Automated Annotation of Human Faces in Family Albums

Lei Zhang
Microsoft Research Asia
49 Zhichun Road
Beijing 100080, China
+86-10-62617711
leizhang@microsoft.com

Longbin Chen*
Institute of Automation
Chinese Academy of Sciences
Beijing 2728#, 100080, P.R.China
+86-10-62647459
lbchen@nlpr.ia.ac.cn

Mingjing Li, Hongjiang Zhang
Microsoft Research Asia
49 Zhichun Road
Beijing 100080, China
+86-10-62617711
{mjli, hjzhang}@microsoft.com

ABSTRACT

Automatic annotation of photographs is one of the most desirable needs in family photograph management systems. In this paper, we present a learning framework to automate the face annotation in family photograph albums. Firstly, methodologies of content-based image retrieval and face recognition are seamlessly integrated to achieve automated annotation. Secondly, face annotation is formulated in a Bayesian framework, in which the face similarity measure is defined as *maximum a posteriori* (MAP) estimation. Thirdly, to deal with the missing features, marginal probability is used so that samples which have missing features are compared with those having the full feature set to ensure a non-biased decision. The experimental evaluation has been conducted within a family album of few thousands of photographs and the results show that the proposed approach is effective and efficient in automated face annotation in family albums.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications - *Computer vision*; I.4.10 [**Image Processing and Computer Vision**]: Image Representation - *Statistical*;

General Terms

Algorithms, Management, Experimentation.

Keywords

Face annotation, content-based image retrieval, face recognition.

1. INTRODUCTION

With the rapid development of digital cameras and scanners, digital photographs are becoming a commodity and automated tools for organizing these photographs become extremely desirable. Unfortunately, though there are many commercial products, annotating semantic content of photographs required in organizing photographs, a tedious task, is still left to users.

As a new research area, content-based image retrieval (CBIR) was motivated by the need to automatically organize and retrieve images from large collection. Partially because it targets at solving the general image retrieval problems, CBIR research efforts have not resulted in practical solutions to automatic family photograph management. However, we argue that to meet the special needs of family album management, we can apply some CBIR methodologies instead of waiting for CBIR to mature.

The most commonly used entries for indexing family photographs are related to *when, where, who* and *what*. With the advance in digital camera technology, date and time as well as location data is or will be readily available in cameras. In this paper, we focus on how to automatically extract "*who*" in family photographs.

To automatically annotate faces in images, face detection and recognition are two essential steps. Over the past few decades, face detection and recognition have been studied extensively by numerous researchers in computer vision. As a result, efficient and robust face detection algorithms [6, 8] have become available. In contrast, though there have significant progress in developing robust face recognition algorithms, the effectiveness of available algorithms are still limited to images of mug shots in which faces are mostly frontal and with reasonably homogenous lighting conditions and small variations in facial expressions [10]. The large variance in illuminations, poses, and expressions of face images in real life family photographs makes accurate face alignment difficult, thus, makes it difficult to accurately extract face features and establish face model of an individual. As a result, no effective face recognition solution has been developed for automated annotation of family photographs.

To overcome the difficulties in face annotation, we have proposed to introduce CBIR techniques into face annotation in a previous work [1]. However, this work is primarily focused on the CBIR technologies related to feature extraction and similarity measure, whereas face recognition technologies are not well integrated.

In this paper, we propose a new framework for semi-automated face annotation in family photo album applications. We have reformulated the face annotation from a pure recognition problem to a problem of similar face search and annotation propagation and have developed a solution to this problem by seamlessly integrating content-based image retrieval and face recognition algorithms in a Bayesian framework.

The rest of this paper is organized as follows. Section 2 presents the proposed framework, with a detailed description of the proposed algorithms. In Section 3, we describe the experiment setting and the evaluation result of the proposed algorithm. Thereafter, we will give concluding remarks in Section 4.

MM'03, November 2-8, 2003, Berkeley, California, USA.
Copyright 2003 ACM 1-58113-722-2/03/0011…$5.00.

* This work was performed at Microsoft Research Asia.

2. BAYESIAN FACE ANNOTATION

2.1 Framework

The user scenario we target at is in typical digital family photo albums, in which only a handful number of people, e.g. ten to fifty, are of our concern and appear frequently.

The framework of the proposed face annotation system is described as follows. First, a multi-view face detector [6] is used to detect faces in new uploaded images or images already in the album. The facial features are extracted from each detected face area, as well as the contextual features. To derive the similarity measure, a large set of training samples are collected offline to train the probability model for each feature. Then these probability models are integrated into a Bayesian framework to measure the similarity between faces. Based on this statistically derived similarity measure, the system generates a list of name candidates for a given query or new face by statistical learning approaches. Either selecting a name from the name list or setting a new name is allowed in the system. To further simplify the face annotation, similar face retrieval and relevance feedback are allowed for labeling multiple faces in a batch way.

