

FACE ANNOTATION FOR FAMILY PHOTO ALBUM MANAGEMENT*

LONGBIN CHEN[†] and BAOGANG HU[‡]

*Institute of Automation, Chinese Academy of Sciences,
P.O. Box 2728, Beijing 100080, China*

[†]lbchen@nlpr.ia.ac.cn

[‡]hubg@nlpr.ia.ac.cn

LEI ZHANG,[§] MINGJING LI[¶] and HONGJIANG ZHANG^{||}

Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, China

[§]l-zhang@microsoft.com

[¶]mjli@microsoft.com

^{||}hjzhang@microsoft.com

Received 7 June 2002

Revised 30 September 2002

In this paper, we propose a framework to semi-automatically annotate faces in family photo albums. The core of the framework is the features used to define face similarity and this results in the learning algorithm used to refine automatic face annotation. We have adopted similarity based search and relevance feedback ideas developed for content-based image retrieval and a set of simple yet effective color and texture based features, in addition to the traditional face recognition features, in performing candidate annotation search. The experimental evaluation of the proposed approach has been conducted with a family album of 1707 photos and the results show that the proposed approach is an effective and efficient one for semi-automatic family photo album annotation.

Keywords: Face Annotation; Face Recognition; Content-based Image Retrieval; Photo Album.

1. Introduction

With the rapid development of digital cameras and scanners, more and more people are taking digital photos and scanning printed photos into their computers. Therefore, there is a strong need of digital photo album tools to help people organize and manage their digital photos. Though there are already many commercial products for this purpose, they all require human annotations, a tedious task that very few will take. It is highly desirable to automate this indexing and organizing process.

*This work was performed at Microsoft Research Asia.

This is one of the key objectives that have motivated content-based image retrieval (CBIR) research efforts in the last ten years.¹ Partially because it targets the general image retrieval problems which require the capability of semantic understanding of image content, CBIR research efforts have not resulted in effective tools useful in automatic family photo organization and management. However, we argue that to meet the special needs of family album management, one does not have to wait for CBIR to solve all the problems. What we need are specially designed tools for this application.

The most commonly used entries for indexing family photos are related to when, where, who and what. That is, our visual memory associated with a photo is when and where the photo was taken, and who is in the photo and in what event. With the advance in digital camera technology, digital photos come with date and time data in most cameras. Very soon, embedded GPS in cameras will be able to provide the location data of a photo as well. Therefore, the most desirable photo index data left for automated extraction is the annotating of “who is in each photo.”

To automatically annotate the faces in the photos, face detection and recognition are the two essential steps. Over the past 20 years, face detection and face recognition have been studied extensively in the computer vision and machine learning field.² There have been a large number of works on face detection. With the advances of the research targeted on this problem, especially with the availability of robust face detection algorithms,^{3–5} the task of face annotation does not seem as formidable as before.

However, as compared to face detection, automated face recognition is a much more difficult problem.^{6,7} The difficulty of face recognition lies in the complexity of human perception. It is still not clear how people recognize the faces, and how they remember the faces they ever saw.⁶ Algorithm-wise, because of the large variance in illuminations, poses, and expressions of faces in real life photos that often break face alignment algorithms, it is difficult to extract accurate face features and to establish an accurate face model of a person. In the family album scenario, it is also not feasible ask a user to collect a large number of well aligned sample faces to train an accurate model of a person. This fact reduces the potential of face recognition algorithms that requires large training data in family photo album applications. As a result, no effective face recognition solution has been developed for the automated annotation of family photos.

Face annotation is a relatively new topic in the field of face detection and recognition. In 1993, Gudivada proposed a framework to retrieve face images from face database based on some semantic attributes.⁸ The focuses of that work was on the retrieval method and face database was used merely as another test data. The MiAlbum system searches images in the photo album using keywords and low level features.⁹ It also provides a function to detect the faces in the image, but it does not provide automated face annotation functions. Baker built a mug-shot search system that adopts the CBIR techniques in searching faces based on eigenfaces in the face database.¹⁰ However, none of these systems provides a robust solution to automated face annotation.

