AN ATTENTION BASED SIMILARITY MEASURE USED TO IDENTIFY IMAGE CLUSTERS

A. Bamidele*, F.W.M. Stentiford†

Content Understanding Group, University College London, Adastral Park Campus, UK.
E-MAIL: a.bamidele@ee.ucl.ac.uk*, f.stentiford@ee.ucl.ac.uk †

Keywords: visual attention, similarity measure, clustering, pattern recognition, image classification.

Abstract

This paper outlines a new attention based similarity measure and describes an application to the problem of identifying image clusters in a 4 class problem. A diverse set of images was obtained using camera phones in 4 separate locations and classification performance was tested against the true location of the images. The approach promises to have application to the unsupervised extraction of unknown numbers of clusters in larger datasets.

1 Introduction

The volume of media gathered by digital cameras, camcorders and camera phones is increasing dramatically. Whilst storage and image capture technologies are able to cope with huge numbers of images, poor image and video retrieval is in danger of rendering many repositories valueless because of the difficulty of access. Many disciplines and segments in industry including telecommunications, entertainment, medicine, and surveillance, need high performance retrieval systems to function efficiently. Visual searches by text alone are often ineffective on images and are haphazard at best. Descriptive text simply does not reflect the capabilities of the human visual memory and does not satisfy users' expectations. Furthermore the annotation of visual data for subsequent retrieval is almost entirely carried out through manual effort. This is slow, costly and error prone and presents a barrier to the stimulation of new multimedia services.

Similarity measures are central to most pattern recognition problems not least in computer vision and the problem of categorising and retrieving large number of digital images. These problems have motivated considerable research into content based image retrieval [13, 21]. There are many approaches to similarity and pattern matching and much of this is covered in several survey papers [22].

CBIR systems normally rank the relevance between a query image and target images according to a similarity measure based on a set of features. The pre-determined features can take the form of edges, colour, location, and texture functions dependent on pixel values [14], shape measures [17], segmented regions [5], interest points [19], and many other intuitively appealing aspects. Subsequent processing can include texture histograms, color histograms and discriminant analysis [9, 11, 24]. Mikolajczyk et al [15] employed the use of edge models to obtain correspondences with similar objects. The advantages and disadvantages of using 3D colour histograms in which bins represent location are investigated by Ankerst et al [1].

Recent research in human perception of image content [16] suggests the importance of semantic cues for efficient retrieval. One method of interpreting human intention uses relevance feedback mechanisms [7]. Relevance feedback is often proposed as a technique for overcoming many of the problems faced by fully automatic systems by allowing the user to interact with the computer to improve retrieval performance. This reduces the burden on unskilled users to set quantitative pictorial search parameters or to select images that come closest to meeting their goals.

Vailaya et al. [25] organizes vacation images into a hierarchical structure of semantically meaningful classes by measuring the saliency of low-level features based on the plot of the inter-class and intra-class distance distributions; vacation images are assigned to classes using a k-nearest neighbor classifier. At the top level, images are classified as indoors or outdoors. Outdoor images are then classified as city or landscape and are further divided into sunset, forest, and mountain classes. Statistical classification methods are applied to low level visual features to derive categories [18, 6]. For example, The SemQuery system [20] categorizes images into different clusters based on their heterogeneous features. The SIMPLIcity system [26] classifies images into graph, textured photograph, or non-textured photograph, and thus narrows down the searching space in a database. Although these classification methods are successful in their specific domains of application, difficulties arise if further unseen material is introduced.

More recently [2], attributes derived from a combination of probable contextual metadata and image similarity have been used to indicate the location at which camera phone images were taken. In this paper we address the problem of image classification by grouping images into visual clusters according to a new similarity measure [23]. The method is applied to a very diverse set of images gathered by camera phones. The algorithms are described in the next section and results and discussion reported in the following sections.

1.1 Cognitive visual attention

Studies in neurobiology and computer vision [8, 12] 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.