2.2 Face Representation

Intuitively, face annotation can be considered as a pure face recognition problem and naturally the features and the algorithm developed for face recognition should be applied. However, as current face recognition algorithms are not robust enough for face annotation in family album systems, in the proposed face annotation framework, contextual features are incorporated in addition to those used in classical face recognition algorithms. Figure 1 illustrated the facial and contextual features used in the proposed framework and the regions from which they are extracted from.

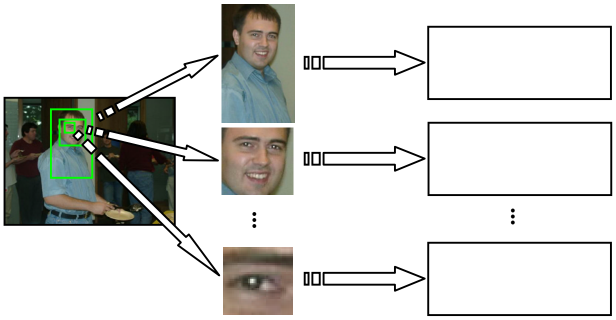

Figure 1. Feature extraction and face representation

2.2.1 *Face appearance feature*

Face appearance features are extracted based on face detection result. Because face appearance features are most reliable when extracted from frontal faces, poses of detected faces need to be accurately determined. For this purpose, the system first applies a texture-constrained active shape model (TC-ASM) [7] trained from frontal faces, and localizes facial points in each detected face. Then, the system calculates the texture reconstruction error of that face. As a strong prior model corresponding to frontal faces is built in the TC-ASM, if the texture reconstruction error of a face is lower than a certain threshold, the system identifies that face to be a frontal face; otherwise, the face appearance feature is treated as a missing feature.

2.2.2 *Contextual features*

The contextual features are extracted from the extended face region as shown in Figure 1. In family photo album scenarios, it is very common that an individual appeared in multiple photographs taken in the same day or event often wore the same clothes. Such contextual features could be used in distinguishing a limited number of individuals in the album.

It is worth noting that it is most likely meaningless to compare the contextual similarity between two pictures taken in two different days far from each other. Thus we restrict that the date difference between two photos must be within two days when comparing their similarity, otherwise the contextual features are treated as missing features.

2.3 Similarity Measure

As training samples for the probability estimation are severely limited for each individual, we cast the standard face recognition from a multi-class classification problem to a two class classification by introducing two classes of face variations, intra-personal variations Ω_{I} and inter-personal variations Ω_{E}, where Ω_{I} corresponds to difference appearances of the same individual and Ω_{E} corresponds to variations between different individuals. The idea is somewhat similar to that Moghaddam proposed in [3]; however, we have to deal with more complex situations that multiple features need to be combined and some of the features may be missing.

Let $\boldsymbol{F}=\{\boldsymbol{f}_i \mid i=1,\ldots,N_f\}$ denote the features set, where each feature $\boldsymbol{f}_i$ is a vector corresponding to a specific feature described in the previous section. By introducing two mutually exclusive classes Ω_{I} and Ω_{E}, the similarity between two faces is defined as:

$$S(\boldsymbol{F}_1,\boldsymbol{F}_2)=p(\Delta\boldsymbol{F}\in\Omega_{\mathrm{I}})=p(\Omega_{\mathrm{I}}\mid\Delta\boldsymbol{F}) \qquad (1)$$

where $\Delta\boldsymbol{F}=(\boldsymbol{F}_1-\boldsymbol{F}_2)$ is the difference between two face features. Using the Bayesian rule, (1) can be rewritten as:

$$S(\boldsymbol{F}_1,\boldsymbol{F}_2)=\frac{p(\Delta\boldsymbol{F}\mid\Omega_{\mathrm{I}})p(\Omega_{\mathrm{I}})}{p(\Delta\boldsymbol{F}\mid\Omega_{\mathrm{I}})p(\Omega_{\mathrm{I}})+p(\Delta\boldsymbol{F}\mid\Omega_{\mathrm{E}})p(\Omega_{\mathrm{E}})} \qquad (2)$$

where $p(\Omega_{\mathrm{I}})$ and $p(\Omega_{\mathrm{E}})$ are the *a priors* which can be estimated empirically. $p(\Delta\boldsymbol{F}\mid\Omega_{\mathrm{I}})$ and $p(\Delta\boldsymbol{F}\mid\Omega_{\mathrm{E}})$ are the likelihoods for a given difference $\Delta\boldsymbol{F}$, and can be estimated from a large set of training data which do not depend on any particular individual.