In this paper, we propose a framework of automated or semi-automated face annotation of family photos. The goal of our system is to provide a semi-automatic way for face annotation in a typical digital family photo album, which usually contains a limited number, say, 10–50 people appearing frequently. The user scenario we target at is the following: When a face is detected in a photo and a user would like to annotate faces, the system will calculate a candidate list of names for the user to annotate the detected face, according to the similarities to the annotated faces. The user might accept the recommendation, or set a new name to that face. Furthermore, the user might also annotate multiple faces in a batch way by the similar face retrieval. The framework is designed to support this scenario. That is, when the user moves the mouse onto a detected face in a photo, a tooltip will popup to suggest the name of the person if the face has been annotated at least once before; otherwise a list of candidate names will appear to help the user annotate. Furthermore, to facilitate the annotation of all the faces in a photo album, a similar-face retrieval tool is provided to help the user annotate multiple faces conveniently.

To achieve the goal, we not only employ the advanced techniques of face detection and recognition to detect the faces and extract the face related features, we also adapt the content based image retrieval techniques and relevance feedback techniques,¹¹ which is suitable for the learning problem with limited samples. In other words, we divide the face annotation into two key issues: feature extraction and learning-based similarity search. We have implemented a face annotation system with this framework. To compare the performance of different features and learning approaches, we also propose a performance measure of the face annotation approaches.

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

2. Face Annotation Framework

2.1. System overview

The framework of our face annotation system is illustrated in Fig. 1. First, a face detector is used to detect all the faces in the photo album or new uploaded photos. The positions of the detected faces are stored and used to crop image patches of faces and bodies (see Appendix A). Then, the face-related feature is extracted for annotation and retrieval. These features include the color, texture and eigen-face feature in the face area. After the feature extraction, the users could annotate the faces, or find the similar faces in the photo album (see Appendix B).

A real-time face detector⁶ is utilized to detect the faces in the photo album. Although the accuracy of the detection result cannot reach 100%, which means that some faces will be missed and some false alarms will occur, the detection result, nevertheless, is a good starting point to build up a practical system. The

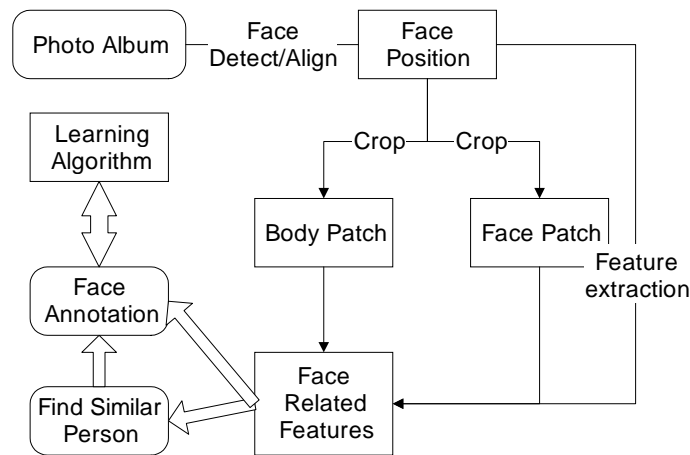
4 *L. Chen et al.*

Fig. 1. Framework overview.

detector we employed here can detect about 85% of the faces in the album. For the missed faces, a friendly user interface can help to locate them through the users' interaction. As the user usually does not care about the false alarms and will not annotate them, these false alarms will never affect the learning performance.

Since the face detection method is not perfect, especially when the detected position is not always at the center of the faces, the face alignment is an important step for feature extraction and thus is able to improve on the accuracy of face recognition.¹²

The user interface in the previous face retrieval system is very complicated, requiring the user to specify which parts of retrieval result are correct or wrong. The users must pay more attention on the result and their patience will soon run out. Furthermore, for non-expert users, the frequently occurred errors will deteriorate greatly the retrieval result. In our system, we try to simplify the user interface to make the annotation easier. If the user moves the mouse onto an annotated face, a tooltip will popup to show the name of this person. If the face has not been annotated before, a popup menu with a list of candidate names will appear. The user can select one of the names to annotate the face, or input a new name (see Fig. 2). To further facilitate the face annotation, the user can specify a face and search the similar faces in the album, then annotate multiple faces in a batch way. Moreover, the relevant feedback⁸ can be used to improve the search result based on the users' interaction.

