USING CONTEXT INFORMATION AND LOCAL FEATURE POINTS IN FACE CLUSTERING FOR CONSUMER PHOTOS

Wei-Ta Chu and Ya-Lin Lee

Dept. of CSIE
National Chung Cheng University
Chiayi, Taiwan
wtchu@cs.ccu.edu.tw, lylin96m@cs.ccu.edu.tw

Jen-Yu Yu

Information and Comm. Research Labs
Industrial Technology Research Institute
Hsinchu, Taiwan
KevinYu@itri.org.tw

ABSTRACT

We introduce local feature points to achieve face clustering for consumer photos. After combining eigenfaces with context information like clothes, we further investigate the usage of local feature points to match face images. The relationships between face images are constructed by feature matching and then described as a graph. Outliers in the results of preliminary clustering are detected and are re-clustered according to matching characteristics. We report complete performance comparison for different datasets and show that the proposed method has superior performance than conventional approaches.

Index Terms— Face clustering, context information, local feature points

1. INTRODUCTION

Managing and browsing massive digital photos have been one of the most common behaviors in the cyberspace, due to the extreme ease of photo taking and sharing. To facilitate efficient access, people may organize photos according to timestamps, topics, or events. Among various perspectives of organization, human faces are one of the most attractive clues for photo organization and browsing.

The goal of this work is to automatically cluster photos that contain the same person. Obviously, if we can perfectly recognize human faces in photos, the clustering task is trivial. However, in the ever-increasing consumer photos taken for daily life, events, or travel, human faces are often taken in different poses and different expressions under different lighting conditions, as shown in Figure 1. Face recognition is much more difficult in consumer photos than in image databases taken in controlled environments.

To tackle with this troublesome problem, some studies have been proposed to exploit not only standard face recognition techniques but also external context information. Zhang et al. [1] extracted three features from the upper part of body, face, and eyes, and proposed a Bayesian framework to describe and predict the identification of each face. Zhao et al. [2] proposed a graphical model to integrate face and clothes information. They also conduct some post-processing to eliminate identification errors. Gallagher and Chen [3] especially paid attention to the property that people belonging to the same group often appear together. They extract facial characteristics and compare performance based on different probabilistic models. Because of measurable commercial values, Google's picasa web album [4] also provides the "name tags" function. Faces that are identified as the same person are grouped together to ease manual annotation.

The major challenges of face clustering in consumer photos are the drastically varied face poses and lighting conditions. Although the prescribed works appeal to face and context information likes clothes color, promising features for this task have not been completely studied. In this work, we propose an approach that further combines local feature points [5], which are widely used in computer vision applications. By checking the spatial correspondence between faces, we can largely improve the accuracy of face clustering.

The rest of this paper is organized as follows. Section 2 provides a system overview. Section 3 describes context information extraction and usage. The idea of local feature points is addressed in Section 4. Performance evaluation and comparison are given in Section 5, and conclusion is stated in Section 6.

Figure 1. The same person in different poses, expression, and lighting conditions.

2. SYSTEM OVERVIEW

Figure 2 shows the proposed framework. Given a set of photos, this system automatically detects faces, clusters faces according to the proposed features and methods, and fix clustering errors by the post-processing module. We briefly describe the components in the following.

2.1. Face detection and eye detection

We construct a face detector based on the features extracted by the Haar wavelet transform. Although the face detection

module provides acceptable results, some objects that are similar to face appearance would be mis-detected as faces. Thus we further apply the eye detection process to the detected faces, and filter out the ones that have no eye.

2.2. Eigenface

Following the standard eigenface approach, we project each detected face region into the pre-trained eigenspace and present each face region by the space's basis, i.e., eigenface. The coefficients corresponding to this basis is concatenated as a feature vector to describe a face region.

2.3. K-means clustering

Based on eigenface coefficients, we apply the k-means algorithm to cluster face regions. Conceptually, the faces that are grouped into the same cluster should correspond to the same person. However, this method works badly in consumer photos, where side-view faces or drastic lighting conditions significantly harm the performance. To improve the accuracy of face clustering, most works appeal to other context information, such as clothes color [1][2] and group's prior probability [3].

2.4. Re-clustering based on clothes information

People in the photos that were taken within a short duration are assumed to be in the same dresses. Clothes information, therefore, provides important clues to correlate faces. For the faces that are grouped in the same cluster at the previous stage, their corresponding clothes are further examined based on clothes' color histograms. If a face's corresponding clothes is significantly different from the average clothes information of the same face cluster, this face is specially selected to be re-assigned.