As the features are extracted from different areas, or of different types, such as color correlogram or appearance gray texture patches, we consider them to be independent with each others. Denote the feature difference by $\Delta\boldsymbol{f}_j=(\boldsymbol{f}_{j1}-\boldsymbol{f}_{j2})$ for each feature between two faces, (2) can be further expressed as:

$$S(\boldsymbol{F}_1,\boldsymbol{F}_2)=\frac{\prod_{j=1}^{N_f}p(\Delta\boldsymbol{f}_j\mid\Omega_{\mathrm{I}})p(\Omega_{\mathrm{I}})}{\prod_{j=1}^{N_f}p(\Delta\boldsymbol{f}_j\mid\Omega_{\mathrm{I}})p(\Omega_{\mathrm{I}})+\prod_{j=1}^{N_f}p(\Delta\boldsymbol{f}_j\mid\Omega_{\mathrm{E}})p(\Omega_{\mathrm{E}})} \qquad (3)$$

This similarity function integrates multiple features into a Bayesian framework. Based on this similarity measure, name candidates for a given unknown face can be derived by statistical learning approaches, such as K nearest neighborhood algorithm.

2.4 Dealing with Missing Features

If all features for any two given faces are available, their similarity can be directly calculated according to (3). However, as we mentioned in Section 2.2, each feature has its particular applicable range, and it is very common that some of the features are not available in some images in the family album scenario. In such cases, we have to deal with the ranking problem, in which a set of samples can be ranked according to their similarities to a given query sample, no matter whether there are any missing features for each sample.

To simplify the derivation and without losing generality, suppose that there are only two features in (3), facial appearance feature and body color feature. The corresponding feature differences are denoted as *f* and *b*. Note that the symbol Δ is ignored for simplicity. If *f* is unknown or missing, the Bayesian rule is applied again to derive how to measure the similarity given the known feature *b*. We can rewrite (3) as follows:

$$S(\boldsymbol{F}_1,\boldsymbol{F}_2)=\frac{p(f\,|\,\Omega_\mathrm{I})p(b\,|\,\Omega_\mathrm{I})p(\Omega_\mathrm{I})}{p(f\,|\,\Omega_\mathrm{I})p(b\,|\,\Omega_\mathrm{I})p(\Omega_\mathrm{I})+p(f\,|\,\Omega_\mathrm{E})p(b\,|\,\Omega_\mathrm{E})p(\Omega_\mathrm{E})} \tag{4}$$

$$=\frac{p(\Omega_\mathrm{I}\,|\,f)\dfrac{1}{p(\Omega_\mathrm{I})}p(b\,|\,\Omega_\mathrm{I})p(\Omega_\mathrm{I})}{p(\Omega_\mathrm{I}\,|\,f)\dfrac{1}{p(\Omega_\mathrm{I})}p(b\,|\,\Omega_\mathrm{I})p(\Omega_\mathrm{I})+p(\Omega_\mathrm{E}\,|\,f)\dfrac{1}{p(\Omega_\mathrm{E})}p(b\,|\,\Omega_\mathrm{E})p(\Omega_\mathrm{E})}$$

Intuitively, given that f is unknown, the posterior probability $p(\Omega_\mathrm{I}\,|\,f)$ can be replaced with the prior probability $p(\Omega_\mathrm{I})$, and $p(\Omega_\mathrm{E}\,|\,f)$ can be replaced with $p(\Omega_\mathrm{E})$, as the only information we are allowed to use for $p(\Omega_\mathrm{I}\,|\,f)$ and $p(\Omega_\mathrm{E}\,|\,f)$ is the value of the prior probabilities. Therefore the similarity function can be further simplified as if there is no feature *f*:

$$S(\boldsymbol{F}_1,\boldsymbol{F}_2)=\frac{p(b\,|\,\Omega_\mathrm{I})p(\Omega_\mathrm{I})}{p(b\,|\,\Omega_\mathrm{I})p(\Omega_\mathrm{I})+p(b\,|\,\Omega_\mathrm{E})p(\Omega_\mathrm{E})} \tag{5}$$

2.5 Likelihood Estimation

For each feature in (3), the conditional probabilities $p(\Delta\boldsymbol{f}_j\,|\,\Omega_\mathrm{I})$ and $p(\Delta\boldsymbol{f}_j\,|\,\Omega_\mathrm{E})$ can be estimated offline from a large set of training data by either eigenspace [3] or SVM [5] approaches.

