Using Context and Similarity for Face and Location Identification

Marc Davis[1], Michael Smith[2], Fred Stentiford[5], Adetokunbo Bamidele[5],
John Canny[3], Nathan Good[1], Simon King[1], Rajkumar Janakiraman[4]

School of Information Management and Systems, U.C. Berkeley, {marc, ngood, simonpk}@sims.berkeley.edu[1]
France Telecom R&D, South San Francisco, CA, michael.smith@rd.francetelecom.com[2]
Computer Science Division, U.C. Berkeley, Berkeley, CA, jfc@cs.berkeley.edu[3]
University College London, Adastral Park Campus, Martlesham Heath, Ipswich, UK {f.stentiford, a.bamidele}@ee.ucl.ac.uk[5]
School of Computing, National University Singapore, Singapore, janakira@comp.nus.edu.sg[4]

ABSTRACT

This paper describes a new approach to the automatic detection of human faces and places depicted in photographs taken on cameraphones. Cameraphones offer a unique opportunity to pursue new approaches to media analysis and management: namely to combine the analysis of automatically gathered contextual metadata with media content analysis to fundamentally improve image content recognition and retrieval. Current approaches to content-based image analysis are not sufficient to enable retrieval of cameraphone photos by high-level semantic concepts, such as who is in the photo or what the photo is actually depicting. In this paper, new methods for determining image similarity are combined with analysis of automatically acquired contextual metadata to substantially improve the performance of face and place recognition algorithms.

For faces, we apply Sparse-Factor Analysis (SFA) to both the automatically captured contextual metadata and the results of PCA (Principal Components Analysis) of the photo content to achieve a 60% face recognition accuracy of people depicted in our database of photos, which is 40% better than media analysis alone. For location, grouping visually similar photos using a model of Cognitive Visual Attention (CVA) in conjunction with contextual metadata analysis yields a significant improvement over color histogram and CVA methods alone. We achieve an improvement in location retrieval precision from 30% precision for color histogram and CVA image analysis, to 55% precision using contextual metadata alone, to 67% precision achieved by combining contextual metadata with CVA image analysis. The combination of context and content analysis produces results that can indicate the faces and places depicted in cameraphone photos significantly better than image analysis or context analysis alone. We believe these results indicate the possibilities of a new context-aware paradigm for image analysis.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation (*e.g.*, HCI)**]: Multimedia Information Systems; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis.

General Terms

Algorithms, Design, Experimentation.

Keywords

Clustering, Similarity, Content-base Image Retrieval (CBIR), metadata, mobility, GPS, Bluetooth Context-Aware, Face Recognition, Cameraphone, SFA, PCA

1. INTRODUCTION

For decades researchers in computer vision and multimedia information systems have attempted to solve basic problems in image content analysis such as face recognition, place recognition, and object recognition [17]. Other than in highly constrained applications (e.g., mugshot databases in law enforcement), purely signal-based approaches to addressing these challenges have not yet achieved a level of performance such that they are used in web image search today. The opportunity to address these unsolved image content recognition challenges in a new way is presenting itself in the recent confluence of mobile media capture, context-sensing, programmable computation, and networking in the form of the nearly ubiquitous cameraphone.

Cameraphones are rapidly becoming the dominant platform for consumer digital imaging worldwide. Technology analysts from Future Image Inc. predict that 500 million cameraphones will be sold worldwide in 2005, which also means that 5 out of every 6 digital imaging devices sold in 2005 will be in cameraphones. The growing ubiquity of cameraphones and the attendant explosion in the number of photos taken worldwide present both a great challenge and opportunity for multimedia researchers.

Cameraphones as a platform for multimedia computing offer us the chance to pursue new approaches to media analysis: namely to combine the analysis of automatically gathered contextual metadata with media content analysis to significantly improve image content recognition and retrieval. To better illustrate the limitations of the current dominant paradigm in computer vision which does not make use of contextual metadata in image analysis, consider this lighthearted parable of human perception:

"You go out drinking with your friends. You get drunk…really drunk. You get hit over the head and pass out. You are flown to a city in a country you've never been to with a language you don't understand and an alphabet you can't read. You wake up face down in a gutter with a terrible hangover…you have no idea where you are or how you got there."

