CLUSTERING CONSUMER PHOTOS BASED ON FACE RECOGNITION

Liexian Gu[1], *Tong Zhang*[2], *Xiaoqing Ding*[1]

[1] Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]Hewlett-Packard Laboratories, Palo Alto, California, USA
lxgu@ocrserv.ee.tsinghua.edu.cn; tong.zhang@hp.com; dingxq@mail.tsinghua.edu.cn

ABSTRACT

The ability of finding photos of a particular person through face recognition is a highly desired feature in indexing, searching and browsing consumer photo collections. In this research, based on an advanced face recognition engine we developed in prior work, one two-pass clustering approach is proposed which groups photos of the same person in a fully automatic way. Firstly, a similarity matrix for all detected faces is computed, with which a semi-supervised clustering is done. Next, larger clusters are selected and modeled as people frequently appearing in the image collection. Then, smaller clusters are recognized against these dominant clusters. Contextual information is used to obtain better results. The approach achieved promising accuracy when tested on an image dataset containing 2316 photos.

1. INTRODUCTION

With people taking more and more digital photos, efficient and effective methods for managing and searching consumer photos such as clustering and indexing photos have become highly demanded. Meanwhile, researches on face detection and face recognition have made significant progresses in recent years – technologies have been developed which can detect and/or recognize human faces with reasonably high accuracy [1][2] in controlled situations. In this paper, we propose a totally automated approach for organizing consumer photos based on the combination of unsupervised face clustering with supervised face model training. That is, with existing face detection and face recognition engines, as well as data clustering methods, a consumer photo collection is automatically clustered into a number of groups, with each group only consisting of photos containing the face of one particular person. Then, the user may visit subsets of the collection like "photos of John", "photos of Mary", or "photos of John & Mary together". Moreover, dominant clusters with frequently appearing people may be deemed as containing major characters of the collection, which enables further understanding of semantic events involved in the image collection. For instance, a picture with two or more major characters standing together can be selected as a one-shot summary of the entire photo set, which is more informative than a randomly picked one.

There has been some prior work on this topic. Systems that are either manual or semi-automatic have been proposed which require the user to label most, if not all of the faces. Such a process can be tedious and time-consuming when the user has to deal with thousands of images [3][4]. One semi-automatic approach was described in [5], in which the user assigns extracted faces to a person, then a face recognition engine forms a model for the person and determines similar faces to the model. Human intervention is limited, but still required here. It may be hard for the user to find a proper photo to start with, especially for a person whose photos are not so many in the collection. Also, this approach may not be suitable for application scenarios such as a web-sharing service or when the photo collection is really large. Das et al. [6] proposed a system that automatically groups images based on face using the nearest neighbor clustering method. However, with unsupervised clustering alone, it is difficult to achieve a high accuracy for face grouping. In [7], Zhao et al. exploited various social context information to combine with face recognition decisions for the sake of person annotation of family photos.

Different from existing approaches, clustering methods and face models are used together in our proposed scheme to achieve a fully automatic and highly accurate process. Also, contextual cues are used to guide the clustering procedure, improving the efficiency and accuracy of the algorithm.

The rest of the paper is organized as follows. In section 2, framework of the proposed scheme is introduced; then details of the method is described in section 3. Experimental results are presented in section 4, followed by conclusion remarks and future work in section 5.

2. FRAMEWORK OF PROPOSED SCHEME

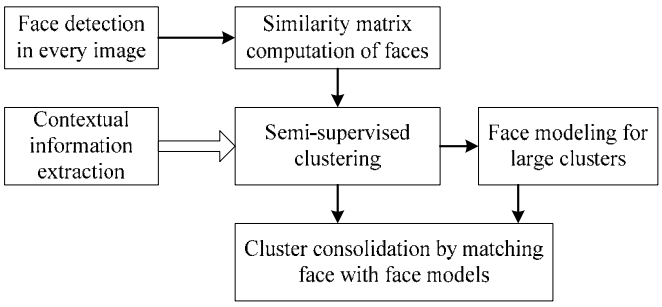

Fig.1: Framework of the proposed scheme.

