A Face Annotation Framework with Partial Clustering and Interactive Labeling

Yuandong Tian *
Shanghai Jiaotong University, China
tydsh@sjtu.edu.cn

Wei Liu
The Chinese University of Hong Kong
wliu5@ie.cuhk.edu.hk

Rong Xiao Fang Wen Xiaoou Tang
Microsoft Research Asia, Beijing, China
{rxiao, fangwen, xitang}@microsoft.com

Abstract

Face annotation technology is important for a photo management system. In this paper, we propose a novel interactive face annotation framework combining unsupervised and interactive learning. There are two main contributions in our framework. In the unsupervised stage, a partial clustering algorithm is proposed to find the most evident clusters instead of grouping all instances into clusters, which leads to a good initial labeling for later user interaction. In the interactive stage, an efficient labeling procedure based on minimization of both global system uncertainty and estimated number of user operations is proposed to reduce user interaction as much as possible. Experimental results show that the proposed annotation framework can significantly reduce the face annotation workload and is superior to existing solutions in the literature.

1. Introduction

Digital photo albums are growing explosively in both number and size due to the rapid popularization of digital cameras and mobile phone cameras in the last decade. These large collections require the annotation of some semantic information to facilitate browsing, manipulation and sharing of photos. In a typical family photo, besides the information of *when* and *where*, *who is in the photo* is essential. Therefore, face annotation is becoming an indispensable part of the management of photos depicting people.

In most commercial systems, such as Adobe Photoshop Elements, iView Media Pro, and ACDSee Photo Manager, face annotation is mainly based on elaborate user-driven UI designs. Although some efforts have been made to simplify photo labeling with a drag-and-drop interface, automatic face recognition technology has not been used in any of these systems, except for the face detection technique. Intensive operations are required to label/group faces. Labeling each photo by hand remains a tedious task. Hence, it is important to develop an automatic/semi-automatic method to enable rapid face annotation.

A straightforward idea for automatic/semi-automatic face annotation is to integrate face recognition algorithms which have been well studied in the last decade [2, 3, 9, 13, 15, 16, 17]. Girgensohn *et al*. used face recognition technology to sort faces by their similarity to a chosen face or trained face model, reducing user workload to searching faces that belong to the same person [7]. However, despite progress made in recent years, face recognition continues to be a challenging topic in computer vision research. Most algorithms perform well under a controlled environment, while in the scenario of family photo management, the performance of face recognition algorithms becomes unacceptable due to difficult lighting/illumination conditions and large head pose variations[21].

In [20], social context information and body information is used to do automatic person annotation. Davis *et al*. also used contextual metadata to help face recognition [6]. Although face recognition performance was significantly improved after integrating contextual information, the recognition rate is still far from the requirement of an automatic face annotation system.

In [18], Lei *et al*. proposed a semi-automatic approach to do face annotation. In their method, they proposed a Bayesian framework to automatically calculate a candidate list of names for the face to be annotated. The major disadvantage of this work is that it requires users to annotate photos one by one. In their later approach [19], they proposed a new approach to do name propagation while annotating multiple photos. In their scenario, they assume that multiple selected photos contain the same person. The name propagation problem is to find the optimal solution to annotate

*This work was done when Yuandong Tian and Wei Liu were visiting students at the Microsoft Research Asia.

1-4244-1180-7/07/$25.00 ©2007 IEEE

this person in these photos automatically. However, their scenario is quite limited, since people still need to browse the whole album to select the photos to be annotated.

A common problem of above semi-automatic frameworks is that users have to manually select photos one by one. Suh *et al*. proposed a framework to allow cluster annotation [12]. The faces were clustered according to time and torso information firstly, and then the user can label a cluster in one operation. However, in this work, once the clusters were determined, the remaining work becomes all manual. All the errors in the clustering need to corrected one by one by the user. In addition, the clustering performance using only time and torso information is very limited. Therefore, the remaining manual work is still intensive.