2.5. Re-clustering based on SIFT

Accurately finding clothes regions is still a challenging work. Moreover, person's clothes are often occluded by other persons in group photo. Therefore, we further extract SIFT (scale-invariant feature transform) features [5] to represent local feature points, and use them to compare detected faces. We demonstrate that this feature works well in matching faces and effectively enhance face clustering.

2.6. Postprocessing

It is obviously wrong if two or more faces in the same photo are grouped in the same face cluster. They are re-examined and surely assigned to different clusters in the post-processing module.

Due to the space limitation, we would like to skip face detection, eye detection, and the standard eigenface module, which can be easily found in literature. In Sections 3 and 4, we describe our major contributions – re-clustering based on clothes information and SIFT features.

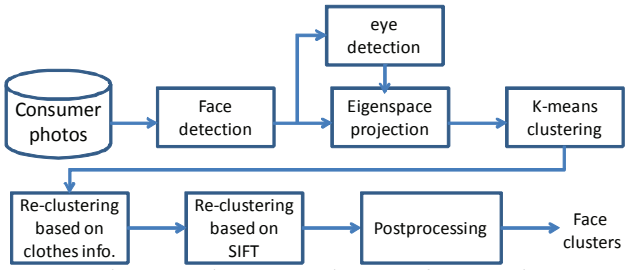

Figure 2. The proposed system framework.

3. RE-CLUSTERING BASED ON CLOTHES INFORMATION

After k-means clustering based on eigenface coefficients, we obtain K face clusters, denoted by $C = \{C_1, C_2, ..., C_K\}$. For each face cluster $C_i, 1 \le i \le K$, we select face images whose corresponding clothes regions are significantly different from the average characteristics of the same group and re-cluster them. Because accurately segmenting each face image's corresponding clothes region is still an on-going research, we simply set the position and area of a rectangle region according to the corresponding face image.

For a the ith face cluster $C_i = \{f_1, f_2, ..., f_n\}$, the corresponding clothes regions $T = \{t_1, t_2, ..., t_n\}$ are extracted. A face image $f_j, 1 \le j \le n$, will be selected to be re-clustered if

$$d(t_j, t_{avg}) > 2\sigma, \quad (1)$$

where t_{avg} is the average clothes of the set T, $d(t_j, t_{avg})$ is the Euclidean distance of RGB color histograms between t_j and t_{avg}, and σ is the standard deviation of the distances between clothes in T.

The selected clothes regions are viewed as the "outliers" of the set T. An outlier t_j, which corresponds to the face image f_j, is re-assigned to the k^*-th face cluster C_{k^*} if

$$k^* = \arg\min_{1 \le k \le K} d(t_j, t_{avg}^k), \quad (2)$$

where t_{avg}^k is the average clothes of the face cluster C_k.

4. RE-CLUSTERING BASED ON SIFT

Although eigenface and clothes information provides helpful clues for clustering face images, luminance changes and pose variation often degrade the robustness of clustering. Recently, local feature points are widely applied to match images in different viewpoints, scales, and lighting conditions. Therefore, we extract local feature points from face regions, describe them by SIFT description, and introduce them to the task of face clustering.

Figure 3(a) shows the SIFT-based matching situations between the same persons. We can easily see that the matched points distribute around the eyebrows, eyes, noses, and mouths. This situation actually matches the cognition theory [6], which describes how we understand and distinguish people. We can also note that feature points can be matched even two face images are in different lighting conditions or in different poses. On the other hand, in Figure

3(b), the number of matched points between different persons is fewer, and the matched points don't uniformly distribute around most of the important facial organs.

One thing worth noting is that the people in Figure 3 actually are members of the same family. Although some of them (father and son, or brothers) look similar, SIFT features still provide good clues to distinguish different persons.

For each face cluster C_k, we describe the relationships between face images as an undirected graph $G_k = (V, E)$. The vertices $V = \{f_1, f_2, ..., f_n\}$ are the face images in this cluster, and the edge set E includes edges that connect "similar" faces. The edge e_{ij} in E connects the *i*th face image f_i and the *j*th face image f_j if the SIFT matched points between them distribute at least more than two facial organs, such as eyes and noses.

If all face images in C_k really belong to the same person, and the SIFT features match perfectly between images, then the graph G_k would be a complete graph of size n. However, the clustering results are not perfect after examining eigenface and clothes information. Moreover, SIFT-based matching is limited in many situations. Therefore, we instead detect cycles in the graph.