2.2. Feature extraction

The performance of this semi-automatic annotation tool depends primarily on two issues besides the performance of face detection. One is the feature extraction and the other is the learning algorithm. Intuitively, we can consider it a face recognition



Fig. 2. A popup menu in the face annotation system.

problem and naturally the features and the techniques used for face recognition should be considered first. However, the face recognition technique is not robust enough for face annotation in family album systems. Most of face recognition algorithm can only deal with the frontal face under ideal illumination, whereas in a family photo album, the poses and illuminations of faces are of great variance. Such large variance makes the extraction of face features unreliable. To improve the performance, other face-related features and image-related features should be incorporated in addition to the face recognition features.

We compared several features used in CBIR, such as color histogram, color moment, correlogram and wavelet texture. Color histogram,¹³ which donates the intensities of RGB channels, is mostly used to represent the color information. Color moment is proposed by Stricker *et al.* to avoid the quantization effects in color histogram.¹⁴ The first three moments (mean, standard deviation and skews) are extracted from the three color channels (HSV space) and therefore form a 9-dimensional feature vector. Auto-correlogram, proposed by Huang *et al.*, is very effective in combining the color and texture features together.¹⁵ We quantize the RGB color space into $4 \times 4 \times 4 = 64$ colors. Then we use the distance set $D = \{1, 3, 5, 7\}$ for computing the auto-correlogram to build a 256-dimension feature. In early 1990's, the wavelet transform was introduced and was applied in various fields.¹⁶ To extract the wavelet based texture, the original image is decomposed into 10 de-correlated sub-bands through 3-level wavelet transform. In each sub-band, the standard deviation of the wavelet coefficients is extracted, resulting in a 10-dimensional feature vector.

Although these features are the low-level features, they are useful in face annotation and they achieve good results in our experiments.



Fig. 3. Block features.

To capture the structural information of the face and body patches, we propose the use of local regional features to capture the spatial information. In our system, a facial area is represented in terms of 4×2 blocks and the features are extracted from these local blocks. (see Fig. 3). In this paper, the 8-block wavelet and 8-block correlogram features are used with 144 and 2048 dimensions respectively.

To extract the contextual information of a person, the face region is extended to the body region. The extended region usually includes the clothes of the person. The body-related feature is useful when a lot of photos are taken under the same scene where the person wears the same clothes. The color and texture of the body region is suitable to capture the body information.

Note that it is critical that the features can be calculated fast and are easy to represent in a practical system.

2.3. Learning algorithm

Learning algorithm is another important issue to give a good candidate name list with respect to the features extracted from the face and the surrounding regions. In content-based image retrieval, the feedback of users to the query result is used for further query. It is called relevance feedback technique.¹¹ We adopt a similar technique in face annotation. Given a number of annotated faces belonging to several persons, the goal of the learning algorithm, for an unlabeled face, is to generate a candidate name list, which is sorted according to the similarities between the unlabeled face and the labeled faces. If several persons' faces are annotated, it can be regarded as a multi-class classification task to give a candidate name list: The requirement of learning algorithm are computational speed and the storage of learning result. Therefore, in this paper, we focus mainly on some fast algorithms, such as Nearest Neighbor (NN) and K Nearest Neighbor (KNN).

Beside the facial features and body-related features, the date and time information of the photos as well as the distribution of faces are very useful. For the photos

taken by digital images, the date and time could be extracted from the meta-data stored in the image file. For the other photos uploaded from the scanners, we can either extract the printed time stamp from the image⁹ or use the images' created time. Based on the time constrain and the low level features of the images, the photos can be grouped into several events.¹⁷ Usually the person in the same event will wear the same clothes. Combined with the date constraints, the body-related features will play an important role in the recognition.