Such a mechanism has been explored [10] and extended to apply to the comparison of two images in which attention is drawn to those parts that are in common rather than their absence as in the case of saliency detection in a single image [3]. Whereas saliency measures require no memory of data other than the image in question, cognitive attention makes use of other stored material in order to determine similarity with an unknown image. This has been further explored in [4]; the approach employs the use of a visual attention model to attach more importance to image regions that are visually salient. The work has indicated that laying emphasis upon areas of images that attract high visual attention can improve retrieval performance. The model of Cognitive Visual Attention (CVA) used in this paper relies upon the matching of large numbers of pairs of pixel groups (forks) taken from patterns A and B under comparison. Let a location $\boldsymbol{x}$ in a pattern correspond to a measurement $\boldsymbol{a}$ where

$$\boldsymbol{x} = (x_1, x_2) \text{ and } \boldsymbol{a} = (a_1, a_2, a_3)$$

Define a function $\boldsymbol{F}$ such that $\boldsymbol{a} = \boldsymbol{F}(\boldsymbol{x})$. Select a fork of m random points S_A in pattern A where

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

Similarly select a fork of m points S_B in pattern B where

$$S_B = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, ..., \boldsymbol{y}_m\} \qquad (3) \text{ Where } \boldsymbol{x}_i - \boldsymbol{y}_i = \boldsymbol{\delta}_j.$$

The fork S_A matches the fork S_B if

$$|\boldsymbol{F}(\boldsymbol{x}_i) - \boldsymbol{F}(\boldsymbol{y}_i)| < \boldsymbol{\varepsilon} \quad \forall i \text{ for some } \boldsymbol{\delta}_j \; j = 1,2, \ldots, N$$

In general ε is not a constant and will be dependent upon the measurements under comparison i.e.

$$\varepsilon_j = f_j(\boldsymbol{F}(\boldsymbol{x}), \boldsymbol{F}(\boldsymbol{y}))$$

In effect up to N selections of the displacements $\boldsymbol{\delta}_j$ apply translations to S_A to seek a matching fork S_B. The CVA similarity score C_{AB} is produced after generating and applying T forks S_A :

$$C_{AB} = \sum_{i=1}^{T} w_i \quad where \quad w_i = \begin{cases} 1 & if\ S_A\ matches\ S_B \\ 0 & otherwise \end{cases}$$

C_{AB} is large when a high number of forks are found to match both patterns A and B and represents features that both patterns share. It is important to note that if C_{AC} also has a high value it does not necessarily follow that C_{BC} is large because patterns B and C may still have no features in common. The measure is not constrained by the triangle inequality.

1.2 Image classification

The similarity values obtained in this way may be used to drive a nearest neighbour classifier in which relative similarities with class representatives or exemplars determine the classification decisions. The selection of representative images or exemplars that characterise the pattern class may be carried out in many different ways.

The most straightforward selection is the visual centre of gravity (centroid) i.e. the pattern G_I to which all others in the class i are most similar, or rather the pattern with which all others in the class share most features in common (matching forks).

$$G_I = \max_{P \in I}{}^{-1} \sum_{Q \in I, P \neq Q} C_{QP} \qquad (1)$$

The classification $Cl(U)$ of an unknown pattern U is then given by

$$Cl(U) = \max_{R} C_{UG_R} \qquad (2)$$

Here $Cl(U)$ identifies the class exemplar that shares the most features with the unknown pattern. It is important to emphasise that pattern separations are not being measured in a conventional feature space in which the features are fixed and extracted in a similar fashion from all patterns. Instead different features are identified for each specific pattern comparison as an integral part of the process of calculating the similarity measure. This approach avoids many of the problems which have to be faced when dealing with high dimensional feature spaces.

1.3 Visual sub-cluster extraction

The selection of an exemplar from a class of training images given yields an exemplar that possesses most features in common with other images in that class. However, the diversity of images that are captured at each location means that the class is probably represented more realistically by several sub-clusters that contain specific similarities within themselves. Adding more exemplars in a conventional feature space does not necessarily guarantee improvements in classifier performance because although some errors are corrected often many new ones are introduced because of the fixed spatial relationships imposed by the metric used. In this work new exemplars will generate errors only if they share comparatively many features (forks) with the error patterns, which would in turn imply some visual similarity and therefore some justification for the errors.