The framework of the proposed scheme is shown in Fig.1. First, face detection is done on each image in the collection and a skin-color filter is employed to screen out false alarms. Facial features are extracted and similarity values between every pair of faces are computed by a face recognition engine to form an affinity matrix. Based on this, as well as contextual information, a semi-supervised

agglomerative clustering is conducted, and the collection is divided into groups by face. Then, larger clusters (e.g. those containing more than three or four images) are modeled as frequently appearing people. Finally, clusters are consolidated by matching faces with each of these face models. Especially, faces in smaller clusters are merged into larger clusters.

3. DESCRIPTION OF THE STEPS

Details of each step in this approach are described below.

3.1. Facial feature extraction and similarity computation

For each picture in the dataset, face regions are detected using an Ada-Boosting detector with Harr-like features [8]. Due to complicated features of the content in consumer images, false alarms are inevitable. To reduce the number of falsely claimed faces, pixels whose color values fall into a pre-defined skin-color range [9] are counted in each candidate face region. If the ratio of skin-like pixels is beneath a threshold, the face region is discarded as a face-detection false alarm.

In the biometrics research community, the open-set face verification problem has been an active topic, in which a matching score is obtained by comparing two faces to judge whether they are from the same person. Although state-of-the-art technologies have performed well in controlled environments [2], faces in consumer photos present much more difficulties due to variations in imaging conditions such as lighting, pose, expression and so on. For such cases, the facial similarity value between two faces alone cannot be reliable enough for identity decision. Nevertheless, in a family photo collection, there are always several principal people who appear most frequently, e.g. members of the family. Thus, among their faces, some instances can be expected to be more reliably similar than others. Then an aggregating effect is achieved through the clustering approach in our system based on similarity matrix, where each item represents the similarity value between one pair of available faces. More details about the face matching algorithm used in this work can be found in section 4.9. of [2].

3.2. Semi-supervised face clustering

If the number of clusters, K, is known, K-Means is the most prevailing clustering algorithm to partition N objects into K groups [10] for its simplicity and effectiveness. However in our problem, the number of people in the photo collection can not be determined in advance. Also, K-Means is only optimal in describing hyperspherically distributed groups, which is not suitable for the variety of human faces. Therefore, we choose the agglomerative clustering framework. It can work directly on any similarity values and the number of clusters is controlled by a stopping threshold. The algorithm begins with an initial partition where each instance forms a singleton cluster, then among them the most similar two clusters are selected and merged to one cluster; the merging operations repeat until the similarity value between the two merging clusters falls below a stopping threshold or a specified number of clusters have been obtained.

The similarity measures between two clusters are defined as [10]:

$$sim_{\text{comp}}(C_k, C_l) = \min_{f_m \in C_k, f_n \in C_l} (sim(f_m, f_n)) \qquad (1)$$

$$sim_{\text{single}}(C_k, C_l) = \max_{f_m \in C_k, f_n \in C_l} (sim(f_m, f_n)) \qquad (2)$$

where $sim(\cdot, \cdot)$ is the similarity measure, f_m and f_n are faces, C_k and C_l are clusters of faces. The measure (1), *complete-linkage*, leads to compact groups, whereas the measure (2), *single-linkage*, forms elongated clusters with "chaining effect" [6][10]. In our clustering procedure, the complete-linkage algorithm is employed first to ensure each cluster only contains the most similar faces of the same person. Then the single-linkage algorithm is used upon obtained clusters to further consolidate similar clusters until the similarity value between the nearest clusters drops below a threshold.

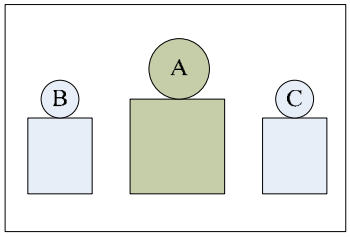

Fig. 2: Illustration of CANNOT-Links. Faces appearing simultaneouly in a picture cannot belong to the same person, namely, A and B, A and C, B and C cannot share the same cluster label.

Photos containing multiple faces can provide additional hints for the clustering. It is obvious that faces appearing simultaneously in a picture must belong to different persons, as shown in Fig. 2. These constraints are so-called CANNOT-Links on instance pairs and can be extracted automatically by analyzing the spatial layout information of detected faces. Clustering with these auxiliary constraints can be done by recently emerged semi-supervised clustering algorithms in the machine learning area.