Edges connect similar face images, and according to transitivity, face images in the same "cycle set" are similar. A cycle set is defined as a set that contains connected cycles. For example, if nodes no. 1, 2, 3 form a cycle, and nodes no. 1, 4, 5 form another cycle, these two cycles would be put into the same cycle set. Note that it doesn't necessarily mean that nodes no. 1, 2, 4 form a cycle, for example. In this work, we detect members in the largest cycle set, and leave the remainder as outliers of this face cluster.

Face images f_i and f_j are neighbors if there is an edge e_{ij} linking them. We develop an algorithm to find the members of the largest cycle set, as shown in Figure 4.

The selected outlier face images are re-clustered based on SIFT features matching. For simplicity, we still use the notation $\{C_1, C_2, ..., C_K\}$ to denote the face clusters in which outliers have been picked out. One outlier face image f_o is re-assigned to the cluster C_{k^*} if

$$k^* = \arg\max_{k=1,2,...,K} \langle f_o, C_k \rangle, \quad (3)$$

$$\langle f_o, C_k \rangle = \frac{1}{|C_k|} \sum_i^{|C_k|} \langle f_o, f_i \rangle, \quad (4)$$

where $|C_k|$ denotes the number of face images in C_k, and $\langle f_o, f_i \rangle$ denotes the number of SIFT matched points between the face image f_o and the *i*th face image in C_k. The value $\langle f_o, C_k \rangle$ means the average number of matched points between f_o and face images in C_k.

After SIFT-based re-clustering, we apply a post-process to generate the final results. It's obviously that the same person would not appear in the same photo more than twice. Therefore, we re-cluster the face images that are clustered together but are from the same photo, according to SIFT matched points.

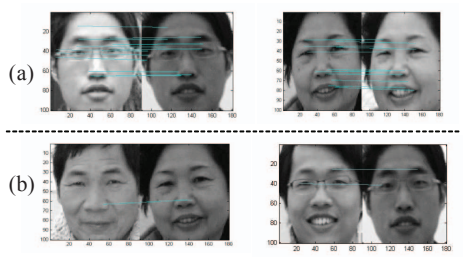

Figure 3. SIFT matches between (a) the same persons' face images. (b) different persons' face images.

Input: The face images in a cluster
Output: The members of the largest cycle set in this cluster

Step 1: For the *i*th face image f_i, find its neighbors $M_i = \{f_s, f_{s+1}, ..., f_{s+m}\}$.
Step 2: Check the relationship between any two images in M_i. If the images f_{s+j} and f_{s+k} are neighbors, $0 \le j \le m,\ 0 \le k \le m,\ j \ne k$, then
(1) if f_i, f_{s+j}, and f_{s+k} never appear in another cycle set, create a set $\Omega_i = \{f_i\} \cup \{f_{s+j}\} \cup \{f_{s+k}\}$
(2) if one of f_i, f_{s+j}, and f_{s+k} once appears in another cycle set, say Ω_p, then $\Omega_p = \Omega_p \cup \{f_i\} \cup \{f_{s+j}\} \cup \{f_{s+k}\}$
Step 3: Go back to Step 1 until all face images in this cluster have been examined.
Step 4: Find the largest cycle set that contains the most members.

Figure 4. The algorithm for detecting members of the largest cycle set.

5. EXPERIMENTS

To evaluate the performance of the proposed approach, we collect consumer photos that record travel or daily life on the internet. We also use a subset of photos from an open photo collection [7]. There are 16 different datasets, and the total number of photo is 1199. The number of persons in a dataset range from two to seven.

We compare the performance of three approaches: only based on eigenface, combining eigenface and clothes-based re-clustering, and combining eigenface, clothes-based re-clustering, and SIFT-based re-clustering. Figures 5 and 6 respectively show precision and recall performance in different datasets. From Figure 5, we can see that combining SIFT-based re-clustering can effectively improve performance, while clothes-based re-clustering doesn't significantly improve as reported in previous literature. It is expectable that improvement varies for different datasets. Overall, the average precisions of three approaches are 0.53, 0.54, and 0.62. Similarly, the average recalls of three approaches are 0.34, 0.34, and 0.41.

One of the reasons about lower recall is the failure of face detection caused by extreme poses. Figure 7 shows

some examples that cause failed face detection. Such extreme poses frequently occur in consumer photos.