The frequency of a person appearing in the photo album is a kind of *prior* information which could be incorporated into the learning process. It is reasonable to assume that a person will appear more frequently if he/she appears much more before. So we can count the previous appearing times for each annotated person as the estimation of this person's appearing frequency. With more and more faces being annotated, the estimation will be more and more accurate. After this *prior* is estimated, a Bayesian learning framework can be used to give a more accurate prediction to annotate the face.

Actually, the *prior* distribution of the faces has already been incorporated in the algorithm of nearest neighbor.¹⁸ Suppose there are C persons in the photo album denoted as w_1, w_2, \dots, w_C , and there have L faces that have been annotated, the corresponding face features will be $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_L$. Given a new face with the feature value \mathbf{f} , the expected error rate $\mathbf{P}_L(e)$ is:

$$P_L(e) = \iint P_L(e|\mathbf{f}, \mathbf{f}') p(\mathbf{f}'|\mathbf{f}) d\mathbf{f}' p(\mathbf{f}) d\mathbf{f}, \quad (1)$$

where \mathbf{f}' is the nearest neighbor of \mathbf{f} , $\mathbf{P}_L(e|\mathbf{f}, \mathbf{f}')$ denotes the error rate for the prediction of \mathbf{f} , i.e. \mathbf{f} and \mathbf{f}' belong to the different classes, $p(\mathbf{f}'|\mathbf{f})$ denotes the probability that given a sample \mathbf{f} , its neighbor is \mathbf{f}' . It can be proved that when $L \rightarrow \infty$, there is¹⁸:

$$\lim_{L \rightarrow \infty} P_L(e|\mathbf{f}) = 1 - \sum_i P^2(w_i|\mathbf{f}), \quad (2)$$

and the average error rate P is:

$$\begin{aligned} P &= \lim_{L \rightarrow \infty} P_L(e) \\ &= \int \lim_{L \rightarrow \infty} P_L(e|\mathbf{f}) p(\mathbf{f}) d\mathbf{f} \\ &= \int \left[1 - \sum_{i=1}^c P^2(w_i|\mathbf{f}) \right] p(\mathbf{f}) d\mathbf{f}, \end{aligned}$$

where $\mathbf{P}(w_i|\mathbf{f})$ is the distribution of \mathbf{f} with its classes.

$$P(w_i|\mathbf{f}) = \frac{P(w_i, \mathbf{f})}{p(\mathbf{f})} = \frac{P(\mathbf{f}|w_i)p(w_i)}{\sum_{\mathbf{k}} p(\mathbf{f}|w_{\mathbf{k}})p(w_{\mathbf{k}})}. \quad (3)$$

From Eq. (3), we can see that the average error rate depends on the distribution of face classes $P(w_i)$, which means that it is not necessary to consider the *prior*

8 *L. Chen et al.*

distribution of different faces in the KNN algorithm. What's more, we can have an estimation of P as P satisfies¹⁸:

$$P^* \leq P \leq P^* \left(2 - \frac{C}{C-1} P^* \right),$$

where P^* is the Bayes error rate:

$$P^* = \int P^*(\mathbf{e}|\mathbf{f})p(\mathbf{f}) d\mathbf{f},$$

here $P^*(\mathbf{e}|\mathbf{f})$ is the Bayes condition error rate:

$$P^*(\mathbf{e}|\mathbf{f}) = 1 - \max_i [P(w_i|\mathbf{f})].$$

3. Experiment Evaluation

3.1. Performance measure

To evaluate the performance of the face annotation, we propose two performance measures. The first measure is the H -Hit rate. Let $F = \{f_1, f_2, \dots, f_N\}$ denote all faces in the photo album \mathbf{D} , and suppose that the faces in the album are sorted one by one in some order. In our experiments, we sorted all the faces by the order that they are detected. To give an estimation of the prediction performance, it is assumed that the faces are annotated by this same order, i.e. f_k will not be annotated until all previous faces $\{f_1, f_2, \dots, f_{k-1}\}$ are annotated and the prediction of f_k will be based on the previous faces. For the face f_k that will be annotated, the system will generate a list of H candidate names based on the previous annotated faces. If the true name of face f_k is in the list, we call this a *successful prediction* and the face is H -Hit by the name list. So we can calculate the average H -Hit rate of all the faces in the album, given the length H of the candidate name list:

$$H\text{-Hit} = \frac{1}{N} \sum_{f \in D} \text{hit}_H(f), \quad (4)$$

where N is the number of the faces in the image database \mathbf{D} . $\text{Hit}_H(f)$ is 1 if f_k is H -Hit by the name list of H names, and 0 otherwise.

