Towards Context-Aware Face Recognition

Marc Davis[1], Michael Smith[2], 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]
School of Computing, National University Singapore, Singapore, janakira@comp.nus.edu.sg[4]

ABSTRACT

In this paper, we focus on the use of context-aware, collaborative filtering, machine-learning techniques that leverage automatically sensed and inferred contextual metadata together with computer vision analysis of image content to make accurate predictions about the human subjects depicted in cameraphone photos. We apply Sparse-Factor Analysis (SFA) to both the contextual metadata gathered in the MMM2 system and the results of PCA (Principal Components Analysis) of the photo content to achieve a 60% face recognition accuracy of people depicted in our cameraphone photos, which is 40% better than media analysis alone. In short, we use context-aware media analysis to solve the face recognition problem for cameraphone photos.

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

Context-Aware, Face Recognition, Cameraphone, SFA, PCA

1. INTRODUCTION

Cameraphones are becoming the dominant platform for digital imaging worldwide and have continued to surpass sales of digital cameras since the first half of 2003. InfoTrends/CAP Ventures predicts that 860 million cameraphones will be shipped in 2009, accounting for 89% of all mobile phone handsets. Most important for multimedia researchers, in addition to their growing global ubiquity, cameraphones offer a unique opportunity to purse new approaches to media analysis and management: namely to combine the analysis of automatically gathered contextual metadata with media content analysis to radically improve image content recognition and retrieval. The fundamental challenges of media content analysis have been understood for several years [7]—using contextual metadata gathered from groups of users capturing and sharing media on cameraphones we can "reclaim the world' by adding *context* and *memory* to multimedia processing and retrieval [3, 4]. In short, we are bridging the semantic and sensory gap by reconnecting multimedia analysis to the context in which the media was captured and to the patterns of media capture and use of individual users and groups of users.

MM'05, November 6–11, 2005, Singapore.
ACM 1-59593-044-2/05/0011.

We capture a plethora of contextual metadata using the sensors available on cameraphones: temporal (exact time served from the cellular network); spatial (Cell ID from the cellular network and 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). Using contextual metadata together with media analysis, we have employed collaborative filtering machine learning techniques to predict the probability that a given user has photographed a given subject (e.g., person or place) in a given spatio-temporal-social context. We achieve 60% face recognition accuracy of people depicted in our cameraphone photos. This result represents a nearly 40% improvement over PCA alone (the best performing publicly available face recognition algorithm) which has a 43% face recognition accuracy on the same dataset.

1.1 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 [9] 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 25,000 cameraphone photos collected by our 66 users over 9 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.

2. RELATED WORK

The research of Naaman, et al., uses similar context features as our work for identifying human subjects [5]. With the exception of features for weekend vs. weekday photo capture, and indoor vs. outdoor capture, they also use event, location, time, and neighboring information. 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. Other research cited in [5] 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 [6], but our context-aware approach utilizes comparatively lightweight computation and offers significantly improved performance today.

3. SYSTEM OVERVIEW

3.1 MMM2: Gathering Data and Metadata

The Mobile Media Metadata 2 (MMM2) system consists of two primary components: the Context Logger, running on the cameraphone, and the server application running on a Linux server [2]. The Context Logger is responsible for capturing contextual metadata and uploading photos and metadata to the server. The server application manages photos, their associated metadata, and user profile information. The server application uses a set of servlets and Java Server Pages to generate customized HTML for display on a PC-based web browser or the Opera web browser on a cameraphone handset.

3.1.1 MMM2 Context Logger

The Context Logger (developed by and modified in cooperation with the University of Helsinki Department of Computer Science Context Project (http://www.cs.helsinki.fi/group/context/) is a Nokia Series 60 application that runs as a background process and continually captures contextual metadata and monitors the phone's file system for newly created media (photos, videos, and audio clips). The Context Logger records most phone actions—when a voice call is received or initiated, when the phone switches to a new cell tower, when the phone is charged. Additionally the logger also uses the phone's Bluetooth device to periodically poll for the presence of other nearby Bluetooth-enabled devices. The Context Logger can also communicate with a Bluetooth-enabled GPS devices (we use the HP iPAQ Bluetooth GPS Navigation System) to record GPS location information. When the Context Logger detects a new media file on the file system it displays a one-screen user interface and begins to upload the media file over HTTP to the server. Appended to the end of the image data is an XML fragment containing the context snapshot. This snapshot includes the time, current cell ID, sensed Bluetooth devices, and GPS location information (if available).

3.2 Creation of Ground-Truth Dataset

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.

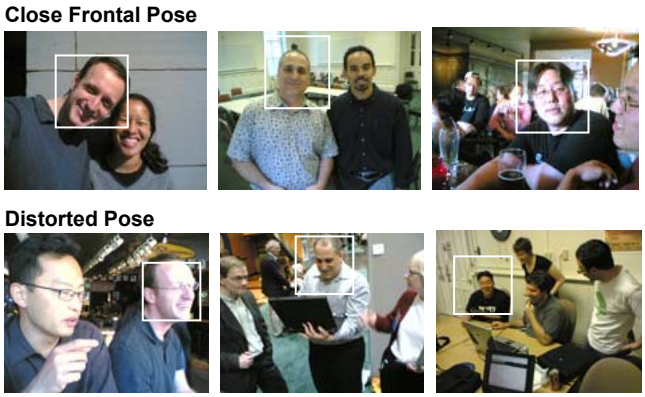

Figure 1. (Top) Subjects with frontal pose, (Bottom) Same subjects with non-frontal or distorted pose.