This is what it's like to be most computer vision systems—they have no *context* (and usually little or no memory and are also sensory-impaired). *Context* is what enables us to understand what we see. Using contextual metadata automatically gathered from

Internet Imaging VII, edited by Simone Santini, Raimondo Schettini, Theo Gevers,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6061, 60610E, © 2006 SPIE-IS&T · 0277-786X/06/$15

cameraphones, we can more closely model the process of human image understanding by adding context to multimedia processing and retrieval [6, 9].

In our own research [6,7,8,21], we capture a variety of contextual metadata using the sensors readily available on consumer cameraphones: *temporal* (exact time served from the cellular network); *spatial* (CellID from the cellular network and GPS location from Bluetooth-connected GPS receivers); and *social* (who took the photo, who sent and/or received the photo when shared, and who was co-present when the photo was taken sensed via Bluetooth MAC addresses mapped to usernames). We have found in our experiments that analysis of such automatically gathered contextual metadata for face recognition[8] (the person(s) depicted in the photograph) and place recognition (the location depicted in the photograph) without any image analysis at all outperforms image content analysis alone. The combination of context and content analysis outperforms either one.

For faces, we demonstrated a significant improvement in retrieval precision from 43% precision for Principal Components Analysis (PCA) image analysis methods for determining the identity of a person depicted in a photo, to 50% precision using Sparse Factor Analysis (SFA) of contextual metadata alone, to 60% precision achieved by combining contextual metadata analysis with image analysis [8]. For location, we achieve a significant improvement in retrieval precision from 30% precision for color histogram and Cognitive Visual Attention (CVA) image analysis methods for determining the location depicted in a photo, to 55% precision using contextual metadata alone, to 67% precision achieved by combining contextual metadata with CVA image analysis. The improvements our results demonstrate over purely signal-based analysis techniques are substantial. The combination of context and content analysis produces results that can indicate the faces and places depicted in cameraphone photos significantly better than image analysis or context analysis alone. We believe these results indicate the possibilities of a new context-aware paradigm for image analysis.

It is important to note that in our analysis we are solving for the location of the *subject of the photo*, not of the cameraphone. *Automatically* gathered contextual metadata can help us determine the location of the cameraphone, but it does not tell us what the photographer is pointing the cameraphone at.
While automatically gathered contextual metadata outperforms image analysis alone, by helping reduce the set of images that content-based analysis need consider, the two approaches together (context and content analysis) can achieve better results than either alone.

In short, context can help content analysis to focus on what are the best subsets of photos to analyze, and content analysis can help context analysis disambiguate various locations that are below the level of its metadata precision.

1.1 Related Work

The research of Naaman, et al., uses similar context features as our work for identifying human subjects [14]. The key differentiator is that we combine contextual analysis with signal-based face recognition to produce a better result than contextual analysis or computer vision alone can provide [8]. Other research cited in [8] has explored methods for face image annotation that focus on image similarity, thumbnail visualization, and intuitive interfaces. Much of this work is focused on annotation interfaces. In prior work, a list of candidates is presented for verification using a compact interface. New methods in face recognition, such as high-resolution images, three-dimensional face recognition, and new preprocessing techniques may offer improved accuracy [11,15,20], but our context-aware approach utilizes comparatively lightweight computation and offers significantly improved performance today.

Timing information has been shown to be effective in clustering personal photo collections [11]. Naaman [13] used locational metadata as well as time of day and weather information to provide contextual cues to users for browsing their photos. Time and content based features (DCT coefficients) are combined by Cooper [5] to produce clusters of photos taken at events. In this paper, we focus on the use of automatically sensed and inferred metadata about the context of photo capture together with computer vision analysis of image content to make accurate predictions about the locations depicted in cameraphone photos. We combine locational metadata derived from the cellID, temporal metadata about the time of day and the day of the week, and image similarity to identify the location at which photographs were taken.

