weighting on any content that may be encountered within the patch and there is a danger that in general saliency will be masked.

The model described above uses a simple translation to obtain the fork S_y to test for a match or mismatch. This may be generalised to enable invariance to orientation and other attentional properties by mapping the pixels in S_y according to the appropriate transform before testing for a match. In the case of the orientation property a rotation by a random angle θ is given by

$$S_y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, ..., \boldsymbol{y}_m\} \text{ where } \boldsymbol{y}_i - \boldsymbol{y}_1 = R_\theta[\boldsymbol{x}_i - \boldsymbol{x}_1] \quad \forall\, i$$

$$\text{and } R_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix},$$

The attention score V($\boldsymbol{x}$) for each pixel $\boldsymbol{x}$ is computed in the following steps:

1. $k = 0$; $V(\boldsymbol{x}) = 0$
2. Generate a fork S_x of maximum radius ε.
3. Select a random pixel $\boldsymbol{y}$.
4. $k = k+1$; if $k > M$ then stop
5. Transform fork S_x to produce fork S_y

6. If S_y mismatches then $V(\boldsymbol{x}) = V(\boldsymbol{x}) + 1$ and loop to 3 retaining same fork.
7. Loop to 2.

This model possesses many of the characteristics of low level visual neurons with centre-surround receptive fields where cell responses are affected by surrounding cells simultaneously responding to the same signals. Similarities between stimuli within the receptive field and outside inhibit responses, whereas differences do not have this effect [33]. The fork matching and mismatching procedure described above captures aspects of this mechanism.

This model has been shown to identify saliency as indicated in many of the behavioural examples given by Treisman. Saliency determined by a single feature such as colour or orientation is identified strongly by the model. Targets defined by the conjunction of features are identified less strongly and is dependent on the diversity of the distractors [34]. Treisman [29] reports, for example, that a parallel line pair is salient against a background of single lines of various orientations, but the reverse is not the case. This asymmetry is reproduced in our attention model where the attention scores for the pixels belonging to the parallel lines are much greater than for the surrounding randomly oriented single lines. On the other hand the attention scores for the target single line are lower than the background parallel pairs indicating no pop out. Fig. 2 and 3 show the high scoring black pixels in both cases.

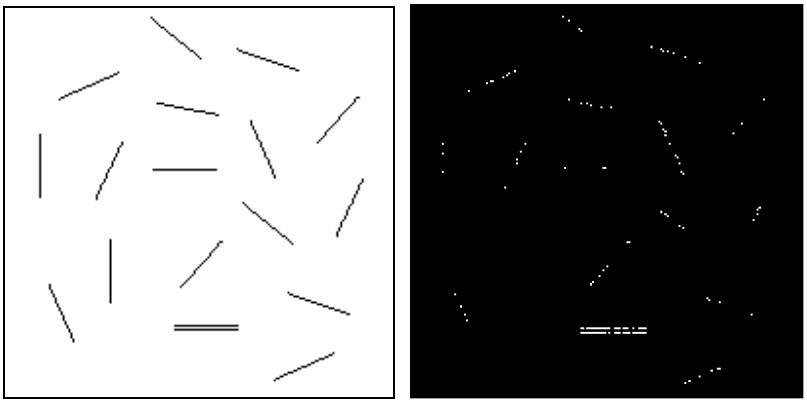

Fig. 2. Target parallel line pair and corresponding high attention scoring pixels

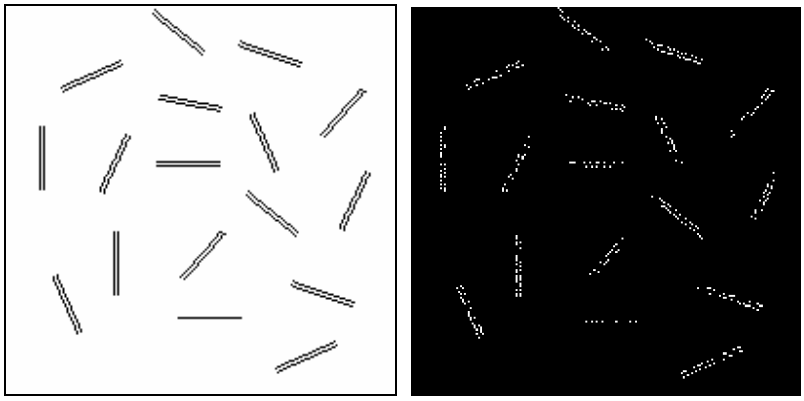

Fig. 3. Target single line and corresponding high attention scoring pixels

4. Similarity Measure

The model of similarity [35] used in this paper is the dual of that for attention and makes the assumption that two images are similar if they possess many features in common. As before no use is made of a priori information in the choice of features. It is to be expected that many features that are shared will be unique to that pair of images. This means that it is never possible to pre-select features in a way that will reflect all aspects of the relationships between a set of objects or the similarity of pairs of images in a database unless complete knowledge of the database is known beforehand. As with the measure of attention described above, the similarity measure proposed here generates features on a competitive basis as part of the calculation of the measure. The features take the form of forks and similarity is dependent upon the number of such forks that can be found to match pairs of images.