Semi-supervised face clustering

Given faces $f_i, i = 1, \cdots, N$, stopping threshold θ and

$\text{CANNOT-Links} = \{(f_j, f_k)\}$,

return clusters $C_k = \{f_i \mid if \ \ f_i \in C_k\}, k = 1, \cdots K$

1. Construct initial *N* clusters, with each cluster containing one single face, i.e. $C_k = \{f_k\}, k = 1, \cdots, N$.

2. Find the most similar pair of clusters, say, C_l and C_m. Form instance pairs $P_{lm} = \{(f_s, f_t) \mid f_s \in C_l, f_t \in C_m\}$. If $P_{lm} \bigcap \text{CANNOT-Links} \neq \phi$, these two clusters cannot be merged. Find the next closest cluster pair.

3. If the similarity score falls below θ or there are no more valid cluster pairs, stop; otherwise merge these two clusters and repeat from 2.

Fig.3. Sketch of the semi-supervised face clustering algorithm.

In this work, a semi-supervised agglomerative clustering method based on similar ideas in [11] is employed. The algorithm is sketched in Fig. 3 and it directly operates on the obtained facial similarity matrix in section 3.1. The CANNOT-Links control the procedure of cluster merging together with similarity comparison. In later aggregating steps, these effects will be further propagated. One of the most difficult problems in face clustering is the misclassification of similar-looking individuals [6], e.g. faces of siblings may be falsely mixed into one group. However, in family photo collections it is not occasional that they happen to appear together in one image, thus our semi-supervised clustering can explicitly discriminate them into different clusters.

The stopping threshold controls the final count of clusters. When deployed in real applications, it may be interactively controlled by the user. At this point, it is conservatively preset to ensure that each cluster only contains faces of the same person. Therefore, it is possible that one person has multiple clusters. These separate clusters may be merged in the next step.

3.3. Face modeling and cluster consolidation

After the semi-supervised clustering, faces are grouped into a number of clusters. For typical family photo collections, the obtained clusters often exhibit strongly imbalanced distribution. That is, there are several dominant clusters which correspond to frequently appearing people. And there are yet many small clusters or singletons, which may be strangers, face detection false alarms or faces belonging to the salient persons but not included into the dominant clusters due to rather low facial similarity scores. Each large enough cluster, for instance, those containing more than three or four faces, and/or those in the top 10% of the clusters, is modeled as a pattern class. Then the remaining small clusters are matched with these patterns in a supervised classification manner. If the recognition succeeds with one model, the small cluster is merged into the corresponding larger cluster. In addition to cluster consolidation, such an architecture also facilitates appending new images into an existing photo collection, where new faces are recognized against present people.

For this purpose, The kNN classifier [12] is employed for its simplicity and effectiveness. Each pattern is represented by all of its member faces. The majority class in the k nearest neighbors of the small cluster/singleton is the desired recognition result. Again a threshold is needed here for the case that it does not belong to any of the modeled patterns.

4. EXPERIMENTAL RESULTS

The proposed method was tested on a typical consumer image collection consisting of 2316 photos. It contains images taken in various family events such as vacations, weddings, graduation ceremony and birthday party. The patterns of people appearance and scenes are highly complex. At first, 2994 faces were found by the face detector. We then manually inspected these detected faces and labeled them to create the ground truth. Among detected faces, there are 1296 false alarms. By applying the skin-color filter, 804 false alarms were claimed; among them 67 are actually true faces. Therefore, the false-alarm reduction effort achieves 92% precision and 57% recall. It is reasonable since we want to retain as many true faces as possible.

Clusters containing more than three member faces are defined as significant clusters in this work. With the 2190 detected faces, 2553 CANNOT-Links constraints are extracted. The obtained clusters by pass-1(i.e. semi-supervised clustering) and pass-2(i.e. face modeling and cluster consolidation) are listed in Table 1.

Table 1. The distribution of obtained clusters.