As users would not like to choose the name from a long list, so the H -Hit is meaningful only for small numbers such as 1, 2, 5, 10.

The H -Hit rate gives an estimation of the prediction performance for a photo album, while another measure, H -Hit $_L$ rate, is proposed to evaluate the prediction performance with the number of labeled samples. Suppose the faces in the image database \mathbf{D} is split into two parts, training part \mathbf{D}_1 and testing part \mathbf{D}_2 , let N_1 and N_2 be the number of faces in \mathbf{D}_1 and \mathbf{D}_2 , respectively. All the faces in \mathbf{D}_1 have been annotated correctly and all the faces in \mathbf{D}_2 are waiting to be annotated. Obviously, the more faces annotated in the image database, the more accurate the

prediction will be. To measure the prediction accuracy with respect to different percentages of annotated faces, we define the performance measure, H -Hit $_L$, as:

$$H\text{-Hit}_L = \frac{1}{N_2} \sum_{f \in \mathbf{D}_2} \text{hit}_{H, \mathbf{D}_1}(f), \quad (5)$$

where L is the percentage of annotated faces, i.e. N_1/N , $\text{hit}_{H, \mathbf{D}_1}(f)$ is 1 if f is hit by the name list of H names, given \mathbf{D}_1 , and 0 otherwise.

Other performance measures, such as the calculation speed, are important as well. A fast algorithm is necessary because users may not be patient enough to wait for a long time before a candidate name list appears. Both the features and the learning algorithm will affect the calculation speed.

3.2. Photo album

The photo album we used in our experiment is a typical family album. There are a total of 1707 photos in the album. The scenes in the album include all kinds of events such as graduation ceremonies, birthday parties, family gathering, visiting and so on. Figure 4 is a typical photo in the album where the faces are detected automatically and are bounded by the white rectangles.

There are many people in the album. We selected 15 people who appeared more than five times as our ground truth.

3.3. Features versus hit rate

In this experiment, we simulate the users' annotation process and calculate the H -Hit rate [see Eq. (1)] automatically based on the annotated ground truth. Various



Fig. 4. Face detection result.

Table 1. Comparison of different features.

Feature name	Size	Accuracy rate		
		1-Hit	2-Hit	5-Hit
Color moment	9	0.437321	0.744498	0.986603
Wavelet	18	0.465072	0.727273	0.985646
Correlogram	256	0.466986	0.761722	0.991388
Block wavelet	144	0.645933	0.872727	0.987560
Block correlogram	2048	0.528230	0.784689	0.986603
Eigen-face value	32	0.490909	0.775120	0.987560

features, including the eigen-face values, are used and the 1-Hit, 2-Hit, 5-Hit rates are calculated and shown in Table 1. From the table, we can see clearly that the block wavelet feature that we proposed performs best in terms of 1-Hit and 2-Hit rate. For 5-Hit rate, all features perform quite well and the hit rates approximate to about 99%, which means that the user could annotate most of the faces only by clicking the mouse, choosing the name from a list of no more than five people.

3.4. Learning algorithm versus hit rate

The Hit rate varies with different learning algorithm. In this experiment, we mainly compared the performance of different learning algorithms, NN and KNN ($K = 10, 20$). The selection of K in the algorithm will affect the result of prediction accuracy (see Table 2). KNN algorithm performs better for larger H -Hit measures. For instance, the 20-nearest neighbor algorithm performs best in terms of 5-Hit rate. It is reasonable because when we try to find the more possible candidate names in the database, we should try to find more faces in the neighbor of the target face. For both the NN and KNN algorithm, the calculation speed is quick enough to build a real time system. We cannot perceive any delay when we run the three algorithms in our system.