Consider two images X and Y containing pixels $\boldsymbol{x}$ and $\boldsymbol{y}$ with colour values $\boldsymbol{a_x} = (a_{x1}, a_{x2}, a_{x3})$ and $\boldsymbol{a_y} = (a_{y1}, a_{y2}, a_{y3})$, respectively.

A fork of m random points S_x is defined as a set of any pixel positions in image X where

$$S_x = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, ..., \boldsymbol{x}_m\}.$$

A randomly positioned fork of m pixels S_y is defined in image Y where

$$S_y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, ..., \boldsymbol{y}_m\} \text{ and } \boldsymbol{x}_i - \boldsymbol{x}_j = \boldsymbol{y}_i - \boldsymbol{y}_j \ \forall \ i,j$$

In other words pixels $\boldsymbol{y}_j$ are a randomly translated version of the $\boldsymbol{x}_j$.

The fork S_y matches image Y if $\left|a_{xj} - a_{yj}\right| \le \delta_j \ \forall \ j$.

The similarity measure R of image X to Image Y is computed in the following steps:

1. $R_{XY} = 0$; $k = 0$
2. Generate a random fork S_x
3. $k = k+1$; if $k > M$ then stop
4. $p = 0$
5. Transform fork S_x to produce fork S_y
6. If S_y matches Y then $R_{XY} = R_{XY} + 1$ and loop to 2.
7. $p = p+1$; if $p > P$ then go to step 9 (generate new S_x)
8. Loop to 4 (no match so try another transform).
9. Loop to step 2
10. Repeat with X and Y reversed. Similarity measure $R = R_{XY} + R_{YX}$

High similarity scores are obtained when many forks in image X are found to correspond and match transformed forks in image Y. If all the forks match a maximum score of M will be obtained.

It is necessary for a high score that forks in image Y correspond to forks in image X, otherwise, for example, cropped versions of images would be classified as being very similar to the whole, although this may be a requirement in some applications. The similarity score R is therefore made symmetric by summing the number of matching forks in both directions.

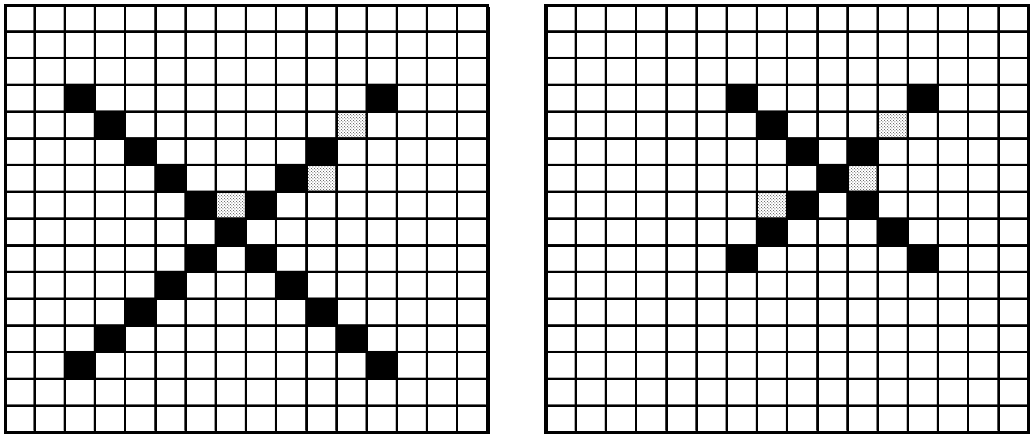

Image X Image Y

Fig. 4. Fork from Image X matching Image Y.

Fig. 4 shows a 3 pixel fork covering one black pixel and two white pixels from image X matching in a position in image Y. In general each of the pixels will possess three colour parameters, so $\boldsymbol{a(x)} = (r, g, b)$ and the fork will fit image Y if the colour parameters of all m corresponding pixels have values within δ of each other. A series of M forks from image X are tested for a match at some position in image Y and the number of such matches forms the similarity measure of image X to image Y.

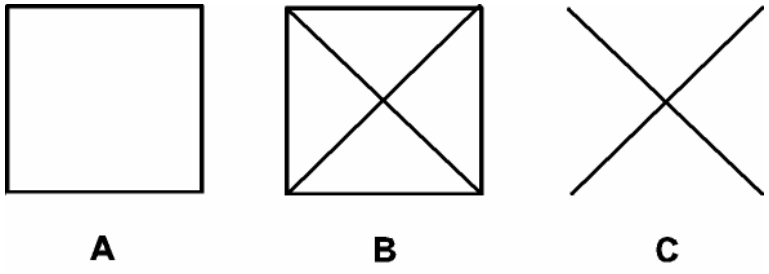

Fig. 5. Shapes possessing similarity measures that do not satisfy the triangle inequality.

The measure does not necessarily satisfy the triangle inequality and is therefore not a metric. In practice there is no requirement for the triangle inequality to be satisfied in terms of human perception and indeed this may be a damaging constraint on the distance measure. A pattern B may contain features in common with A and C and be similar to both. However, A and C may have no features in common at all, and no conclusion regarding A and C can necessarily be inferred from the similarities of AB and BC (Fig. 5).