2. SYSTEM OVERVIEW

2.1 MMM2: Gathering Data and Metadata

The MMM2 system uses a client-server software architecture. A single Java-based HTTP server running on a Linux machine aggregates data and coordinates photo distribution for multiple client applications running on Nokia Series 60 handsets. The server application stores photo metadata and user profile information in a relational database accessed by Java servlets which interface via HTTP with the Context Logger and the web browser running the client handsets.

2.1.1 MMM2 Context Logger

The Context Logger is a Symbian application developed by and modified in cooperation with the University of Helsinki Department of Computer Science Context Project (http://www.cs.helsinki.fi/group/context/). The Context Logger runs continuously on the handset and obtains rough location information by logging switches between cell towers. In addition, the logger periodically accesses the handset's Bluetooth radio to poll for the presence of nearby Bluetooth devices and to obtain precise location information from Bluetooth-enabled GPS devices (if available). Finally, the Context Logger monitors the phone's file system to detect new photographs. When a new photo is detected the Context Logger displays a simple user interface and begins to uploading the new photo to the MMM2 server in a background process. If the user selects to share the photo immediately, the Context Logger launches the phone's web browser to display an HTML-based user interface generated by the server. In the case of immediate sharing, photo metadata is transferred to the server via a URL query string. If the user opts not to share immediately, the metadata is attached to the end of the photo upload stream in XML format. The metadata snapshot associated with each photo consists of photo capture time, nearby Bluetooth device IDs, cell tower ID and GPS location (if available).

2.2 Creation of Ground-Truth Dataset

The data used for the location dataset came from the previously described MMM2 system. The details of the data are listed below.

Table 1. Metadata Types, Sensors, Features, Feature Details for MMM2 System

Type	Sensor	Feature	Feature Detail
Temporal	Cellular Network	Time of Capture	Weekend/weekday
			Day of week
			Minute of day
			Timeslot (24 overlapping 2-hour intervals)
			Detail
Social	MMM2 Server	Photo Owner	Phone owner (likely to be photographer)
		Other Users	Which potential recipients are system users at share time
Spatial	Cellular Network	Cell ID	Cell ID where photo was captured

1. Time of Capture—binary values indicating whether or not the image was captured on the weekend. Also, the capture timeslot (hour of day, 24 values), day of week, and minute of the day
2. Photo Owner—the identity of the photo owner, 66 values
3. Other Users—recipient is a User, 66 values
4. Cell ID—the cell ID (426 values)

Each entry in the table describes the type of information, what sensor was used to collect it, and what the exact feature was. The type column describes where the information sits in relation to our view of contextual information. The sensor column describes how the information is actually collected from a technical perspective. The cellular network provides the phone time information and location in the form of the cell tower and cell id that the phone connects to. We gathered social information (contacts, shares and share frequencies) from the MMM2 server that we created.

To create a set of photos annotated with labeled faces we used a custom-built Java applet that can be accessed on the web, linked from the MMM2 website. The applet allows a user to select a region of a photo and associate a person's name with this region. Selecting a region associated with each face rather than simply a single point with the face allows this metadata to be used for face detection as well as recognition. Users were instructed to select regions of the photo containing faces (from ear to ear, forehead to chin), which were at least 20-30 pixels wide and in which the face is visible enough for the human annotator to recognize it. In an effort to create a dense set of annotated photos in which many faces appear many times, rather than a sparse set in which many faces appear only a few times, we had several MMM2 users (primarily from the development team) use this annotation tool to annotate as many photos as possible.

