An Attention Based Similarity Measure for Colour Images

Li Chen and F. W. M. Stentiford
University College London, Adastral Park Campus, UK.
{l.chen, f.stentiford}@adastral.ucl.ac.uk

Abstract. Much effort has been devoted to visual applications that require effective image signatures and similarity metrics. In this paper we propose an attention based similarity measure in which only very weak assumptions are imposed on the nature of the features employed. This approach generates the similarity measure on a trial and error basis; this has the significant advantage that similarity matching is based on an unrestricted competition mechanism that is not dependent upon a priori assumptions regarding the data. Efforts are expended searching for the best feature for specific region comparisons rather than expecting that a fixed feature set will perform optimally over unknown patterns. The proposed method has been tested on the BBC open news archive with promising results.

1 Introduction

Similarity matching is a basic requirement for the effective and efficient delivery of media data and for the identification of the infringement of intellectual property rights. Considerable effort has been devoted to defining and extracting image signatures, which are based on the assumption that similar images will cluster in a pre-defined feature space [1-5]. It is common for unseen patterns not to cluster in this fashion despite apparently possessing a high degree of visual similarity.

In this research we propose an attention based similarity matching method with application to colour images based on our previous work [6,7]. The approach computes a similarity measure on a trial and error basis; this has the significant advantage that features that determine similarity can match whatever image property is important in a particular region whether it is a colour, texture, shape or a combination of all three. Efforts are expended searching for the best feature for the region rather than expecting that a fixed feature set will perform optimally over unknown patterns in addition to the known patterns. In this context, the proposed method is based on the competitive evolution of matching regions between two images rather than depending on fixed features which are intuitively selected to distinguish different images or cluster similar images. In addition the proposed method can cope with different distortions of images including cropping, resizing, additive Gaussian noise, illumination shift and contrast change. These functions potentially help detect copied images.

The remainder of this paper is arranged as follows. In Section 2, the cognitive visual attention model is presented. Experiments are conducted on BBC open news archive and results are shown in Section 3. Conclusions are addressed in Section 4.

2 Visual Attention Similarity Measure

Studies in neurobiology [8] suggest 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. In this context of visual attention, we argue that humans are able to spot anomalies in a single image or similarity between two images through a competitive comparison mechanism, where similar and dissimilar regions are identified and scored by means of a new similarity measure.

Our model of visual attention is based upon identifying areas in an image that are similar to other regions in that same image [6,7]. The salient areas are simply those that are strongly *dissimilar* to most other parts of the image. In this paper we apply the same mechanism to measure the similarity between two different images. The comparison is a flexible and dynamic procedure, which does not depend on a particular feature space which may be thought to exist in a general image database.

Let a measurement $a = (a_1, a_2, a_3)$ correspond to a pixel $x = (x_1, x_2)$ in image A and a function *F* is defined so that *a = F(x).*

Consider a neighbourhood N of x where

$$N = \{x' \in N \ \ if\ and\ only\ if\ \left|(x_{\mathrm{i}} - x'_{\mathrm{i}}) \le \varepsilon_{\mathrm{i}}\right|\}.$$

Select a set (called a fork) of *m* random pixels S_A from *N* where

$$S_A = \{x'_1, x'_2, ..., x'_{\mathrm{m}}\}\ .$$

Select another random pixel *y* in image B and define the fork S_B

$$S_B = \{y'_1, y'_2, ..., y'_{\mathrm{m}}\}\ \text{where}\ x - x'_{\mathrm{i}} = y - y'_{\mathrm{i}}\ \ \forall\ i\ .$$

The fork S_A matches S_B if

$$\left|F_j(x_i) - F_j(x'_i)\right| \le \delta_{\mathrm{j}} \quad \forall\, i, j\ .$$

That is, a match occurs if all colour values (suffix *j*) of corresponding pixels in S_A and S_B are close. The similarity score of a pixel *x* is incremented each time one of a set of *M* neighbourhoods S_A matches a neighbourhood S_B surrounding some *y* in pattern B. This means that pixels *x* in A that correspond to large numbers of matches between a range of *M* neighbouring pixel sets S_A and pixel neighbourhoods somewhere in B are assigned high scores. In Fig. 1, *m = 3* pixels *x'* are selected in the neighbourhood of a pixel *x* in pattern A and matched with 3 pixels in the neighbourhood of pixel *y* in pattern B.