4.1. Translation, Rotation and Scale

The similarity algorithm makes up to P attempts to discover a match in image Y for each of M transformed forks that are generated. The distribution of locations of high attention scoring pixels in forks that fit both images gives an indication of the location of those regions in both images that hold the features that largely determine the overall image similarity.

The centroids of pixels in images X and Y that are present in the M' pairs of forks $S_x^k, S_y^k\, k = 1, \ldots, M'$ that match is given by $\bar{x}$ and $\bar{y}$ where

$$\bar{x} = \sum_{k, x_i \in S_x^k} {x_i}/{M'} \qquad \bar{y} = \sum_{k, y_i \in S_y^k} {y_i}/{M'} \qquad M \geq M' \succ 0$$

The extent of the region of similarity around $\bar{x}$ and $\bar{y}$ may be estimated by the standard deviation of x_i and y_i where

$$\sigma_x = \sqrt{\frac{1}{M'-1}\left(\sum_i (x_i - \bar{x})^2\right)} \quad \text{and} \quad \sigma_y = \sqrt{\frac{1}{M'-1}\left(\sum_i (y_i - \bar{y})^2\right)}$$

In general an affine relationship between a shape in image X and image Y may be identified by allowing the forks S_x to be randomly rotated between $\pm\alpha$ and scaled in x and y by s to produce the S_y and studying the distribution of angles and scales of pixel co-ordinates of forks that fit both images.

In this case the transform applied to pixels in S_x is given by

$$S_y = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \ldots, \mathbf{y}_m\} \text{ where } \mathbf{y}_i - \mathbf{y}_1 = s\, R_\theta [\mathbf{x}_i - \mathbf{x}_1] \quad \forall\, i$$

$$\text{and } R_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

As with the attention measure described in section 3 there is every justification for tailoring the processing as much as possible before further analysis if sufficient is known about the data to be sure that this improves performance and does not destroy information. This applies to the normalization of image data according to position, orientation and scale which will yield good results if the data is well behaved within the rules followed by the normalisation process. However, as before there is no guarantee that this is the case and a particular normalization process may not be appropriate to all items in a dataset and may cause information to be lost especially in complex and noisy data. The approach described here makes no assumptions about the properties of the data but rather relies upon evidence in the form of mappings (matching forks) discovered between image pixels to reflect orientation, positional and scale relationships between data items.

4.2. Parameter Sensitivity

The behaviour of the similarity measure is clearly dependent on the values of two key parameters, M the number of forks generated to produce a score, and m the number of pixels in each fork, M determines the volume of computation and m the precision with which matches are made.

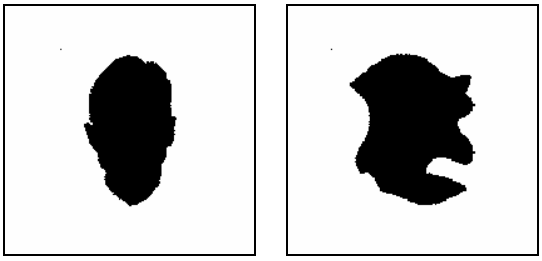

Image X Image Y

Fig. 6. Test image pair X, Y

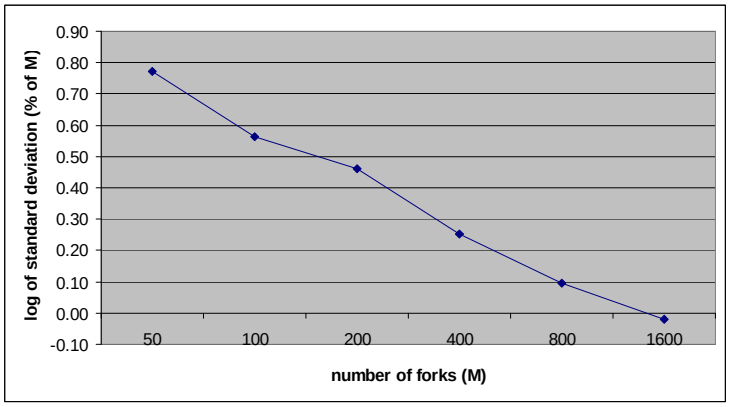

Fig. 7. Variation in standard deviations of similarity measure against no. forks generated.

M	50	100	200	400	800	1600
Standard deviation(σ)	5.90	3.66	2.90	1.79	1.25	0.95
Mean score	81.64	83.05	83.03	83.33	82.86	83.02
log(σ)	0.77	0.56	0.46	0.25	0.10	-0.02

Table 1. Variation in scores against numbers of forks generated (*M*)

In order to test the stability and the statistical confidence of the similarity measure, a number of repeated calculations were carried out on the same pair of patterns (Fig. 6) for each of several values of *M* with *m = 10*. 100 calculations of the similarity measure were computed for values of *M* from 50 to 1600 and the results are shown in Table 1 and Fig. 7. The standard deviation decreases with *M* approximately as the power of -0.6 in this range indicating that the precision of the similarity measure increases with increasing computation.