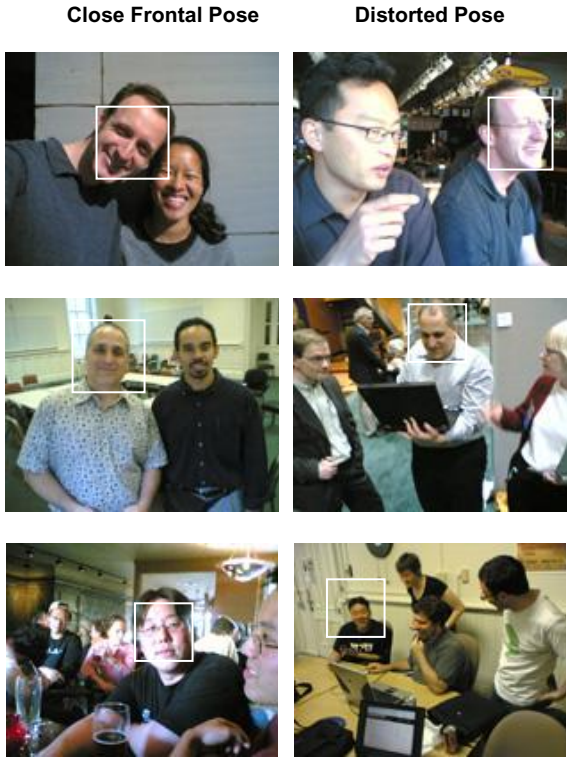

Figure 1. (Left) Subjects with frontal pose, (Right) Same subjects with non-frontal or distorted pose.

Eleven users total used the annotation tool, seven of which each annotated at least 20 photos. The result is a dataset of 1057 photos with faces, covering 173 different faces with 31 faces occurring at least 10 times each and 58 faces appearing at least 5 times each. While only 1057 photos had faces, the annotation process also produced a set of nearly 2000 additional photos known not to contain faces. While these additional photos are of no use in machine vision face recognition, the data can still be used in attempting to determine the contexts in which a user is likely to be photographing people rather than non-person subjects. Examples of the photos taken are shown in Figure 1. Frontal pose images represent a small fraction of our face images.

3. CONTENT ANALYSIS

3.1 Image Analysis

Similarity measures are central to most pattern recognition problems especially 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 [17] and many commercial and laboratory systems are described in the literature [e.g. 4]. There are many approaches to similarity and pattern matching and much of this is covered in survey papers [20].

3.2 Cognitive Visual Attention

Studies in neurobiology and computer vision [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 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 [1, 2, 18]. 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

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

Define a function $\boldsymbol{F}$ such that $\boldsymbol{a} = \boldsymbol{F(x)}$.

Select a fork of m random points S_A in pattern A where

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

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

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

$$\boldsymbol{x}_i - \boldsymbol{y}_i = \boldsymbol{\delta}_j$$

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 ie.

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

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 :

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

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.

3.3 Training and Classification

The similarity values obtained in this way may be used to drive a nearest neighbor classifier in which relative similarities with class representatives or exemplars determine the location classification decisions. The training process consists of the selection of representative images or exemplars that characterise the pattern class.

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).

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

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

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

Here $Cl(U)$ identifies the class exemplar that shares the most features with the unknown pattern. It is important to emphasize 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.

Table 1. Test set classification errors

	Color Histogram	CVA	Random	Metadata	Metadata & Histogram	Metadata & CVA
Number of errors / 630 photos	440	434	386	283	248	207
% Error	70	69	61	45	39	33
% Reduction in histogram errors	_	2	12	36	44	53

3.4 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 [1]. However, many different photos 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 (Fig. 2) may be produced:

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

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

3.5 Color Histogram Techniques

One of the most popular techniques used in content based image retrieval employs color histograms mainly because of its simplicity and computational speed [17]. Pixel color 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 color bins. $B = 64$ in the results (Table 1).

Color histograms are used as an alternative similarity measure with which to compare the performance of the CVA algorithm.

Figure 2. Visual sub cluster example corresponding to an exemplar (first image).

3.6 Face Recognition

Face recognition has long been the standard for identifying humans in images. Current methods attempt to detect key facial features such as eyes, nose, and lips, and match these features to known templates for human faces. Evaluation of these methods usually occurs with frontal facing images, such as those shown in the left column of Figure 1. Problems occur when facial imagery is not frontal as in the right column in Figure 1. Most of the images in this research were taken in natural settings with limited frontal pose. We tested 4 publicly available face recognition systems implemented by Colorado State University (CSU - http://www.cs.colostate.edu/evalfacerec/):