	Pass-1	Pass-2
Number of clusters	1568	1463
Significant clusters	38	38
Faces in significant clusters	488	605

To objectively measure the clustering performance, we select 10 most frequently appearing people in the image collection as the ground truth and check whether they are in the obtained significant clusters. It is a retrieval-style evaluation and the average recall and precision rates are computed with equation 3:

$$Recall = \frac{N_c}{N_t}, \quad Precision = \frac{N_c}{N_r} \tag{3}$$

where N_t is the number of faces in ground truth, N_r is the number of retrieved faces and N_c is the number of correctly retrieved faces. The evaluation results are reported in Table 2.

Table 2. Performance of major character retrieval. The Truth column is the number of faces of the person in ground truth. The #C column is the number of significant clusters corresponding to the person. The r and c columns are the number of faces in the significant clusters and the number of correctly retrieved faces, respectively.

People	Truth	Pass1			Pass2		
		#C	r	c	#C	r	c
P0	212	2	122	122	2	163	157
P1	155	1	88	88	1	120	112
P2	119	2	25	25	2	48	45
P3	65	1	36	36	1	42	41
P4	47	2	28	28	2	29	29
P5	39	2	10	10	2	10	10
P6	37	1	13	13	1	15	15
P7	37	1	17	17	1	21	21
P8	29	1	9	9	1	11	11
P9	23	1	4	4	1	4	4
total	763	14	352	352	14	463	445

Pass-1: $Recall = 0.46, \quad Precision = 1.0$

Pass-2: $Recall = 0.58, \quad Precision = 0.96$

From Table 1 and Table 2, it can be seen that pass-2 of face modeling and cluster consolidation has improved the performance significantly. Many small clusters are merged into large clusters and the average recall rate is improved substantially with only minor degradation in the precision rate.

The high precision rate of the clustering indicates that faces within one cluster in most cases do belong to just one person. Nevertheless faces of the same people may be split into multiple groups. In consumer image management, such a situation is more desirable than to have fewer groups at the expense of mixing wrong faces into clusters, because interactively merging two groups is easier than splitting groups. Another advantage of our clustering results is that the face-detection false alarms are grouped

together into their own three clusters in the 38 significant clusters, instead of being mixed with real faces.

In terms of speed of the process, altogether it took two hours for the procedures of JPEG decoding, face detection, feature extraction and similarity matrix computation of the 2316 images on a 2.4GHz mainstream PC. This would be acceptable since the indexing process may be run in the background. The clustering and consolidation stages are accomplished in less than 2 minutes.

We have developed a photo organizing user interface to demonstrate the usage of face clustering. A snapshot is shown in Fig. 4, where one cluster of the top major character in Table 2 is displayed. Note that many difficult situations such as very dark, small or tilted faces can be handled well by our clustering algorithm. The user can easily navigate, preview and search the whole image collection which is automatically grouped and labeled based on faces. When an image is clicked on, an enlarged and cropped version of it is shown at the left-lower corner, with all detected faces indicated together with the number of faces in corresponding clusters. The user may then choose to switch to other clusters by clicking on one face in the picture.

The example shown in Fig. 5 demonstrates the benefit of using semi-supervised clustering, where faces of the two little children exhibit high resemblance because they are brothers. However since they have appeared in the same picture, the semi-supervised clustering is capable of grouping them into disjoint clusters respectively.

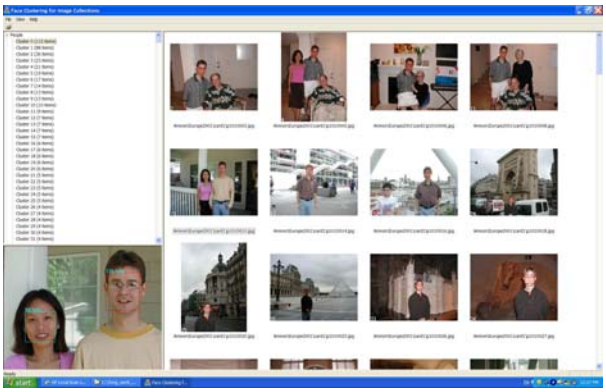

Fig. 4. Demonstration of photo browsing based on face clustering results.