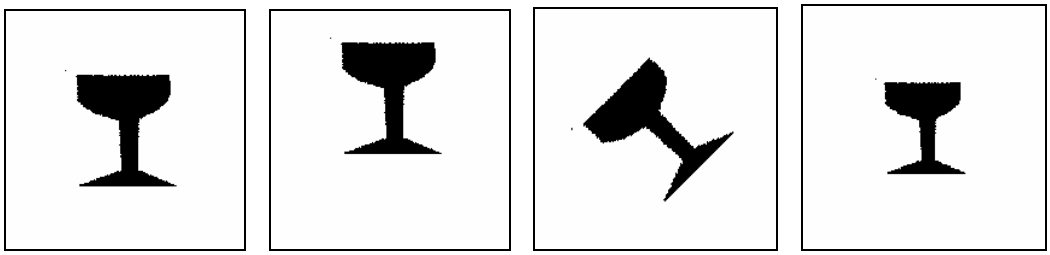

Image X Image Y1 Image Y2 Image Y3

Fig. 8. Original (X), shifted (Y1), rotated (Y2) and scaled (Y3) shapes.

Fig. 8 shows image X transformed by translation (Y1), rotation (Y2) and scale (Y3). Histograms of *x*,*y* values of the centroids of forks fitting image Y1 are shown in Fig. 9 and reflect the upward movement of the shape by about 27 pixels by a corresponding shift to the left of the peak in the *y* histogram. In a similar fashion Fig. 10 shows a histogram of the angles of rotation of forks that fit image Y2 with a peak at +45º which is a good indication of the angular displacement in image Y2. Image Y3 is a 20% isotropic reduction of image X which is again reflected in the histogram of scale values of forks that fit Y3 which has a peak at about 0.8 (Fig. 10). Parameter values used were *m = 10* and *M = 2000*.

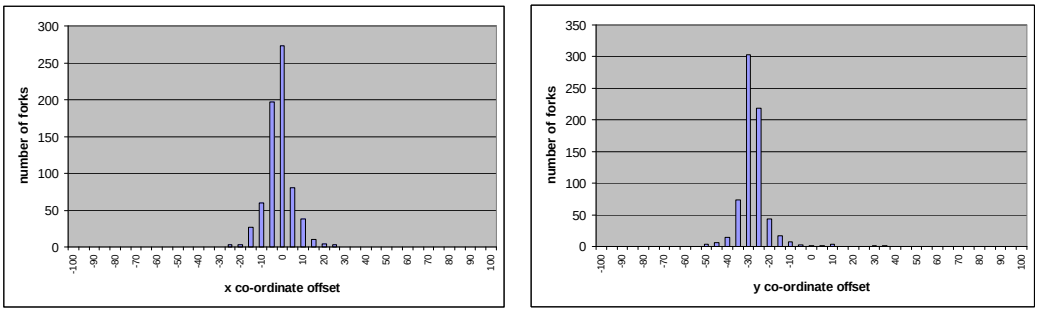

Fig. 9. *x* and *y* histograms of forks fitting image Y1 (*m=10, M = 2000*).

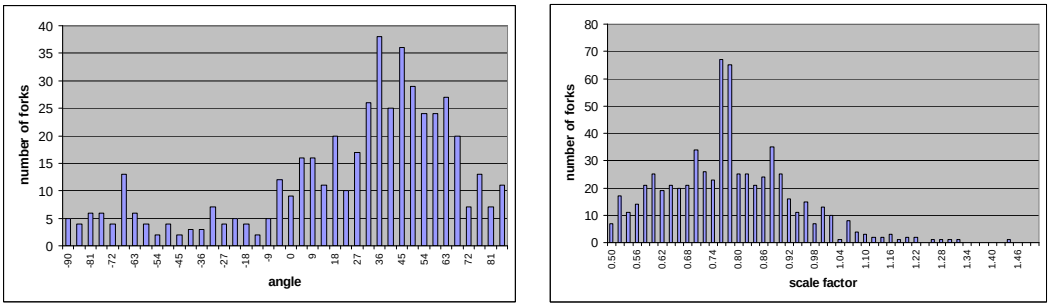

Fig. 10. Rotation and scale histograms of forks fitting images Y2 and Y3, respectively (*m=10, M = 2000*).

The calculations were repeated for several numbers of fork pixels (*m*) on each of the three images with *M=2000* to determine the effect on the stability of the results (Tables 2-4). In all cases the standard deviations of the estimates of position, angle and scale decrease as extra pixels are added to the forks. In addition the peakiness (kurtosis) of the distributions increases with the fork pixel number (*m*). Increased accuracy and precision is achieved at the expense of additional computation which increases with *m*.

m	max	σ	kurtosis
2	-27.4	34.93	0.41
4	-27.2	19.87	4.44
6	-27.2	14.07	10.65
8	-26.6	9.50	12.62
10	-27.1	7.10	20.08
12	-26.9	5.72	43.91

Table 2. Variation in y shift estimate against number of fork pixels (*m*). *M=2000*

m	max°	σ (rad)	kurtosis
4	45	0.86	-1.00
6	45	0.88	-0.92
8	54	0.85	-0.48
10	50	0.75	0.36
12	41	0.72	1.22
14	45	0.69	1.91

Table 3. Variation in angle estimate against number of fork pixels (*m*). *M=2000*

m	max	σ	kurtosis
2	0.76	0.26	-0.72
4	0.78	0.21	0.18
6	0.74	0.19	0.38
8	0.76	0.16	0.69
10	0.77	0.15	0.85
12	0.76	0.14	1.04

Table 4. Variation in scale estimate against number of fork pixels (*m*). *M=2000*

The results show that it is possible to extract relationships between patterns without making specific pre-defined measurements. Location, orientation and size can all be measured indirectly without placing a ruler on the object being measured. Structure that is common between pairs of patterns may be identified by observing the statistics of randomly generated transformed forks that have few rules for their production.