- PCA: Eigenfaces principle components analysis based on linear transformations in feature space. PCA requires a short training time and uses a relatively small dimensionality of feature vectors. Many distance measures can be used, but we received the best accuracy with Euclidean and Mahalanobis.
- LDA+PCA Combination: Linear discriminant analysis based on the University of Maryland algorithm in the FERET tests. LDA training requires multiple images and first using PCA to reduce the dimensionality of the feature vectors.
- Bayesian MAP: Maximum a posteriori (MAP) difference classifier based on the MIT algorithm developed by Moghaddam and Pentland. This algorithm examines the difference image between two photos to determine whether the two photos are of the same subject.
- Bayesian ML: Maximum likelihood (ML) classifier based on the same MIT algorithm above.

3.7 Predicting Faces Using SFA

We used Sparse Factor Analysis (SFA) to combine the various contextual factors and metadata to predict which faces occurred in each photo. SFA is a linear probabilistic model that deals correctly with missing data. SFA was shown in [3] to be the most accurate method on standard collaborative filtering data (the EachMovie dataset). SFA is a generative probabilistic model, whose parameters are computed using expectation maximization (EM). Since it is a full generative model, it has a prior which serves as a regularizer. On instances where there is less evidence, the prior distribution exerts more influence and the algorithm is more conservative in its predictions. This works well on datasets with missing information. Our dataset contained cases were data was missing or not always available, for example the Bluetooth co-presence information.

Because SFA is a linear model, it also gives a direct means to infer the probable influence of particular metadata on the predictions. SFA is described formally as a model:

$$Y = m X + N$$

Where Y is a vector of (partially) observed values, X is a latent vector representing user preference, m is the "model" predicting user behavior, and N is a noise function. Y and X are assumed to be real-valued vectors. N is assumed to be multivariate, independent Gaussian noise. X is assumed to have a Gaussian prior distribution. All observed data are encoded as fields in the Y vector. All the data except for the computer vision output were discrete values. Each k-valued discrete input was encoded as k fields in Y that were binary value predicates. For instance, if there were 22 possible users, the third user would be represented as a 22-tuple in Y as (0, 0, 1, 0, 0, … 0). This departs somewhat from the ideal model for SFA, but the model was still able to produce useful predictions, as we will see in a moment.

We use SFA to combine all the contextual factors, metadata and data we gathered from computer vision. Our algorithm uses the data gathered from computer vision as another input for generating predictions. The computer vision data were real values, which correspond to the similarity metric between a face image in the test dataset, and another image in the training dataset. As well as known values, any field in a Y vector can be presented to the algorithm as an "X" meaning the value is unknown. This is how partial or missing data is received.

The SFA method requires two phases. In the learning phase, the EM recurrence is run on a training set of Y vectors to determine the most likely value of the model parameters including the matrix m. Training data will include all metadata fields, the results of computer vision algorithms, and the actual user identity (assuming this is known). In use, the algorithm will receive all contextual metadata and the results of computer vision analysis of the photo. From these partial observations Y and from the model parameters, a single E-step is used to determine the expected value of X for that instance. This X value is then used to predict all missing Y-values directly from the model equation above. These missing values will predict the identities of faces in the photo. This prediction is the MAP prediction for the missing data given the model.

3.8 GPS Clustering

Our algorithm was unable to directly utilize the GPS coordinate of our geo-referenced photos, so we had to change them to a more suitable format. We decided to create two sets of clusters of GPS coordinates, one using k-means and the other using farthest first clustering. We ran both of these algorithms over the entire dataset with the goal of creating 100 clusters for each algorithm. We chose the algorithms and the number of clusters to provide clusters whose centroids approximated the geographical spread of the geo-referenced photos. After clustering, we calculated the geographical distance between each geo-referenced photo and the various cluster centroids and used that value to connect each photo to its nearby clusters.