A parameter *s* is introduced to limit the area in pattern B within which the location *y* is randomly selected. *s = 2* defines the dotted region in Fig. 1. This improves the efficiency of the algorithm in those cases where it is known that corresponding regions in the two images are shifted by no more than *s* pixels. In effect *s* represents the maximum expected mis-registration or local distortion between all parts of the two images.

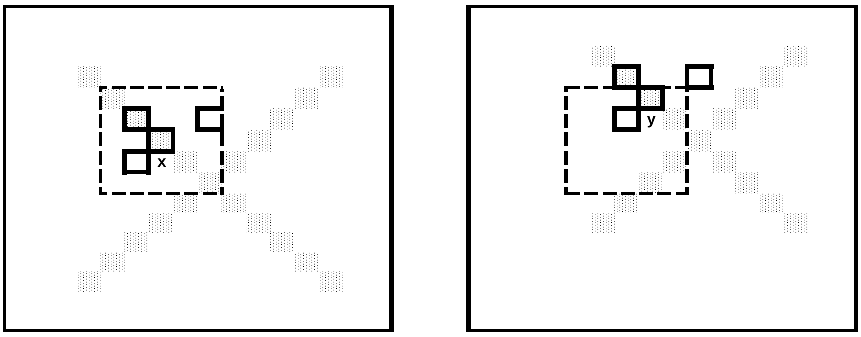


Fig. 1. Neighbourhood at location *x* matching at location *y*

The similarity contributions from all pixel regions in A are summed and normalized to give the total similarity score C_{AB} between images A and B:

$$C_{AB} = \frac{1}{M * \|A\|} \sum_{x \in A} (\sum_{M, y \in B} (1 \,|\, S_A \; matches \; S_B, \; 0 \,|otherwise)$$

3 Experiments

Experiments were carried out on images downloaded from BBC open news archives [9]. 75 videos (more than 230,000 frames) cover different topics including conflicts and wars, disasters, personalities and leaders, politics, science & technology, and sports. Since many frames within a scene differ only slightly, and to utilise the diversity in the database, 2000 non-contiguous frames were extracted by taking every 100th frame from these videos to form the database for image retrieval. 21 images were randomly chosen from the database and 4 distorting transforms were applied to each image (see Fig. 2) including additional Gaussian noise, contrast change, crop and shift, and resize. These distorted images were then added to the image database making a total of 2084 images.

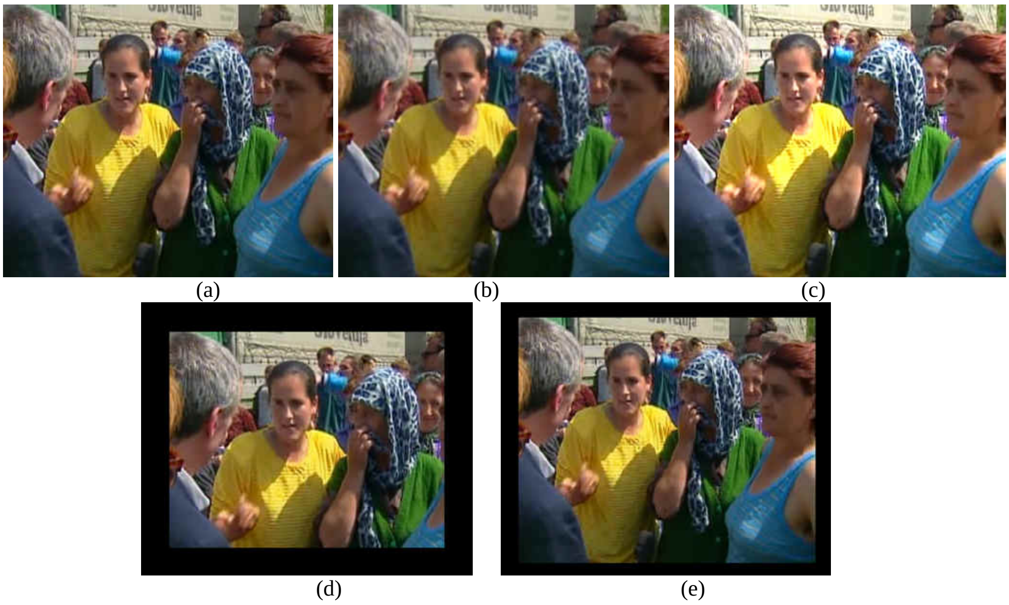
Fig. 3 illustrates the precision and recall performance of the proposed method with 15 queries of the database and M = 20. Recall is the ratio of the number of relevant images retrieved to the total number of relevant images in the database; and it is expressed as:

$$recall = \frac{the \;\; number \;\; of \;\; relevant \;\; images \;\; retrieved}{the \;\; number \;\; of \;\; relevant \;\; images \;\; in \;\; the \;\; database}$$

Precision is the ratio of the number of relevant images retrieved to the total number of irrelevant and relevant images retrieved, and it is defined as:

$$precision = \frac{the\ \ number\ \ of\ \ relevant\ \ images\ \ retrieved}{the\ \ number\ \ of\ \ images\ \ retrieved}$$

Only very similar frames that were immediately adjacent in time to the query frame were considered to be relevant images.