5. Similarity Tests

In order to test how the triangle inequality might be violated by the similarity measure, the three shapes in Fig. 5 were processed. The similarity scores are shown in Table 5. The composite shape B is rated as similar to A and less so to C with scores 374 and 58, respectively. However, A possesses few features in common with C and is given the much lower score of 10. The higher similarity between A and B than between B and C is due to the length of the lines in common.

A	1388		
B	374	1530	
C	10	58	564
	A	B	C

Table 5. Similarity scores of shapes in figure 5.

The similarity calculation may be repeated on the same pair of images, but with greater constraints placed on the transform parameters so that random selections are made in a range that yields the best results. This means that if, for example, a peak in fork fitting frequency is identified at an orientation of 27º, the calculation is repeated selecting forks within the range of rotations $[27° - \alpha, 27° + \alpha]$ where α is a small angle before being applied to image B. In this way rotations are constrained to lie within a smaller range

allowed in the first iteration. In a similar fashion forks may be constrained in location during the second iteration by requiring pixels to lie within a distance D of $(\bar{x}, \bar{y})$ and in effect focusing the computation on the region of highest similarity. These strategies have the effect of enhancing the likelihood of finding forks that fit a pair of similar patterns by focusing attention in the vicinity of the correct transform parameter value.

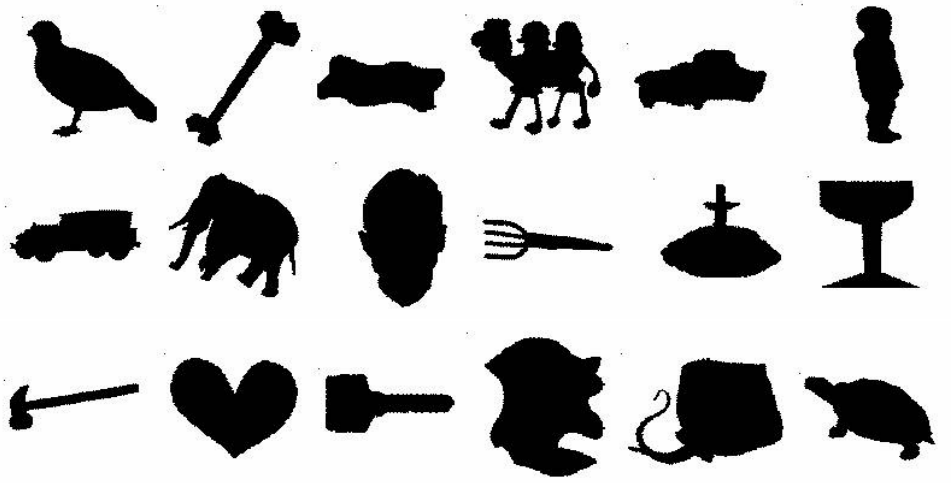

Fig. 11. 18 MPEG-7 shapes.

18 distinct binary shapes (200x200) were taken from the MPEG-7 set [18] (Fig. 11) and the similarity measure R calculated between all pattern pairs with $m=10$ and $M=800$. Forks were allowed to be rotated randomly between $\pm\pi/4$ with D = 150. Patterns matched against themselves obtain the highest scores. Pattern pairs showing strong similarity include (3,7), (5,7), (7,18), and (5,18) (Table 6).

1	968																	
2	47	744																
3	513	41	1038															
4	507	94	452	844														
5	431	73	**753**	406	928													
6	255	169	60	221	67	960												
7	384	71	**734**	369	**859**	58	904											
8	517	68	449	545	443	288	439	1392										
9	281	120	183	475	208	536	174	518	1086									
10	314	12	423	276	412	40	389	244	123	1546								
11	445	41	695	407	699	192	617	465	326	335	988							
12	240	87	302	395	241	276	276	315	**414**	197	169	810						
13	214	42	338	225	371	30	334	213	86	262	238	142	716					
14	511	63	554	569	493	200	482	644	383	300	456	266	251	1394				
15	412	45	568	333	522	91	478	324	157	495	435	156	477	406	1548			
16	336	72	190	470	225	322	210	563	692	148	313	359	130	384	192	1072		
17	415	122	352	469	412	229	371	695	503	202	442	255	154	595	284	519	882	
18	341	115	609	351	**742**	117	**708**	485	248	324	498	219	323	473	473	252	444	940
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**

Table 6. Similarity scores of shapes in Fig. 10.