Exemplars representing sub-clusters of similar images may be extracted from the separation matrix C_{PQ} by identifying those images in each class in the training set that give rise to errors using Equation (2) and generate new exemplars in the same manner as before but applying Equation (1) to the class error set. The new exemplars then are centred on error clusters if they exist. The motivation of this approach is similar to that of support vectors which take account of additional structure contained in the error sets.

1.4 Dataset

1045 images were gathered using Nokia 7610 camera phones at a variety of times by a number of different volunteers without any specific instructions or guidance. The images were labeled according to the 4 locations where they were taken. A training set of 385 images was randomly selected from the total set and the remaining 660 images served as a test set.

No sifting of the data was carried out and many of the images were blurred or taken in very poor light. The image set therefore represented an extremely diverse set of images which are indicative of the material that an amateur photographer might generate and which requires categorisation to become informative. Each image is in the Bitmap format and has a resolution of 216x144. The distribution of images in each class is shown below in Table 1.

2 Experimental results

6 exemplars were extracted using the 385 training data set.. The exemplars represent potential visual sub-clusters across the four classes of locations. Table 1 shows the classification-error-rate using the extracted exemplars. Errors in the training set reduced to 179 using a single exemplar for each of the 4 classes. The addition of two more exemplars for class 1 and class 4 reduced the errors down on both the training and test set, but perhaps more significantly new visual sub-clusters were extracted that were present in the training and test sets (Fig 3a,4b). The images are ranked in accordance with the strength of the similarity with the exemplar image. This confirmed that some of the images that were classified incorrectly by the first selection of exemplars possessed other features in common that independently characterised their class identities. Further exemplars derived from the errors did not improve the results and reflected the unstructured and dissimilar nature of the remaining data.

The similarity matrix C_{AB} for the training set of 385 images took about 1 hour to compute when distributed across five Dell Pentium 4 2.8GHz processors. The exemplars were extracted in a few seconds.

		Class 1	Class 2	Class 3	Class 4	Class 1	Class 4
Training Set		1	2	3	4	5	6
Class 1	173	0	0	13	30	27	30
Class 2	50	50	29	40	41	41	41
Class 3	83	83	83	55	59	60	60
Class 4	79	79	79	79	49	49	38
	385	212	191	187	179	177	169
Test Set		1	2	3	4	5	6
Class 1	294	0	9	40	23	18	39
Class 2	49	49	33	43	42	42	42
Class 3	251	251	251	114	110	109	110
Class 4	66	66	66	66	76	76	43
	660	366	359	263	251	245	234

Table 1: Classification errors

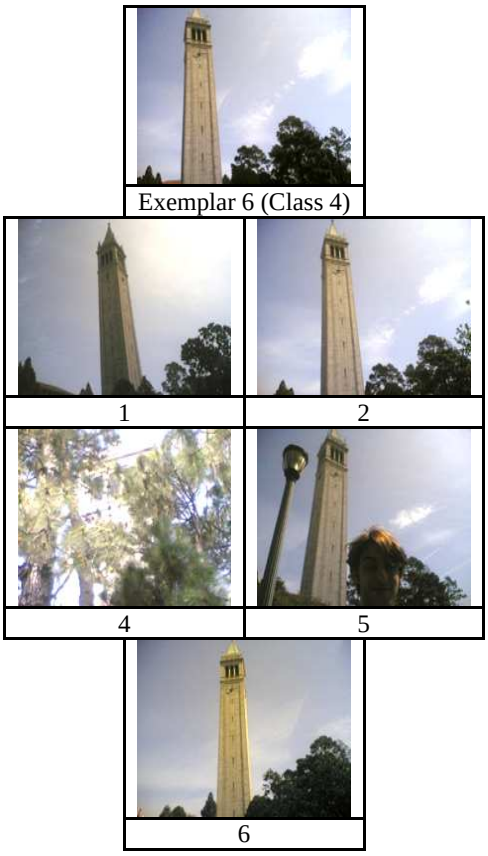

Figure 3a: Training set cluster corresponding to exemplar 6 (Class 4).