3.5. Body feature performance

For the photos taken in the same day or in a short period of time, the person tends to wear the same clothes, which is a prominent feature in distinguishing him. Some image folders, in which the photos are taken in the same day or same event, are

Table 2. Comparison of different learning algorithms.

Learning algorithm	Hit rate		
	1-Hit	2-Hit	5-Hit
Nearest neighbor	0.645933	0.872727	0.987560
10-nearest neighbor	0.620858	0.885965	0.980507
20-nearest neighbor	0.600390	0.875244	0.998051

Table 3. Body feature versus facial feature.

Image folder	Person number	Facial feature			Body feature		
		1-Hit	2-Hit	3-Hit	1-Hit	2-Hit	3-Hit
Birth	5	0.580645	0.838710	0.903226	0.629032	0.903226	0.983871
Picnic	3	0.750000	0.916667	1.000000	0.833333	0.916667	1.000000
Party	8	0.750000	0.927083	0.989583	0.635417	0.958333	1.000000

selected to illustrate the effectiveness of the body features for face annotation in an event. We compared the facial feature and the body features, using the nearest neighbor algorithm. For facial features, the block wavelet features is used because it has the best performance as shown in Table 1. For body features, the block correlogram feature is used because we want to capture more color information on clothings.

From Table 3, we can clearly see that the body features performance is much better than facial features in the same event. Consequently, if combined with time constraint, the body features will outperform the facial features in the large photo album.

4. Conclusion and Future Work

In this paper, we have presented an effective and efficient face annotation framework for family photo organization and management. The framework incorporates face detection, face recognition and contented-based image retrieval techniques. In addition to the classical face feature extraction algorithms, we have incorporated additional supporting features to improve the annotation accuracy. The experimental evaluation shows that the framework is effective and efficient, with significantly improved accuracy in candidate recommendation and retrieval.

A remaining issue is how to incorporate the time information into the annotation process in a more effective way. If all the photos in a photo album are sorted by date and time, the persons appearing in the photo sequence could be regarded as a stochastic process. It is reasonable to assume that a face that has appeared frequently and recently will appear with a large probability in the near future. It is also reasonable to assume that some faces appearing in the same photo frequently will appear in another photo later with a large probability. These time-related features will greatly improve the prediction and retrieval accuracy in our system. We will carry out the research along this direction in the future.

Acknowledgment

The authors would like to thank Rong Xiao, Long Zhu of Face Group in Microsoft Research Asia for providing the face detector and Yi Zhou, Shuicheng Yan for providing the face alignment tools used in the face annotation system.

Appendix A. Face Patch and Body Patch



Fig. A.1. Face patches and body patches.

Appendix B. Similar Face Retrieval Result

Figure B.1 shows the target face and the first eight resulting photos. The caption below the photo are the file names and the distance values between the resulting photos and the target photo.



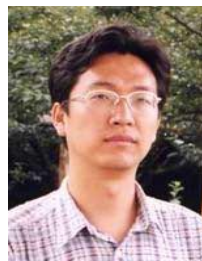
Fig. B.1. Similar face retrieval result.