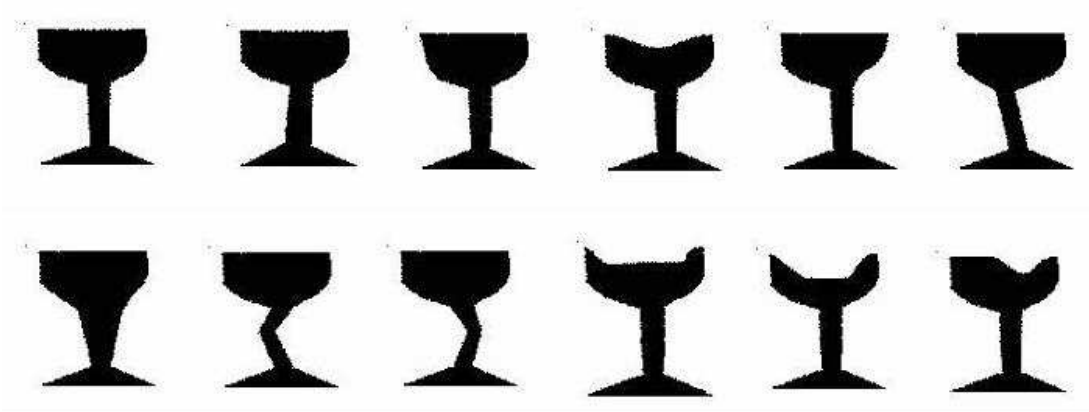

Figure 12. Cluster of similar shapes

1	926											
2	961	996										
3	895	886	746									
4	813	802	901	886								
5	846	785	859	796	896							
6	803	904	892	797	777	918						
7	858	932	876	694	786	810	842					
8	842	882	828	746	766	796	822	1464				
9	880	889	846	764	784	719	784	801	916			
10	827	800	770	734	709	773	749	736	780	712		
11	625	682	589	677	593	632	590	571	**545**	624	1534	
12	939	825	752	843	852	751	750	791	764	755	703	886
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**

Table 7. Similarity scores of shapes in Fig. 11.

To be sure of a successful and rapid search through a database of shapes a hierarchy of such clusters have to be identified in which intra-cluster similarities are much larger than those between different clusters. With this in mind a cluster of very similar shapes to pattern 12 in Fig. 11 shown in Fig. 12 were analysed to determine how well the similarity measure would reflect the tightness of the cluster. Pattern 12 in Fig. 11 is in fact identical to pattern 1 in Fig. 12. Similarity measures were obtained under the same conditions as before and are shown in Table 7. The lowest similarity to pattern 1 with a score of 625 occurs with pattern 11 and significantly exceeds the highest similarity to pattern 12 in Table 6 which took a value of 414 with pattern 9. This means that it would be unlikely for any of the shapes in figure 12 and any other unseen variants to be confused with the wrong 'vantage' shape in Fig. 11.

The similarity scores for the very similar shapes in Fig. 12 are generally much higher than those for the dissimilar shapes in Fig. 11. This reflects the "tightness" of the cluster of similar shapes in comparison with the significant differences indicated in Table 6 by the wide range of scores.

6. Discussion

The results presented here only refer to black and white shapes, but the mechanisms also apply to colour images. Pixel matching for colour requires that all three colour components independently have values within δ_j of each other. However, preliminary experiments indicate that the additional discriminatory information in the colour requires that the number of pixels in forks be reduced from 10 to 3 or 4 to facilitate matching. This also reduces processing time.

A symmetry transform used with the attention model above has been successful in extracting reflective symmetry from colour images [36]. This suggests that additional invariance in the similarity measure may be obtained when comparing shapes that are similar with respect to a given operation. Invariance under different illuminations is a further possibility building upon results using a colour transform in a similar fashion to obtain colour constancy in images [37]. Extracting such parameters simultaneously by applying all the transforms to each fork before testing for a fit may require an impractical increase in the number of forks generated (M) to obtain statistical significance. On the other hand extracting the parameters separately is consistent with the understanding that early visual analysis results in separate maps for separate properties [29]. These pre-attentive properties include orientation [38], size [39], and colour [40] which do not perform well in conjunction.

The "curse of dimensionality" is avoided in this approach because the similarity measure is not dependent upon vectors derived from a n-dimensional feature space. Instead features specific to each pattern pair are extracted as part of the similarity calculation itself and these features may bear no relation to other patterns in the database. On the other hand conventional approaches are forced to limit the numbers of features either because of constraints on the size of training sets, or because of decreasing performance due to the lack of independence amongst larger feature sets [41].

The approach to measuring similarity is based on a trial and error process of generating features (forks) that are present in pairs of patterns. No guidance is given during this process as this would inevitably damage the performance on certain classes of shapes by precluding relevant features from being produced. The negative effects of such guidance cannot be predicted in advance of experiment and would not be discovered until sufficient data is explored.

The matching of forks can be carried out in parallel as each operation is independent of the next. However, although the computational steps are very simple there are a large number of them and a shape comparison takes about 1 second on a 1.8GHz machine running in C++. In fact the computations are inappropriate for a serial machine which is designed to carry out fast arithmetic operations with little opportunity to harness parallel architecture.