For most features, Gaussian assumption is too restricting. Although Gaussian mixture model can be adopted, the cluster number is a magic one which is difficult to estimate. Instead, we prefer casting the SVM classifier output $f(\Delta\boldsymbol{f})$ to the posterior probability [5]. Instead of estimating the class-conditional densities $p(\Delta\boldsymbol{f}\,|\,\Omega)$, a parametric model is adopted to fit the posterior $p(\Omega_\mathrm{I}\,|\,f(\Delta\boldsymbol{f}))$ directly.

$$p(\Omega_\mathrm{I}\,|\,f(\Delta\boldsymbol{f}))=\frac{1}{1+\exp(A\Delta\boldsymbol{f}+B)} \tag{11}$$

To apply this posterior probability in the similarity function (3), similar derivations can be utilized as in (4).

2.6 Candidate Name List Generation

As we discussed in Section 2.3, given a number of labeled faces, the goal of the learning algorithm of face annotation is to generate a candidate name list for an unlabeled face, which is sorted according to the similarity between the unlabeled face and the labeled faces. Based on the nearest neighbor or *K*-nearest neighbor algorithms, given an unknown face, among its *K* nearest labeled faces, the name candidates can be generated by sorting the names according to the sum of similarities to the unknown face.

2.7 Batch Annotation

Given the face similarity measure function, the system also provides the function of similar face retrieval. In this way, users are allowed to search similar faces by specifying either a face or a name and then annotate multiple faces in a batch way. If a name is specified, the system will first expand the query to all the faces labeled as that name, and then search other similar faces based on multi-instance search algorithm. That is, all faces in the album are ranked according to the maximum similarity to each query faces. Furthermore, relevance feedback techniques can also be introduced to refine the retrieval result.

3. EXPERIMENTS

In this section, we present experimental evaluations of the proposed face annotation framework with a large set of family photographs that contain at least one face.

3.1 Implementation Issues

In our experiments, two features are combined to improve the face annotation. One is color and texture feature used as contextual feature and the other is face appearance feature. These two features are generally considered to characterize the technologies from CBIR and face recognition.

The contextual feature in the experiment is the 50 dimensional banded auto-correlogram [2] and 14 dimensional color texture moment [9]. We first extend each face region and divide the extended region into two blocks to include face and body, respectively, and then extract two 64 dimensional color and texture features in these two blocks. We refer this feature as body feature for simplicity.

To estimate the posterior probability of intra- and inter-personal variations of body features, we labeled a small family album with about 200 images taken by digital cameras. Among these images, about 967 pairs of intra-personal features and 2865 pairs of inter-personal features are extracted as training data. Each pair is extracted from two images taken within 24 hours. Then an SVM classifier is trained from the training data and the classification output is fitted to the posterior probability $p(\Omega_\mathrm{I}\,|\,f(\Delta\boldsymbol{b}))$ [5].

The face appearance feature is extracted from the cropped face image of size 25x45. To estimate the distributions of intra- and inter-personal variations, we used FERET [4] database as the training data. 600 difference face images are combined together to calculate the eigenfaces. Then the difference images are projected to the first 71 eigenfaces. We refer this feature as face feature for simplicity. SVM is applied again to derive the posterior probability $p(\Omega_\mathrm{I}\,|\,f(\Delta\boldsymbol{f}))$ of face appearance feature.

3.2 Test Dataset

The test data set we used for the performance evaluation in our experiment is a typical family album, consisting of about 1030 photographs taken by digital cameras. The photographs in the album are taken during 12 months. The scenes in the album include various kinds of events such as birthday party, wedding, family gathering, and sightseeing.

There are so many individuals appeared in the album. We labeled 27 most frequently appeared individuals as the ground truth because other individuals only appeared for no more than 3 times.

3.3 Performance Measure

To evaluate the performance of the proposed face annotation algorithm, we have proposed to use the measure of H-Hit rate [1]. Suppose that faces in the album are annotated sequentially in a random order. Given an input unlabeled face, the system will generate a list of H name candidates based on the historical labeled faces. If the true name of this face is in the list, we call that the face is hit by the name list. Thus we can calculate the average hit rate of all the faces in the album, given the length H of the candidate name list.

3.4 Performance Evaluation

In the experiment, we simulated the users' annotation process by annotating faces one by one in a random order. The system generates name candidates based on previous labeled faces, and calculates the H-Hit rate automatically based on the ground truth.