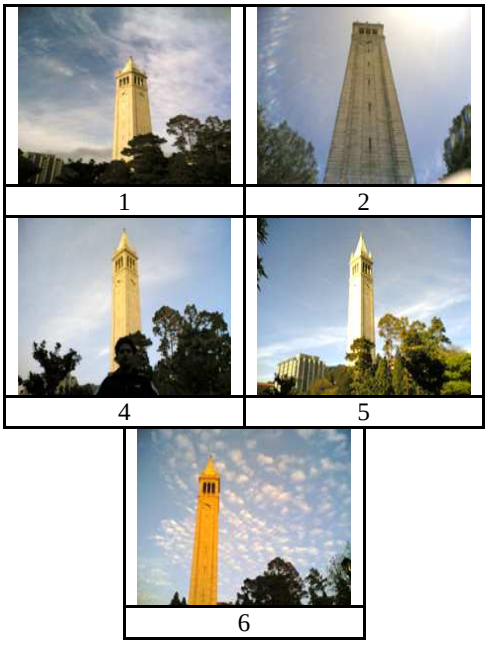

Figure 3b: Test set cluster corresponding to exemplar 6 (Class 4).

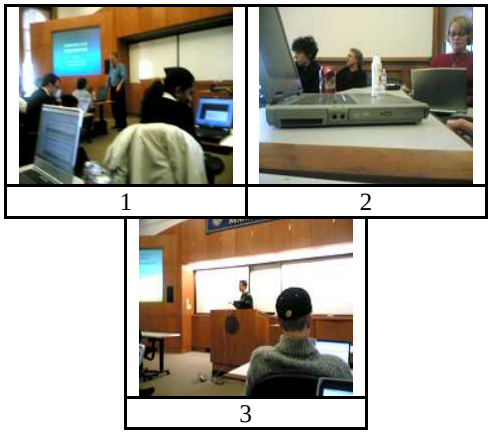

Figure 4a: Training set cluster corresponding to exemplar 3 (Class 3)

Figure 4b: Test set cluster corresponding to exemplar 3 (Class 3)

Figures 4a and 4b show a further example of visually similar clusters detected in the training and test sets. The images are rank ordered according to their similarity with the closest exemplar.

3 Discussion

The error rates are extremely high in this work but this really reflects the variability of the images within each class. Indeed there are many instances of images that contain no obvious clue as to the identity of the location. It is significant that despite this diversity, the approach is able to extract visually similar clusters of images that can be classified according to location. The similarity measure to a certain extent is able to identify features in common that are probably derived from actual structures present in each location and therefore present in the images, although this is not proven.

The computation requirements for the similarity matrix go up as the square of the number of images. It is envisaged that larger classes of data would be divided up into parts before analysis and a hierarchy of exemplars generated.

Retrieval tasks would make use of such a hierarchy to identify clusters likely to contain target images rather than attempt to carry out an exhaustive search.

It is expected that as an image collection expanded new classes would be introduced and new clusters would emerge based on entirely different sets of features in common..

4 Conclusions & future work

This paper has described an approach to the identification of clusters within an extremely diverse set of images. Matrices of inter-image similarities were generated without the use of an *a priori* selection of features, but the measure was nevertheless dependent upon the number of properties that images possessed in common. It has been shown that the selection of exemplars for classification is sufficiently convergent to yield corresponding results in a test set with corresponding clusters being visually similar clusters to those in the training set. Future work will investigate the effectiveness of unsupervised clustering with application to much larger bodies of data where the number of clusters is unknown and the data is not labelled.

Acknowledgements

The authors wish to acknowledge the support of Research and Venturing within British Telecom and colleagues at the University of California at Berkeley [27].

The work also falls within the scope of the MUSCLE Network of Excellence in the European 6th Framework [28].

References

[1] M.Ankerst, G.Kastenmuller, H-P.Kriegel, and T.Seidl, “3D histograms for similarity search and classification in spatial databases.” Proceedings of the International Symposium on Spatial Databases, Hong Kong, (1999).