The human vision system possesses a clever attention mechanism that protects us from danger and gives considerable evolutionary advantage. It is tempting to speculate that those features that distinguish foreground from background are remembered and help us to recognise important objects in a scene at a later date. Of course fresh scenes do not appear exactly as they did in the past and it is necessary to match new inputs onto stored data in ways that preserve the notion of similarity. Desimone at al [42] suggest that this is achieved by favouring inputs competitively that match the description of the information currently needed. It is also interesting to suggest that a form of surround

suppression operates in time and space throughout the brain. Indeed there is no reason to believe that different parts of the brain including vision use different computational tools [43]. Signals in the cortex that vary may be suppressed locally if they constitute a background that possesses a certain self-similarity. On the other hand signals that are locally anomalous are not so inhibited. Signals that create peaks of activity by resonating with a stored response that was retained as a result of earlier anomalous activity will again be anomalous and the activity will be reinforced by the surround suppression and represent an act of recognition.

7. Conclusions

This paper has described a measure of similarity between patterns that does not require the prior specification of features on which that measure is based. The approach is derived from an attention model that identifies saliency by comparing regions in a single image and recognising differences. Similarity is measured by applying the dual operation to two images and seeking features in common. The similarity calculation represents an unsupervised learning process in which pattern clusters are identified without requiring pattern class labels or numbers of clusters. Large numbers of features are generated without guidance and their presence in pairs of patterns determines the similarity measure.

More results are needed to prove the effectiveness of the measure for shape indexing and retrieval in a large database. However, the approach has the significant advantages that training sets are not required and problem solutions are not precluded by the *a priori* selection of features. It is shown that this measure is stable and is not constrained by the triangle inequality which can prevent proper discrimination. Greater precision may be obtained with more computation using larger forks or increased numbers of forks. The measure provides a plausible mechanism for comparing shapes and other patterns which focuses computation on optimising transform parameters that expose similarity.

Future work will be directed at larger sets of shapes and natural images where it is proposed that attention based similarity measures can be used in Content Based Image Retrieval and copy detection applications.

Acknowledgements

This research has been conducted with the support of BT Research and Venturing and within the framework of the European Commission funded Network of Excellence "Multimedia Understanding through Semantics, Computation and Learning" (MUSCLE) [44].

References

[1] Y. Rui, T.S. Huang, S.F. Chang, Image retrieval: current techniques, promising directions and open issues, J. Visual Communication and Image Representation, 10(1), (1999) 39 – 62.

[2] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Mach. Intell. 22(12) (2000) 1349-1379.

[3] R. C. Veltkamp, M. Tanase, Content-based image retrieval systems: a survey, March 2001, http://www.aa-lab.cs.uu.nl/cbirsurvey/cbir-survey/.

[4] E. Izquierdo, V. Mezaris, E Triantafyllou, L-Q. Xu, State of the art in content-based analysis, indexing and retrieval, IST Project IST-2001-32795, Network of Excellence in Content-Based Semantic Scene Analysis and Information Retrieval, Sept. 2002, http://www.iti.gr/SCHEMA/library/index.html

[5] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, Query by image and video content: the QBIC system, IEEE Computer (1995).

[6] A. Pentland, R. W. Pickard, S. Sclaroff, Photobook: content-based manipulation of image databases," Proceedings of SPIE Storage and Retrieval of Images and Video Databases, Vol. 2185, San Jose, CA, USA, 1994.

[7] C. Carson, S. Belongie, H. Greenspan, J. Malik, Blobworld: image segmentation using expectation-maximisation and its application to image querying," IEEE Trans. Pattern Anal. Mach. Intell. 24(8) (2002) 1026-1038.

[8] J.R. Smith, S-F. Chang, VisualSEEk: a fully automated content-based image query system, Proceedings of the ACM International Conference on Multimedia, Boston MA, USA, 1996, pp 87-98.

[9] W-Y. Ma, B.S. Manjunath, NeTra: a toolbox for navigating large image databases," Multimedia Systems, 7 (1999), 184-198.

[10] A. Gupta, R. Jain, Visual information retrieval, Communications of the ACM, 40(5) (1997) 70-79.

[11] D. Vernon, Cognitive vision – the development of a discipline, European Research Network for Cognitive AI-enabled Computer Vision Systems - KA4 - IST-2001-35454, ECVision project report, August 2004.

[12] J.W.H. Tangelder, R.C. Veltkamp, A survey of content based 3D shape retrieval methods, Proceedings of Shape Modelling International Conference, 2004, pp. 145-156.

[13] D. Zhang and G. Lu, Review of shape representation and description techniques, Pattern Recognition, 37 (2004) 1-19.

[14] R Manmatha, S. Ravela and Y. Chitti, On computing local and global similarity in images, Proceedings of SPIE Human and Electronic Imaging III, 1998.

[15] K. Mikolajczyk, A. Zisserman, C. Schmid, Shape recognition with edge-based features, Proceedings of the British Machine Vision Conference, Norwich, UK, 2003.

[16] W.A. Almageed, C. Smith, Hidden Markov models for silhouette classification, Proceedings of the World Automation Congress, Orlando, 2002.