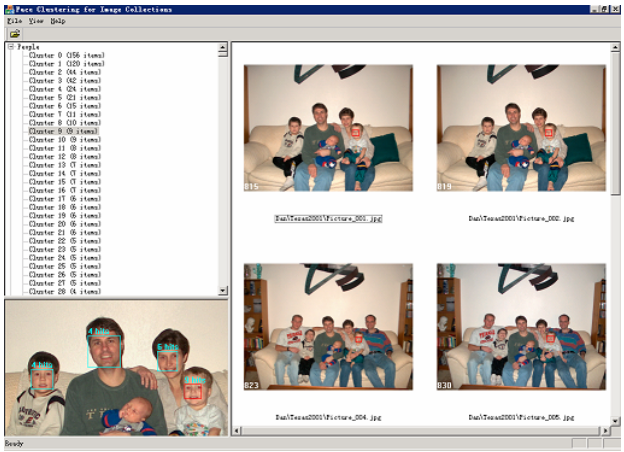

Fig. 5. Sibling children discrimination by semi-supervised clustering.

5. CONCLUSION AND FUTURE WORK

A fully automatic approach for organizing consumer photos based on people is presented in this paper. Faces are detected in every picture and a semi-supervised clustering algorithm is employed on the facial similarity matrix to group faces into clusters while at the same time incorporating spatial constraints. Dominant clusters are modeled as significant people and small clusters are recognized against them to further improve the grouping performance. Promising results have been achieved in consumer photo datasets with high accuracy and fast speed.

In the next step of this work, we will exploit contextual information such as clothes color and hair style to improve the people grouping performance. Metadata contained in digital pictures, for instance, the shooting time and camera imaging parameters, are also helpful to understand semantic contents of photos.

6. REFERENCES

[1] W. Zhao, R. Chellapppa, P. Phillips, *et al.*, "Face recognition: A literature survey," *ACM Computing Surveys*, vol.35, no.4, pp.399-458, Dec. 2003.

[2] K. Messer, J. Kittler, M. Sadeghi, *et al.*, "Face authentication test on the BANCA database," *Proc. of International Conf. on Pattern Recognition*, vol. 4, pp. 523-532, Aug. 2004.

[3] H. Kang, B. Shneiderman, "Visualization methods for personal photo collections: browsing and searching in the PhotoFinder," *Proc. of IEEE Conf. on Multimedia and Expo*, vol.3, pp. 1539-1542, Aug. 2000.

[4] L. Zhang, L. Chen, M. Li, et al., "Automated annotation of human faces in family albums," *Proc. of ACM International Conf. on Multimedia*, pp. 355-358, Berkeley, CA, Nov. 2003.

[5] A. Girgensohn, J. Adcock, and L. Wilcox, "Leveraging face recognition technology to find and organize photos," *Proc. of Workshop on Multimedia Information Retrieval*, ACM Press, pp. 99-106, New York, USA, Oct. 2004.

[6] M. Das and A. Loui, "Automatic face-based image grouping for albuming," *Proc. of IEEE International Conf. on Systems, Man and Cybernetics*, vol. 4, pp. 3726-3731, Oct. 2003.

[7] M. Zhao, Y. W. Teo, S. Liu, et al., "Automatic Person Annotation of Family Photo Album," *Proc. of 5th International Conf. on Image and Video Retrieval*, pp.163-172, AZ, USA, July 2006.

[8] Y. Ma, X. Ding, "Real-time rotation invariant face detection based on cost-sensitive AdaBoost", *Proc. of International Conf. on Image Processing*, vol.3, pp.921-924, Barcelona, Spain, 2003.

[9] P. Kakumanu, S. Makrogiannis and N. Bourbakis, "A survey of skin-color modeling and detection methods", *Pattern Recognition*, vol. 40, no.3, pp.1106-1122, March 2007.

[10] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice Hall, 1988.

[11] I. Davidson and S. S. Ravi, "Agglomerative Hierarchical Clustering with Constraints: Theoretical and Empirical Results," *Proc. of 9th European Conf. on Principles and Practice of Knowledge Discovery in Databases*, Porto, Portugal, Oct. 2005.

[12]. R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd Ed, John Wiley & Sons, 2001.