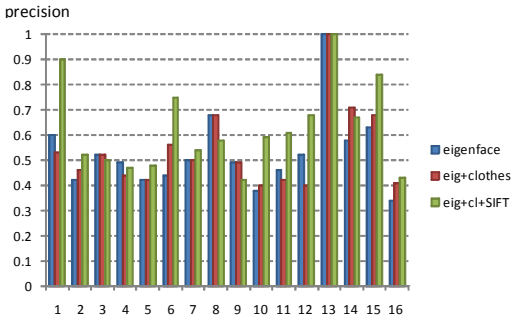

Figure 5. Precision of face clustering in different datasets.

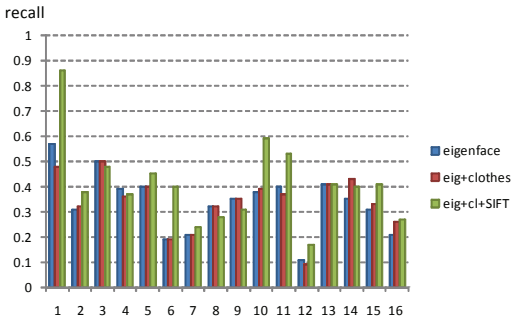

Figure 6. Recall of face clustering in different datasets.

Figure 7. Some examples that cause failed face detection.

In the task of face clustering, precision plays a more important role than recall. Users are more likely to have more small clusters, in each faces really belong to the same person, than fewer large clusters, in each faces may belong to different persons. Recently, Google's Picasa web album [4] provides a function of automatic face clustering to facilitate face annotation. It tends to provide many small face clusters to users, and achieves high precision performance. However, if a dataset actually contains k (e.g., four) different persons, Picasa often provides a much larger number of clusters than k, e.g. twenty.

In our experiments, the actual number of face clusters is decided by users and the proposed approach clusters faces accordingly. To verify whether the precision values get higher when we make more clusters, as Picasa does, we evaluate precision values based on grouping face images into k, k+1, k+2, and k+3 clusters. The value k corresponding to each dataset is the actual number of persons in it. Figure 8 shows the results in different clustering situations. Overall, the average precision values are 0.62, 0.62, 0.67, and 0.68. This result confirms the tendency of more clusters, higher precision, and provides us guidelines of designing a face-centric photo browsing system.

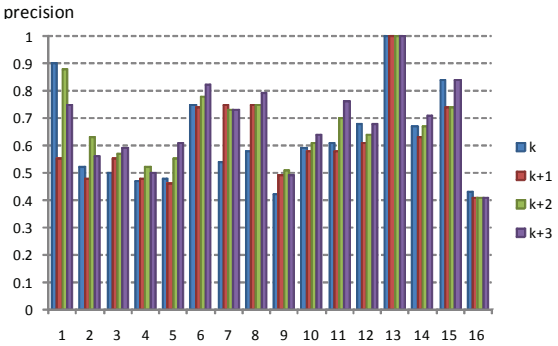

Figure 8. Precision in different clustering situations.

6. CONCLUSION

Face clustering for consumer photos is a challenging work due to the data with extremely different characteristics captured from uncontrolled environments. In this paper, we introduce local feature points in face clustering. With the help of number of SIFT matched points and their spatial distribution, we can match faces in different poses or lighting conditions. The experimental results demonstrate the effectiveness of the proposed method, and further confirm the design guidelines of a face clustering system. In the future, we would investigate the fusion scheme of different modules and compare different face clustering methods based on the same dataset.

7. ACKNOWLEDGMENTS

This work was partially supported by the National Science Council of the Republic of China under grants NSC 97-2221-E-194-050.

8. REFERENCES

[1] L. Zhang, L. Chen, M. Li, and H.J. Zhang, "Automated annotation of human faces in family albums," Proceedings of ACM Multimedia, 355-358, 2003.

[2] M. Zhao, Y.W. Teo, S. Liu, T.-S. Chua, and J. Ramesh, "Automatic person annotation of family photo album," Proceedings of International Conference on Image and Video Retrieval, 163-172, 2006.

[3] A.C. Gallagher and T. Chen, "Using group prior to identify people in consumer images," Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 1-8, 2007.

[4] Picasa web albums, http://picasaweb.google.com

[5] D. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60(2), 91-110, 2004.

[6] F. Nahm, A. Perret, D. Amaral, and T. Albright, "How do monkeys look at faces?" Journal of Cognitive Neuroscience, 9, 611–623, 1997.

[7] The Gallagher Collection Person Dataset, http://amp.ece.cmu.edu/people/andy/GallagherDataset.html