References

1. Y. Rui, T. S. Huang, and S. Chang, "Image retrieval: current techniques, promising directions and open issues," *J. Visual Commun. Image Representation* **10**, 39 (1999).
2. M. H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Trans. Pattern Anal. Machine Intelligence (PAMI)* **24**(1), 34 (2002).
3. P. Viola and M. J. Jones, "Robust real-time object detection," *Tech. Rep.* (COMPAQ Cambridge Research Laboratory, Cambridge, MA, February 2001).
4. S. Z. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum, "Statistical learning of multi-view face detection," In *Proc. 7th European Conf. Computer Vision* (Copenhagen, Denmark, May 2002).
5. R. L. Hsu, M. A. Mottaleb, and A. K. Jain, "Face detection in color images," In *IEEE Trans. Pattern Anal. Machine Intelligence* **24**(5), 969 (2002).
6. R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," In *Proc. IEEE* **83**, 705 (1995).
7. W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips, "Face recognition: a literature survey," *Tech. Rep.* (Maryland University, CfAR CAR-TR-948, 2000).
8. V. N. Gudivada, V. V. Raghavan, and G. S. Seetharaman, "An approach to interactive retrieval in face image databases based on semantic attributes," In *Proc. Third Ann. Symp. Document Anal. Information Retrieval* (Las Vegas, 1993).
9. W. Liu, Y. Sun, and H. Zhang, "MiAlbum — a system for home photo management using the semi-automatic image annotation approach," *ACM MULTIMEDIA 2000 — 8th ACM Int. Multimedia Conf.* (Los Angeles, California, October 30–November 3, 2000).
10. E. Baker, "The mug-shot search problem — a study of the eigenface metric, search strategies, and interfaces in a system for searching facial image data," PhD Thesis (Division of Engineering and Applied Sciences, Harvard University, January 1999).
11. Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool in interactive content-based image retrieval," In *IEEE Trans. Circuits Syst. Video Technol., Special Issue on Segmentation, Description, and Retrieval of Video Content* **8**(5), 644 (1998).
12. Z. M. Hafeed and M. D. Levine, "Face recognition using the discrete cosine transform," *Int. J. Computer Vision* **43**(3), 167 (2001).
13. L. Zhang, F. Lin, and B. Zhang, "A CBIR method based on color-spatial feature," *IEEE Region 10th Ann. Int. Conf. 1999 (TENCON'99)*, (Cheju, Korea, 1999) p. 166.
14. M. Stricker and M. Orengo, "Similarity of color images," *SPIE Proc.* **2420** (1995).
15. J. Huang, S. R. Kumar, M. Mitra, W. J. Zhu, and R. Zabih, "Image indexing using color correlograms," In *IEEE Conf. Computer Vision and Pattern Recognition* **762** (1997).
16. T. Chang and C.-C. J. Kuo, "Texture analysis and classification with tree-structured wavelet transform," *IEEE Trans. Image Processing* **2**(4), 429 (1993).
17. J. C. Platt, M. Czerwinski, and B. Field, "PhotoTOC: automatic clustering for browsing personal photographs," *Microsoft Res. Tech. Rep.* MSR-TR-2002-17 (2002).
18. Z. Bian and X. Zhang, *Pattern Recognition*, Second Edition (Tsinghua Press, Beijing, 2000).



Longbin Chen received his BS degrees in Computer Sciences from Special Class and Gifted Young (SCGY) department of University of Science and Technology of China (USTC) in 2000. He is currently a Master candidate student in the Pattern Recognition and Intelligent System at the Institute of Automation, Chinese Academy of Sciences. His research interests include machine learning, image processing, content-based image retrieval and multimedia system applications.



Lei Zhang received his BS and MS degrees in Computer Science from Tsinghua University in 1993 and 1995 respectively. After two years of working in the industry, he returned to the Tsinghua University and received his PhD degree in Computer Science in 2001. Then he joined Media Computing group in Microsoft Research Asia as an associate researcher. Dr. Zhang's research interests include machine learning, content-based image retrieval and classification, image processing and face detection.



Mingjing Li received his BS in electrical engineering from the University of Science and Technology of China in 1989. He received his PhD in Pattern Recognition from the Institute of Automation, Chinese Academy of Sciences in 1995. He joined Microsoft Research Asia in July 1999. His research interests include handwriting recognition, statistical language modeling, search engine, and multimedia information retrieval.



Hong-Jiang Zhang received his BS from the Zhengzhou University, China in 1982, and his PhD from the Technical University of Denmark in 1991, both in electrical engineering. His research interests include video and image analysis and processing, content-based image/video/audio retrieval, media compression and streaming, computer vision and their applications in consumer and enterprise markets. He has published over 120 articles in the above area. He is a senior member of IEEE. He also serves on the editorial boards of five professional journals and also in a dozen committees of various international conferences.