(a) (b) (c)

(d) (e)

Fig. 1. An example of an image and four transforms: (a) original image (b) with additional radius-1 Gaussian blur (c) with contrast increased 25% (d) cropped and shifted to the right (e) resized to 80% of original image

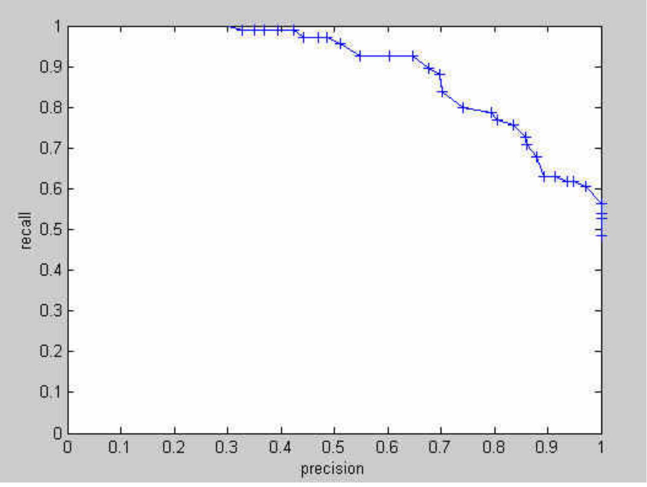
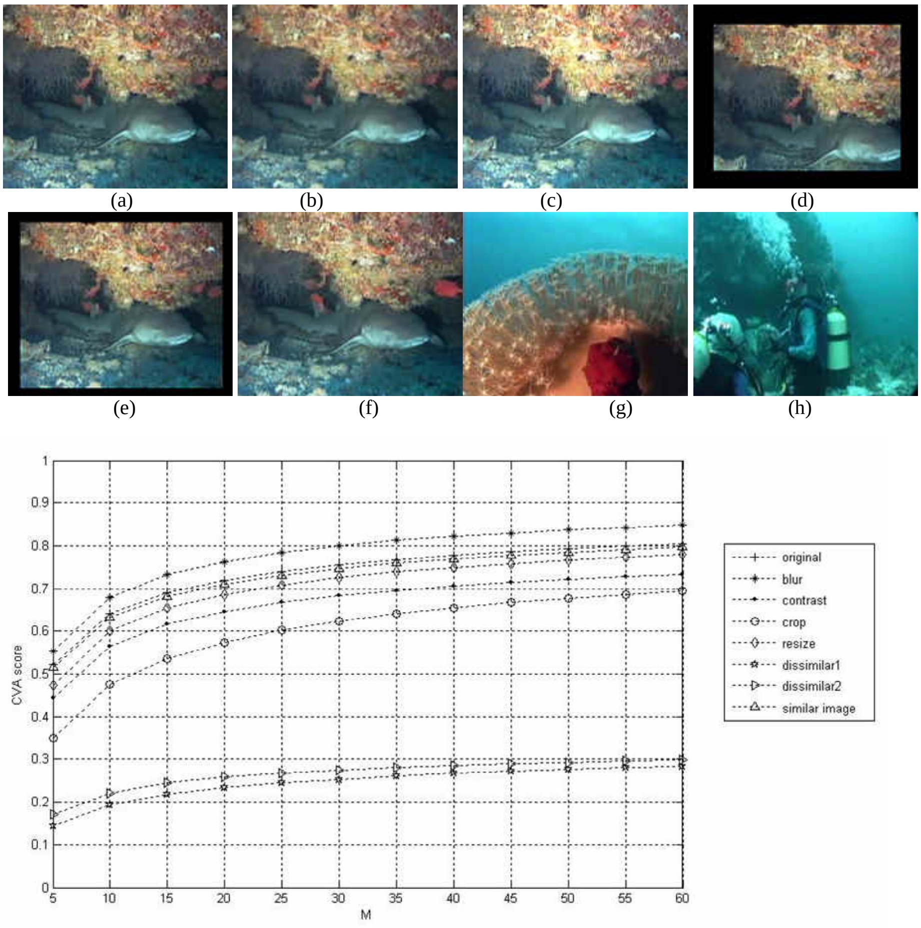