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.

3.3 Content Analysis: 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 top row of Figure 1. Problems occur when facial imagery is not frontal. Most of the images in this research were taken in natural settings with limited frontal pose, as shown in the bottom row of Figure 1. 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 Mahalinobis.
- 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 in the FERET tests.
- Bayesian ML: Maximum likelihood (ML) classifier based on the same MIT Algorithm above.

3.4 Predicting Faces Using SFA

To combine a diverse set of data and metadata for face identity prediction, we used a general-purpose inference algorithm called

Sparse-Factor Analysis (SFA). SFA is a linear probabilistic model that deals correctly with missing data. SFA was shown in [1] 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. In our case, several types of metadata (such as co-present Bluetooth devices) were only available in some instances. Since it 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.

SFA is used as a "master method" to combine data from all cues, including computer vision. That is, it is used for pure context predictions, and also for vision+context where the vision data is another input. 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.

3.5 GPS Clustering

Since our algorithm could not directly utilize the GPS coordinates of our georeferenced photos, we had to convert the coordinates into a suitable format. We created two sets of clusters of GPS coordinates using two different algorithms: k-means and farthest first clustering. Both algorithms were run over the entire dataset with the aim of finding 100 clusters each. Both algorithms and the number of clusters were empirically chosen to provide clusters whose centroids approximated the geographical spread of the georeferenced photos. After clustering, we calculated the geographical distance between each georeferenced photo and the various cluster centroids and used that value to connect each photo to its nearby clusters.

4. 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 1.

Table 1. 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, Bayes ML, LDA and PCA comparison metrics for each candidate face	173 real values per algorithm

4.1 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.

5. RESULTS AND EVALUATION

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 2. 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.

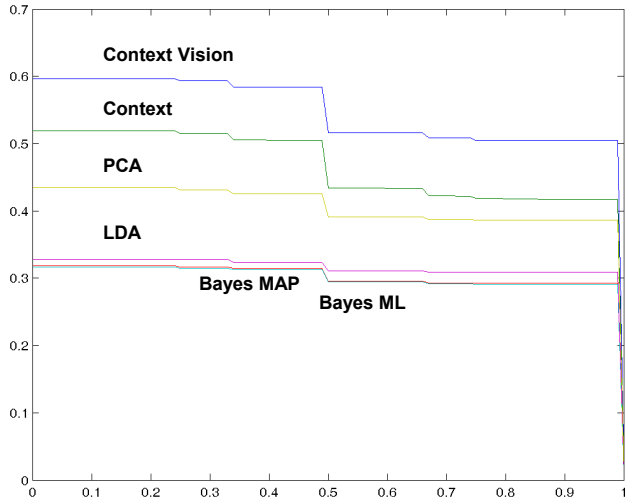

Figure 2. Face Recognition Results from Different Algorithms.

6. IMPLICATIONS

The prognosis for context-assisted face prediction is very good. Although this was a small-scale test (only 11 users), the set of potential faces (173) was realistic for that number of users. The advantage of Context+Vision was quite definitive in this study. The good news is that context-based inference should scale well with the number of users, because it is based on sparse data: each individual user will only have taken photos of a limited number of people in the past and that number is not affected by the total number of users. Nor is the time of day or any of the other metadata items. On the other hand, computer vision alone will have a difficult time as the number of photos increases. Without contextual metadata, each test photo must be matched against every photo in the training gallery, and as this set grows, the likelihood of false positives grows in rough proportion. But with context-based inference, the set of likely matches will be considerably narrowed. Context-only inference gave 50% initial precision, and the precision decreased only slowly after that. That means the perplexity of context-inference is two for the first face, and grows only slowly for subsequent faces. So context inference can be applied first to generate a small set of candidate faces, and then computer vision can be applied to make a final selection.

Based on the matrix weights, the most useful prediction factors were: weekend/weekday, photo owner, indoor/outdoor, GPS location using k-means clustering, and PCA. Of these, photo owner and weekend/weekday were weighted strongest. This suggests that photo owner and weekend/weekday were good predictors overall. These features are readily available from mobile phone carriers (phone user and network time) and may therefore enable context-aware photo management service offerings (e.g., face recognition for cameraphone photos).

7. FUTURE WORK

Our future experiments will include expanded datasets and face detection for automated partitioning of training sets, and we will investigate torso-matching for detecting subjects in multiple photos taken at the same location and time [8]. We will also combine our context-aware face recognition research with our context-aware place recognition research to create a comprehensive solution for mobile media management [2].

8. REFERENCES

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

[2] Davis, M., Van House, N., Towle, J., King, S., Ahern, S., *et al*. MMM2: Mobile Media Metadata for Media Sharing. *Ext. Abstracts of CHI, Portland, Oregon*, 2005.

[3] Davis, M., King, S., Good, N., Sarvas, R. From Context to Content: Leveraging Context to Infer Media Metadata. In: *Proc. of ACM MM 2004 in New York, New York*, 2004.

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

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

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

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

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

[9] Van House, N.A., *et al*. Uses of Personal Networked Digital Imaging: An Empirical Study of Cameraphone Photos and Sharing. *Ext. Abstracts of CHI, Portland, Oregon, 2005.*