4. EXPERIMENTAL DATA

4.1 Face Recognition on Cameraphone Data

Face recognition is highly dependent on frontal pose and not well suited for cameraphone photos. The quotidian use and portability of a cameraphone leads to a capture environment that is often more varied than that of photos of human subjects captured with conventional cameras. Cameraphone users often take spontaneous photos [8, 21] often with non-frontal subjects, as shown in the bottom row of Figure 1. The low resolution and slow shutter speed of current cameraphones, which creates motion blur, or grainy photos in poor lighting conditions, also reduces face recognition accuracy. The much higher accuracy of the same vision algorithms we use in our study in face recognition trials using the NIST FERET dataset (http://www.frvt.org/FERET/default.htm) may be attributed to the "mug shot" quality of the photos in the NIST FERET corpus, i.e., each photo is of people depicted in full frontal view in a head-and-shoulders shot. Our MMM2 corpus of over 27,000 cameraphone photos collected by our 66 users over 10 months shows much greater variability of photo conditions and often has multiple people depicted per photo—as such, our study attempts to test the "real world" accuracy of face recognition algorithms and approaches.

4.2 Photographic Location Data

1209 images were taken using Nokia 7610 cameraphones in 12 different locations and 30 cell 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 photos were blurred, mis-oriented, 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 for evaluation purposes.

5. EXPERIMENTAL DESIGN

To evaluate context-enabled face prediction, we used a dataset gathered from 11 users over 9 months. We built a table of Y vectors, each representing a photograph taken by a user. There were 1057 photos total. The set of photos was randomly partitioned into a training set of 337 photos and a test set of 720 photos. The total number of faces was 1402, with 424 used for training and 978 for testing. For each face, 8 photos were taken at random and 4 photos were selected manually for the training set. Manual selection was done to insure a sufficient number of visible

faces in the training set. We will automate this process in future work. The photos in the training set were hand-labeled with the names of actual individuals in each photo. The resulting set of images will be the "training gallery." There were a total of 173 different individuals pictured in the training gallery. The test images were similarly annotated and partitioned. Each photo contained images of 1 to 4 people. The training gallery contained 2-4 images of each subject on average.

For the face recognizers, the test images were again partitioned into distinct faces images. Each photo record has 173 fields (for a particular recognizer), which correspond to the possible subjects in the image. In field number k, we place the value of the metric distance between a face in the test image and face number k from the training gallery. Since there may be multiple faces in the photo, we used the min of distances between all images in the photo, and training gallery image k. In almost all cases, the actual best match between gallery images and test photo involves that lowest weight edge.

The context data is listed in Table 2.

Table 2. Context Data Features

Feature	Value
1. Weekend or Weekday capture	Binary
2. The capture timeslot (hour of day)	24 binary values
3. The identity of the photo owner	11 binary values
4. Was the photo taken indoors or outdoors	Binary
5. The cameraphone cell ID	426 binary values
6. GPS location value, farthest first clustering	100 binary values
7. GPS location value, k-means clustering	99 binary values
8. Identities of people in the photo	173 values
9. The ID of the photo sharing recipient	ID value
10. Bayes MAP comparison metrics for each candidate face	173 real values
11. Bayes ML comparison metrics for each candidate face	173 real values
12. LDA comparison metrics for each candidate face	173 real values
13. PCA comparison metrics for each candidate face	173 real values

5.1 Running the Experiments

We trained the SFA model in two different ways: first using all contextual metadata and the face recognizer outputs, secondly using the contextual metadata only. To evaluate the results, we used precision-recall plots. We also formed precision-recall plots for each of the computer vision algorithms individually, using the negative of the metric distance as the face predictor. In both cases, the model dimension used was 40. Training time was about 2 minutes. Training for the Bayesian classifiers took about 7 hours. PCA and LDA classifiers trained in less than 10 minutes. Face recognition testing takes less than one minute for all 4 algorithms.

6. EVALUATION

6.1 Location by Contextual Metadata

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} for each location i were extracted from the 579 items where

$$\begin{aligned} \text{Hours of the day} \quad & a^i_{1k} \quad k = 1, ..., 24 \\ \text{Weekday} \quad & a^i_{2k} \quad k = 1, ..., 7 \\ \text{Cell identity} \quad & a^i_{3k} \quad k = 1, ..., 30 \end{aligned}$$

Normalized distributions were extracted across the 12 locations for each attribute value. Photos were then classified by summing the attribute distributions across locations corresponding to the metadata values of the photo 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.

