# Fusing Contextual Metadata and Visual Similarity in Mobile Media Location-Based Classification

A Bamidele and F W M Stentiford

CONTENT UNDERSTANDING GROUP, UNIVERSITY COLLEGE LONDON, ADASTRAL PARK CAMPUS, IPSWICH, UK.

**Abstract:** This paper describes a new approach to the automatic identification of the location of mobile images. New methods for determining image similarity are combined with analysis of automatically acquired contextual metadata to produce location information. Results are reported on a database of 1209 real images collected on Nokia 7610 camera phones by different users in 12 different locations across multiple mobile carrier cell identities.

## 1. Introduction

The numbers of images being taken by digital cameras has increased dramatically in the last few years, but this is likely to be overtaken by camera phones which also serve as a communications device and are even more pervasive in society. Camera phones represent the convergence of multiple technologies and are able to generate and assign semantic metadata to images from several sources. Technical details such as the capture time and date, and the exposure and aperture settings are normally encoded into the image files, and more recently location and Bluetooth contact identities which give co-presence information are also available. This means that a great deal of automatically generated contextual information about each image is available to deduce the content and to retrieve semantic information. Timing information has been shown to be effective in clustering personal image collections [1]. Naaman [2] used locational data as well as time of day and weather information to provide contextual cues to users for browsing their images. Time and content based features (DCT coefficients) are combined by Cooper [3] to produce clustering of images taken at events. In this paper we address the problem of classifying images taken by a variety of people according to their location. Although the mobile phone cell coverage identity is available it does not discriminate between locations within the same cell. More precise data is available from the Global Positioning System (GPS), but again this does not function reliably in or near buildings. We combine and compare metadata derived from the cell identity, the time of day, the week day and image similarity to identify the location on the Berkeley Campus and surrounds.

## 2. Background in Image Analysis

Similarity measures are central to most pattern recognition problems not least in computer vision and the problem of categorising and retrieving huge number of digital images. These problems have motivated considerable research into content based image retrieval [4] and many commercial and laboratory systems are described in the literature [5]. There are many approaches to similarity and pattern matching and much of this is covered in several survey papers [6]. Many approaches involve the use of pre-determined features such as edges, colour, location, texture and functions dependent on pixel values e.g. [7]. Mikolajczyk et al [8] 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 [9].

### 2.1. Cognitive Visual Attention

Studies in neurobiology [10] 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. Such a mechanism has been explored [11] 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 [12]. 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.

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

$$\boldsymbol{x} = (x_1, x_2) \text{ and } \boldsymbol{a} = (a_1, a_2, a_3) \qquad (1)$$

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

$$S_A = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3 \ldots \boldsymbol{x}_m\}. \qquad (2)$$

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

$$S_B = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \ldots, \boldsymbol{y}_m\}$$

where $\boldsymbol{x}_i - \boldsymbol{y}_i = \boldsymbol{\delta}_j$. The fork $S_A$ matches the fork $S_B$ if

$$|\boldsymbol{F}(\boldsymbol{x}_i) - \boldsymbol{F}(\boldsymbol{y}_i)| < \boldsymbol{\varepsilon} \quad \forall i \text{ for some } \boldsymbol{\delta}_j \quad j = 1,2,\ldots,N \qquad (3)$$

In general $\varepsilon$ is not a constant and will be dependent upon the measurements under comparison ie.

$$\varepsilon_j = f_j(\boldsymbol{F}(\boldsymbol{x}), \boldsymbol{F}(\boldsymbol{y})) \quad (4)$$

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 $S_A$ :

$$C_{AB} = \sum_{i=1}^{T} w_i \quad where\ w_i = \begin{cases} 1 & if\ S_A\ matches\ S_B \\ 0 & otherwise \end{cases} \qquad (5)$$

$C_{AB}$ in (5) 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.

## 2.2. 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 location classification decisions. The selection of exemplars that characterise the pattern class may be carried out in many different ways and are considered later. The most straightforward selection is the visual centre of gravity 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).

$$G_I = \max_{P \in I}{}^{-1} \sum_{Q \in I, P \neq Q} C_{QP} \qquad (6)$$

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

$$Cl(U) = \max_R C_{UG_R} \qquad (7)$$

$Cl(U)$ *(7)* 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.

## 2.3. Visual Sub-Cluster Extraction

The method of selection of a single exemplar from a class of training images given in 2.2 will yield an exemplar that represents the most self-similar group of images within that class. However, many different images may be captured at each location and the location class is represented more realistically by several sub-clusters that contain similar content but are different from each other. 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 that are dissimilar to exemplars already selected but similar to others in the class. We generate a difference similarity matrix

$$C'_{PQ} = C_{PQ} - C_{PG_1}$$

where $G_1$ is the first exemplar image. Positive values in this matrix indicate similarities between images that have few features in common with $G_1$. Images which have many such associations are candidates for a sub-cluster exemplar $G_2$. Let

$$G_2 = \max_{P \in I}{}^{-1} \sum_{Q} C'_{QP}$$

$G_2$ corresponds to the image having the greatest column total and therefore the largest number of features in common with others whilst having little similarity with $G_1$.In a similar fashion a succession of sub-cluster exemplars may be produced:

$$C''_{PQ} = C_{PQ} - C_{PG_1} - C_{PG_2}$$

$$G_3 = \max_{P \in I}{}^{-1} \sum_Q C''_{QP}$$

### 2.4. Colour Histogram Techniques

One of the most popular techniques used in content based image retrieval employs colour histograms mainly because of its simplicity and computational speed [4]. Pixel colour distributions are generated that form feature vectors corresponding to each image. In its simplest form the distances between the feature vectors give an indication of the similarity of the respective images. This approach is quite effective on some image databases, but unless scene geometry is also incorporated it is easy to see that large classes of different patterns will not be separated by this approach. As before classification of image *U* is given by

$$Cl(U) = \min_R D_{UG_R}$$

where

$$D_{QP} = \sum_{i=1}^{B} \left| h_i^Q - h_i^P \right|$$

$$h_i^R = {}^{i^{th}\ pixel\ colour\ bin\ count} \big/ {}_{no.\ pixels\ in\ R}$$

and *B* is the number of colour bins. *B* = 64 in the results below. Colour histograms are used as an alternative similarity measure with which to compare the performance of the CVA algorithm.

| | Colour Histogram | CV A | Random | Metadata | Metadata & Histogram | Metadata & CVA |
|---|---|---|---|---|---|---|
| Number of errors / out of 630 | 440 | 434 | 386 | 283 | 248 | 206 |
| % Reduction from histogram errors | - | 1 | 12 | 35 | 43 | 53 |

Table 1. Test set classification errors

## 3. RESULTS

### 3.1. Location by Contextual Metadata

1209 images were taken using Nokia 7610 camera phones in 12 different locations and 30 different cell coverage identities in and around the Berkeley Campus at a variety of times, by a number of different people without any specific instructions. No sifting of the data was carried out and many of the images were blurred or taken in very poor light. All locations were covered by more than one cell. The metadata associated with each data item was extended to include a manually generated location label.

A training set of 579 images were randomly selected from the total set of 1209 and 33 exemplars selected representing visual sub-clusters across the locations. The remaining 630 data items were used as a test set in the results described below. In addition the distributions of metadata attributes $a^i_{jk}$ (table 2) for each location *i* were extracted from the 579 items where

| | |
|---|---|
| **Hours of the day** | $a^i_{1k} \quad k = 1,...,24$ |
| **Weekday** | $a^i_{2k} \quad k = 1,...,7$ |
| **Cell identity** | $a^i_{3k} \quad k = 1,...,30$ |

Table 2. Contextual Metadata used in the classification process

Normalized distributions were extracted across the 12 locations for each attribute value. Images were then classified by summing the attribute distributions across locations corresponding to the metadata values of the image *U* and selecting the location with the highest value:

$$Cl(U) = \max_R{}^{-1} \sum_j a^R_{jK}$$

where K corresponds to the respective attribute values of U.

### 3.2. Location by Metadata and Vision

Contextual metadata or visual features alone are incapable of determining location, but in combination a better result should be attainable than either approach in isolation. The metadata attribute distributions for candidate images U were augmented with a normalised and weighted vector $v_i$ *i = 1,…, 12*

where

$$v_i = \alpha . \max_{R=i} C_{UG_R} \Big/ \sum_i \max_{R=i} C_{UG_R}$$

## 4. DISCUSSION

Images were classified using histogram classifier, the CVA classifier and metadata classifiers alone. Not surprisingly the vision systems performed badly because the image set was extremely diverse and contained many images which from their appearance could have been taken in any of the 12 locations. In fact a random classifier based on the frequencies of images taken at each location would have performed better. The metadata performed surprisingly well not just because the cell identities helped to limit the errors, but also because it was apparent that activities were taking place at certain times of the day and days of the week that distinguished between locations. Both vision classifiers reduced errors mostly in locations with overlapping cell coverage. Detailed study of the errors indicated that many images were visually dissimilar to all the exemplars and so the visual attributes that were extracted did not contribute towards the classification decision. It would be expected that as the image collection accumulated, more visual sub-clusters would emerge and performance would improve. It should be emphasised that the material in these experiments is not from personal collections where time and date of image capture provides a natural attribute for accurate clustering.

## 5. CONCLUSIONS & FUTURE WORK

This paper has described a new approach to the automatic identification of the location of mobile images. It has show that the combination of attributes derived from both contextual metadata and image processing produces a measure that can indicate the location at which images were taken. It is planned to make use of data obtained from Bluetooth contacts made at the time of image capture as this could provide further location information if certain groups are more likely to congregate in certain locations than others. In addition this information will also enable the location of images taken by those contacts at approximately the same time to be assigned even if the other metadata is nonexistent.

The authors wish to acknowledge the support of Research and Venturing within British Telecom and the members of the Mobile Media Metadata project at the University of California at Berkeley [14]. The work also falls within the scope of the MUSCLE Network of Excellence in the European 6th Framework [13].

## References.

[1] A. Graham, H.Garcia-Molina, A.Paepcke and T.Winograd, “Time as essence for image browsing through personal digital libraries,” JCDL ’02, July 13-17, Portland, 2002.

[2] M.Naaman, S.Harada, O.Wang, H.Garcia-Molina, and A.Paepcke, “Context data in geo-referenced digital image collections,” ACM Multimedia ’04, October 10-16, New York, 2004.

[3] M.Cooper, J. Foote, A.Girgensohn, and L.Wilcox, “Temporal event clustering for digital image collections,” ACM Multimedia ’03, November 2-8, Berkeley, 2003.

[4] 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, Sept. 2002, http://www.iti.gr/SCHEMA/library/index.html

[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) (2002) 1026-1038.

[6] J.R.Smith, and 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.

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

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

[9] 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.

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

[11] 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.

[12] 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.

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

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