[17] M-S Dao, F.G.B. De Natale, A. Massa, Edge potential functions and genetic algorithms for shape-based image retrieval, Proceedings of the International Conference on Image Processing, Barcelona, 2003.

[18] T.B Sebastian, P.N. Klein B.B. Kimia, Recognition of shapes by editing shock graphs, Proceedings of the International Conference on Computer Vision, Vancouver, 2001, pp. 755-762.

[19] N. Iyer, S. Jayanti, K. Lolu, Y. Kalyanaraman, K. Ramani, A multi-scale hierarchical 3D shape representation for similar shape retrieval, Proceedings of the International Symposium on Tools and Methods of Competitive Engineering, Lausanne, 2004.
[20] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell., 24(4) (2002) 509-522.
[21] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, Proceedings of Computer Vision and Pattern Recognition, Vol. 1, 2001, pp 511-518.
[22] A. Mojsilovic, J. Hu, E. Soljanin, Extraction of perceptually important colors and similarity measurement for image matching, retrieval, and analysis, IEEE Trans. on Image Processing, 11(11) (2002) 1238-1248.
[23] B. J. Super, Fast correspondence-based system for 2-D shape classification, Pattern Recognition Letters, 25(2) (2004) 217-224.
[24] M.H.C. Law, M.A.T. Figueiredo, A.K. Jain, Simultaneous feature selection and clustering using mixture models, IEEE Trans. on Pattern Anal. Mach. Intell. 26(9) (2004) 1154-1166.
[25] L. Itti, Automatic foveation for video compression using a neurobiological model of visual attention, IEEE Trans. on Image Processing, 13(10) (2004) 1304-1318.
[26] S. Frintrop, A. Nuchter, H. Surmann, Visual attention for object recognition in spatial 3D data, Proceedings of IEEE International Conference on Intelligent Robots and Systems, Sendai, Japan, 2004.
[27] C. Grigorescu, N. Petkov, M.A. Westenberg, Contour detection based on nonclassical receptive field inhibition, IEEE Trans. on Image Processing, 12(7), (2003) 729-739.
[28] R. Desimone, Visual attention mediated by biased competition in extrastriate visual cortex, Phil. Trans. R. Soc. Lond. B, 353, (1998) 1245 – 1255.
[29] A. Treisman, “Preattentive processing in vision,” in “Computational Processes in Human Vision: an Interdisciplinary Perspective,” Ed. Z. Pylyshyn, Ablex Publishing Corp., Norwood, New Jersey, 1988.
[30] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis and F. Nuflo, “ Modeling visual attention via selective tuning,” Artificial Intelligence, 78, pp 507-545, 1995.
[31] T. Lindeberg, “Detecting salient blob-like image structures and their scales with a scale-space primal sketch: a method for focus-of-attention,” Int. J. of Computer Vision, 11:3, 1993.
[32] F.W.M. Stentiford, An estimator for visual attention through competitive novelty with application to image compression,” Proceedings of Picture Coding Symposium, Seoul, Korea, 2001.
[33] H. E. Jones, K. L. Grieve, W. Wang, and A. M. Sillito, “Surround suppression in primate V1,” Journal of Neurophysiology, vol. 86, no. 10, pp 2011-2028, 2001.
[34] F. W. M. Stentiford, “The measurement of the salience of targets and distractors through competitive novelty,” 26th European Conference on Visual Perception, Paris, September 1-5, 2003.
[35] F.W.M. Stentiford, An attention based similarity measure with application to content based information retrieval, Proceedings of SPIE Storage and Retrieval for Media Databases, Vol 5021, Santa Clara, CA, USA, 2003.
[36] F W M Stentiford, “Attention based symmetry detection in Colour Images,” IEEE Workshop on Multimedia Signal Processing, Shanghai, October 30-November 2, 2005.
[37] F. W. M. Stentiford, “Attention based colour correction,” SPIE Human Vision and Electronic Imaging XI, San Jose, Jan. 2006.

[38] J. M. Wolfe, S. R. Priedman-Hill, M. I. Stewart, and K. M. O'Connell, "The role of categorization in visual search for orientation," Journal of Experimental Psychology: Human Perception and Performance, 18(1), PP 34-49, 1992.

[39] A. Treisman and G. Gelade, "A feature integration theory of attention," Cognitive Psychology, 12, pp 97-136, 1980.

[40] S. Engel, X. Zhang, and B. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," Nature, 388(6637), pp 68-71, 1997.

[41] B. Chandrasekaran, Independence of measurements and the mean recognition accuracy, IEEE Trans. Inform. Theory 17(4) (1971) 452-456.

[42] R. Desimone, J. Duncan, Neural mechanisms of selective visual attention, Ann. Rev. of Neuroscience, 18, (1995) 193-222.

[43] V. B. Mountcastle, An organising principle for cerebral function: the unit model and the distributed system," in: G.M. Edelman, V. B. Mountcastle (Eds.), The Mindful Brain, MIT Press, Cambridge, Mass, 1978.

[44] Multimedia Understanding through Semantics, Computation and Learning, Network of Excellence, EC 6th Framework Programme, FP6-507752, http://www.muscle-noe.org/general.html.