Four schemes are compared and the results are shown in Figure 2. The first scheme called "*L2*" is based on body feature only, but we use Euclidean distance (L2 norm) to measure the similarity between two faces. This scheme is the same as that presented in [1] and considered as the baseline. The second scheme called "*prior*" is based on the prior probability estimation. That is, given an unknown face, the name list is generated based on the prior of each individual estimated from the historical labeling result. This scheme can be considered as the simplest approach. The third scheme called "*body*" and the fourth scheme called "*body+face*" are both based on the proposed new probability framework. In the third scheme, the similarity measure depends only on the body feature. In the fourth scheme, the similarity measure depends on both body features and face features.

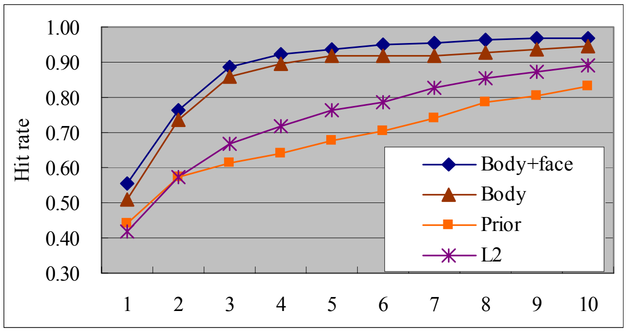

Figure 2. Performance comparison of different schemes. The horizontal axis represents the number of names in the name list.

As there are 27 individuals labeled in the album, on average at least 37% (=10/27) 10-Hit rate can be obtained even if 10 name candidates are randomly provided. Such *random* scheme can be easily overcome by the simplest scheme *prior*. And by introducing color and texture features extracted from the extended face region, the scheme *L2* outperforms the scheme *prior*.

However, from the results, it can be seen that the third and the fourth schemes based on the proposed Bayesian framework are much more effective than both *prior* and *L2*. The results are very promising in that the 5-Hit rate is approaching to 94%, which means that users could annotate most of the faces only by clicking the mouse, choosing the name from a list of no more than five name candidates. And it can be seen that by combining face recognition technologies, the scheme *body+face* still increases the H-Hit rate for about 3%-5%. Note that the face recognition we adopted here is merely a very simple implementation of a classical approach, in which only 300 intra- and 300 inter-personal variation samples are collected for training and these training samples are totally independent with the family photographs used in the evaluation.

4. CONCLUSION AND FUTURE WORK

We have presented a Bayesian framework in this paper to automate the process of face annotation in family photo albums. In this framework, methodologies of both content-based image retrieval and face recognition are seamlessly combined and high accuracy is achieved using a statistically learned similarity measure. In particular, since the framework is able to deal with missing features, various algorithms can be integrated by identifying when to use an algorithm so that their weaknesses can be avoided. The experimental evaluation shows that the framework is effective and efficient. With significantly improved accuracy in candidate recommendation, users' labeling efforts are greatly reduced.

5. REFERENCES

[1] Chen L., Hu B., Zhang L., Li M. and Zhang H.J., "Face annotation for family photo album management", International Journal of Image and Graphics, p.1-14, Vol. 3, No. 1, 2003.

[2] Huang J., Kumar S. R., Mitra M., Zhu W. J. and Zabih R., "Image indexing using color correlograms", In IEEE Conf. on Computer Vision and Pattern Recognition, p. 762, 1997.

[3] Moghaddam, B., Jebara, T. and Pentland, A., "Bayesian Face Recognition", Pattern Recognition, Vol 33, Issue 11, p.1771-1782, 2000

[4] Phillips P., Moon H., Rizvi S. and Rauss P., "The FERET evaluation methodology for face-recognition algorithms", IEEE Trans. PAMI, vol.22, p.1090-1103, 2000

[5] Platt J. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In Advances in Large Margin Classifiers. MIT Press, 1999.

[6] Xiao R., Li M.J., Zhang H.J., "Robust Multi-Pose Face Detection in Images", to be appeared in IEEE Trans. on CSVT Special Issue on Biometrics, 2003

[7] Yan S.C., Liu C., Li S.Z., et al., "Texture-Constrained Active Shape Models", In Proc. International Workshop on Generative-Model-Based Vision, Denmark. May, 2002.

[8] Yang M. H., Kriegman D. and Ahuja N., "Detecting Faces in Images: A Survey", IEEE Trans. PAMI, p. 34, 24(1),2002

[9] Yu H., Li M., Zhang H. and Feng J., "Color texture moment for content-based image retrieval", Proc. IEEE Intl Conf. on Image Processing, September, 2002

[10] Zhao W., Chellappa R., Rosenfeld A. and Phillips P., "Face recognition: A literature survey", Technical Report, Maryland University, CfAR CAR-TR-948, 2000