6.2 Location by Metadata and Vision

Contextual metadata or visual features alone are incapable of precisely determining location, but in combination a better result should be attainable than either approach in isolation, Table 1.

The metadata attribute distributions for candidate images U were augmented with a normalized and weighted vector v_i $i = 1,...,12$ where

$$v_i = \alpha . \max_{R=i} C_{UG_R} \Big/ \sum_j \max_{R=j} C_{UG_R}$$ and α is a constant.

7. DISCUSSION AND RESULTS

7.1 Face identification Experimental Results

The margins in precision/recall among the different methods are quite large. Context+Vision does better than any individual predictor. Its initial precision is about 60% and is fairly flat across the recall range, as seen in Figure 3. The precision-recall curve has an unusual shape, but that is caused by the very small number of faces to be retrieved for each photo (one to four). The curve's flatness shows that the precision does not decrease very much between the best match and second, third, fourth best. Steps in the curve appear at 1/2, 1/3, 2/3, 1/4 etc. corresponding respectively to photos with 2, 3, 3, 4 users. The sharpest drop is at ½, which is intuitive given that there are quite a few more images of two people than three or four, and also the sharpest accuracy drop is likely to occur between the best and second-best face images.

Context-only prediction (without the aid of computer vision) has 50% initial precision, and a similar slow fall-off. The best vision method was PCA, which was much better than the other vision predictors at around 43%. The other three are quite similar to each other, with LDA doing a little better than the two Bayes predictors, which were around 30%. The PCA Euclidean measure was the simplest and performed the best. This was surprising

considering this measure performed roughly 15% worse with earlier CSU experiments using the NIST FERET data. It may be that PCA is more robust for use with real world datasets; this hypothesis deserves further study.

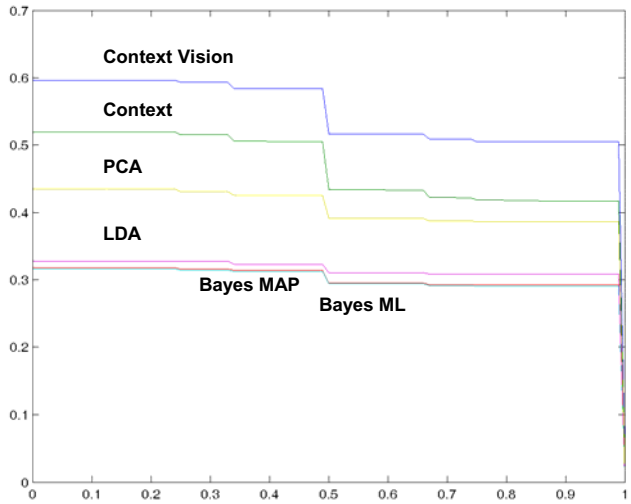

Figure 3. Face Recognition Results from Different Algorithms.

7.2 Location Identification Experimental Results

Images were classified using the 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 photos taken at each location performed better. The metadata performed surprisingly well not just because the cell IDs 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 among various locations for those activities.

Table 3. Error Rate Increase Per Feature Removed

Feature Removed	Rise in error rate
Weekday	0.046
Hour of day	0.116
Cell ID	0.117

The relative value of the contextual categories was tested by removing one at a time and noting the percentage increase in error rate for the metadata plus CVA classifier. Removal of the weekday category gave rise to a 4.6% rise in error rate, and the time and cell identity led to rises of 11.6% and 11.7%, respectively (Table 3). So both time and the cell identity were important and contributed equally towards correct location decisions.

Both combined vision and metadata classifiers reduced errors mostly in locations with overlapping cell coverages. Detailed study of the errors indicated that many images were visually dissimilar to all the exemplars and so in these cases the visual attributes did not contribute towards the classification decision. We expect that as the image collection accumulates, more visual sub-clusters will emerge and performance will improve with the addition of more exemplars.