Fig. 2. Recall and precision for retrieval performance

Fig. 4 shows the relationship between the similarity score and the computation (*M*) for the original image when compared with itself, the blurred, contrast shifted, cropped and resized versions, a similar image taken from the same video ahead of example frame by 70 frame distance, and two other different images in the database. It is interesting that for low values of *M* the original is seen to be less similar to itself than to the distorted versions. This repeated an earlier result [6] when it was found that some similarities were found more easily in blurred images.


(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 3. CVA scores against computation (M). (a) original image (b) image with Gaussian blur (c) image with 25% contrast increase (d) cropped image (e) resize down to 80% of original image (f) similar image ahead of original frame by 70 frame distance (g) and (h) dissimilar images taking from the same video

The approach is further illustrated in Fig. 5 where image (a) has been pasted into another image giving a composite version (b). Image (c) shows the fork pixel locations where matching has taken place during the computation of the similarity score between images (a) and (b). This indicates that the mechanism is potentially able to detect sub-images with application to copy detection.

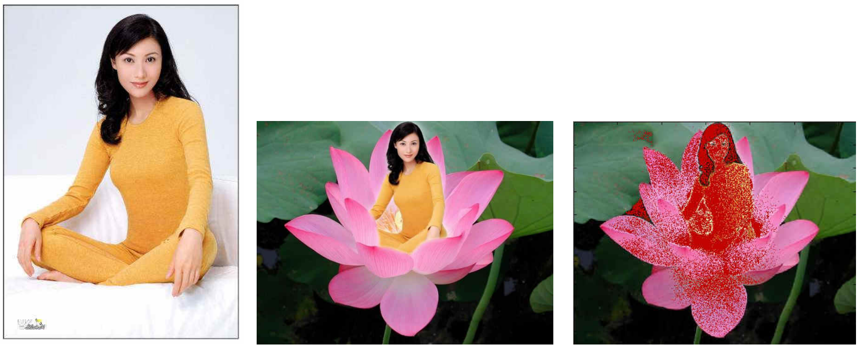


Fig. 4. (a) Image, (b) composite,(c) matching fork pixel locations

4 Conclusions

This paper has shown that a new similarity measure that is not based on pre-selected feature measurements can be used to obtain promising retrieval performance. The similarity is determined by the amount of matching structure detected in pairs of images. Such structure that is found to be in common between specific pairs of images may not be present elsewhere in the database and would be unlikely to be taken into account by a fixed set of features applied universally. The work also provides evidence in support of a mechanism that encompasses notions of both visual attention and similarity.

More results are needed to obtain statistical significance in the precision and recall performances.

Acknowledgement

This research has been conducted within the framework of the European Commission funded Network of Excellence “Multimedia Understanding through Semantics, Computation and Learning” (MUSCLE) [10].

References

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

[2] Fu, H., Chi, Z., Feng, D.: Attention-driven image interpretation with application to image retrieval, Pattern Recognition, 39, no. 7 (2006)

[3] Itti, L.: Automatic foveation for video compression using a neurobiological model of visual attention, IEEE Trans. on Image Processing, 13(10) (2004) 1304-1318

[4] Treisman, A.: Preattentive processing in vision. In: Pylyshyn, Z. (ed.): Computational Processes in Human Vision: an Interdisciplinary Perspective, Ablex Publishing Corp., Norwood, New Jersey, (1988).

[5] Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning, Artificial Intelligence, 78 (1995) 507-545

[6] Stentiford, F.W.M.: An attention based similarity measure with application to content based information retrieval. In Yeung, M.M., Lienhart, R.W., Li, C-S.(eds.): Storage and Retrieval for Media Databases, Proc SPIE Vol. 5021 (2003) 221-232

[7] Stentiford, F.W.M.: Attention based similarity, Pattern Recognition, in press, 2006.

[8] Desimone, R.: Visual attention mediated by biased competition in extrastriate visual cortex, Phil. Trans. R. Soc. Lond. B, 353 (1998) 1245 – 1255

[9] http://creativearchive.bbc.co.uk/

[10] Multimedia Understanding through Semantics, Computation and Learning, EC 6th Framework Programme. FP6-507752. (2005) http://www.muscle-noe.org/