[2] A. Bamidele and F.W.M Stentiford, “Fusing Contextual Metadata and Visual Similarity in Mobile Media Location-Based Classification”, London Communications Symposium, Multi-media systems and applications, University College London, (Sep. 2005).

[3] A. Bamidele, F. W. M Stentiford and J. Morphett, “An Attention-Based Approach to Content Based Image Retrieval”, British Telecommunications Advanced Research Technology Journal on Intelligent Spaces, Springer Verlag Book edition, (Nov. 2004).

[4] A. Bamidele, “An Attention-Based Model to Colour Histogram Indexing”, European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, London, U.K, (Nov. 2004).

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

[6] Y.Chen, J.Z. Wang and Robert Krovetz, “CLUE: Cluster-based Retrieval of Images by Unsupervised Learning”, IEEE Transactions on Image Processing, (2004).

[7] I.J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, .The Bayesian image retrieval system, PicHunter: theory, implementation, and Psychophysical experiments. IEEE Trans. On Image Processing, Vol. 9, No 1(Jan. 2000).

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

[9] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, “Efficient and Effective Querying by Image Content," J. Intell. Inform. System. vol. 3, no. 3-4, pp. 231-262, (1994).

[10] C.Grigorescu, N.Petkov, and M.A.Westenberg, “Contour detection based on nonclassical receptive field inhibition.” IEEE Trans. on Image Processing, 12(7), (2003) 729-739.

[11] A. Gupta and R. Jain, “Visual Information Retrieval”, Commun. ACM, vol. 40, no. 5, pp. 70-79, (1997).

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

[13] E.Izquierdo, V.Mezaris, E.Triantafyllou, and 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,(September 2002) http://www.iti.gr/SCHEMA/library/index.html

[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, and C.Schmid, “Shape recognition with edge-based features.” Proceedings of the British Machine Vision Conference, Norwich, UK, (2003).

[16] T. V. Papathomas, T. E. Conway, I. J. Cox, J. Ghosn, M. L. Miller,T. P. Minka, and P. N. Yianilos, “Psychophysical studies of the performance of an image database retrieval system,” in Proc. IS&T/SPIE Conf. Human Vision Electronic Imaging III, San Jose, CA, pp. 591–602, (July 1998).

[17] R.W. Picard and T.P. Minka, “Vision Texture for Annotation”, Journal of Multimedia Systems, vol.3, no. 1, pp.3-14, (1995).

[18] S. Sarkar and P. Soundararajan, "Supervised Learning of Large Perceptual Organization: Graph Spectral Partitioning and Learning Automata", IEEE Trans. Pattern Anal. Machine Intelligence, vol. 22, no. 5, pp. 504-525, (2000).

[19] C. Schmid and R. Mohr, “Local Gray value Invariants for Image Retrieval”, IEEE Trans. Pattern Analysis. Machine Intelligence, vol. 19, no. 5, pp. 530-535, (1997).

[20] G. Sheikholeslami, W. Chang, and A. Zhang, "SemQuery: Semantic Clustering and Querying on Heterogeneous Features for Visual Data", IEEE Trans. Knowledge and Data Engineering, vol. 14, no. 5, pp. 988-1002, (2002).

[21] W. M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, “Content-Based Retrieval at the End of the Early Years”, IEEE Trans PAMI, Vol 22, No 12, pp 1349-1379 (December 2000).

[22] J.R.Smith, and S-F.Chang, “VisualSEEk: a fully automated content-based image query, ACM International Conference on Multimedia, Boston MA, USA, pp 87-98, (1996).

[23] 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).

[24] M.J. Swain and B.H. Ballard, “Color Indexing”, Int'l J. Computer Vision., vol. 7, no. 1, pp. 11-32, (1991).

[25] A. Vailaya, A. K. Jain, and H. J. Zhang, “On image classification: City images vs. landscapes,” Pattern Recognition., vol. 31, no. 12, pp.1921–1936 (1998).

[26] J.Z.Wang, J.Li, and G.Wiederhold, “SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries”, IEEE Transactions on pattern analysis and machine intelligence, vol. 23, no. 9 (September 2001).

[27] http://garage.sims.berkeley.edu/

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