It should be emphasized that the material in these experiments is not from personal collections, where time and date of image capture alone provides a natural attribute for accurate clustering, but from an image collection of many users in which contextual metadata (spatial, temporal, and social) can be used to determine both individual and group clusters in space-time.

8. CONCLUSIONS & FUTURE WORK

This paper has described a new approach to the automatic identification of human faces and location in mobile images. It has shown that the combination of attributes derived from both contextual metadata and image processing produces a measure that can indicate the location at which photos were taken. In particular, the CVA similarity measure together with the unsupervised cluster extraction promises to be applicable to larger sets of data.

Our future experiments will investigate torso-matching for detecting subjects in multiple photos taken at the same location and time [19] as well as the integration of additional context and content analysis techniques. We will also plan to combine our context-aware face recognition research with our context-aware location recognition research to create a comprehensive solution for mobile media management.

9. ACKNOWLEDGMENTS

The authors wish to acknowledge the support of Research and Venturing within British Telecom, Hewlett-Packard, France Telecom, Nokia, Ricoh Innovations, Opera Software, TEKES, the University of California Discovery Grant for Digital Media, and the students at the UC Berkeley School of Information Management and Systems for contributing their photos to this project. The work also falls within the scope of the MUSCLE Network of Excellence in the European 6th Framework [11].

10. REFERENCES

[1] A. Bamidele, and F.W.M. Stentiford, An attention based similarity measure used to identify image clusters, *Semantic Multimedia analysis*, *2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology,* London, U.K, 30 Nov - 1st Dec., 2005,

[2] 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).

[3] J. Canny, Collaborative Filtering with Privacy via Factor Analysis. *ACM SIGIR in Tampere, Finland*, 2000.

[4] 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

[5] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, Temporal event clustering for digital photo collections, ACM Multimedia '03, November 2-8, Berkeley, 2003.

[6] M. Davis, S. King, N. Good and R. Sarvas, From Context to Content: Leveraging Context to Infer Media Metadata. In:

Proc. of ACM MM 2004 in New York, New York, 2004.

[7] M. Davis, et al. "MMM2: Mobile Media Metadata for Media Sharing." In: *Proceedings of 13th Annual ACM International Conference on Multimedia (MM 2005) in Singapore*, ACM Press, 2005.

[8] M. Davis, M. Smith, J. Canny, N. Good, S. King, and R. Janakiraman. "Towards Context-Aware Face Recognition." In: *Proceedings of 13th Annual ACM International Conference on Multimedia (MM 2005) in Singapore*, ACM Press, 483-486, 2005.

[9] N. Dimitrova, Context and Memory in Multimedia Content Analysis. *IEEE Multimedia, 11(4),* 2004.

[10] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd, Time as essence for photo browsing through personal digital libraries, JCDL '02, July 13-17, Portland, 2002.

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

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

[13] M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke, Context data in geo-referenced digital photo collections, ACM Multimedia '04, October 10-16, New York, 2004.

[14] M. Naaman, R.B. Yeh, H. Garcia-Molina, A. Paepcke, Leveraging Context to Resolve Identity in Photo Albums. *ACM/IEEE-CS Joint Conference on Digital Libraries,* 2005.

[15] P.J. Phillips, Overview of the Face Recognition Grand Challenge. *IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA*, 2005.

[16] R. Sarvas, E. Herrarte, A. Wilhelm, and M. Davis. "Metadata Creation System for Mobile Images." In: *Proceedings of the Second International Conference on Mobile Systems, Applications, and Services (MobiSys2004) in Boston, Massachusetts*. ACM Press, 36-48, 2004.

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

[18] 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.

[19] B. Suh and B. Bederson, Semi-Automatic Image Annotation Using Event and Torso Identification, *Tech Report HCIL-2004-15, University of Maryland, College Park, MD,* 2004.

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

[21] N. Van House, M. Davis, M. Ames, M. Finn, and V. Viswanathan. "The Uses of Personal Networked Digital Imaging: An Empirical Study of Cameraphone Photos and Sharing." In: *Extended Abstracts of the Conference on Human Factors in Computing Systems (CHI 2005) in Portland, Oregon,* ACM Press, 1853-1856, 2005.