Recently, Riya [1] developed an iterative framework for face annotation. In every iteration, the user was asked to manually label some faces, then the system used these labeled information to recognize faces that belong to the same person, and proposed them for user confirmation. Few technical details are available about Riya's algorithm, but from experiments we can see that it still requires a lot of manual labeling to obtain final annotation results.

Inspired by above approaches, we propose a novel interactive framework to further reduce the face annotation workload. This is critical for performance improvement of interactive photo annotation system [5]. Due to current status in face recognition, it is not practical to expect a framework that eliminates all user interaction. Our goal is to:

1) achieve relatively high performance without user interaction;

2) when user interaction is included, reduce it to an acceptable level.

We implement these two criteria in two stages, an unsupervised stage and an interactive stage, as illustrated in Figure 1.

There are two main contributions of our framework. The first contribution is the formulation of a partial clustering algorithm. This algorithm aims at reduction of user labor rather than overall clustering performance. It is designed to deliberately bias toward evident clusters, so that a user can quickly label them to offer the framework a large amount of labeled information with very little effort.

The second contribution is the interactive labeling procedure in the interactive stage. In this procedure, both global system uncertainty and estimated number of user operations are modeled via entropy notation. In each iteration step, a particular group of unlabeled faces that most likely belong to one person and are most informative to decrease global entropy is pop up for the user to label. Therefore, the user's workload in the interactive stage will be reduced as much as possible.

2. Partial Clustering for Face Annotation

As discussed above, the current situation in face recognition limits the maximum performance that an unsupervised algorithm can achieve. So we are not expecting overall good performance for a clustering algorithm, which is, however, the ultimate goal for most machine learning methods, but aim at finding initial good clusters for a user to label easily.

To achieve this goal, we try to properly bias the cluster results so that only "evident" clusters are kept, while other faces, that are not grouped tightly enough, remain in the litter bin. These evident clusters usually contain only one identity, hence a user can do batch labeling with only one click. Then with this easily obtained labeled information at hand, an interactive labeling procedure follows to reduce user interaction.

In the following, we first discuss how features are extracted and combined to form similarity. Then describe the detail of the partial clustering algorithm.

2.1. Spectral Embedding Using Face similarity measure

In a photo album, a set of faces $X = \{x_i\}, i = 1 \dots N$ is extracted for each individual. For each face $x \in X$, $f(x)$ is its representation in a facial feature space. We also extract contextual features for each face, including texture features $c(x)$ which are extracted on clothes areas, and time feature $t(x)$ which is the time when the photo was taken.

For any two faces x_i and x_j, we define the following distances: $d^F_{i,j} \equiv d(f(x_i), f(x_j))$ is the distance in facial feature space, $d^C_{i,j} \equiv d(c(x_i), c(x_j))$ is the distance of cloth feature, $d^T_{i,j} \equiv d(t(x_i), t(x_j))$ is the time distance.

Using the Bayesian rule, face similarity can be formulated as:

$$P(\Omega_I|d^F, d^C, d^T) = \frac{P(d^F, d^C|\Omega_I, d^T)P(\Omega_I|d^T)}{P(d^F, d^C|d^T)}, \quad (1)$$

where Ω_I indicates patches x_i and x_j are from the photos of the same individual.

Using the assumption of a *Time Prior* that a person of the same identity tends to wear the same clothes during a short period of time, the dependence between d^F and d^C only comes from the knowledge of the time prior. Therefore we have

$$P(d^F, d^C|\Omega_I, d^T) = P(d^F|\Omega_I, d^T)P(d^C|\Omega_I, d^T). \quad (2)$$

Given Ω_I, d^F is independent of d^T, and Ω_I is independent of d^T, we have

$$P(d^F|\Omega_I, d^T) = P(d^F|\Omega_I), \quad (3)$$

$$P(\Omega_I|d^T) = P(\Omega_I). \quad (4)$$

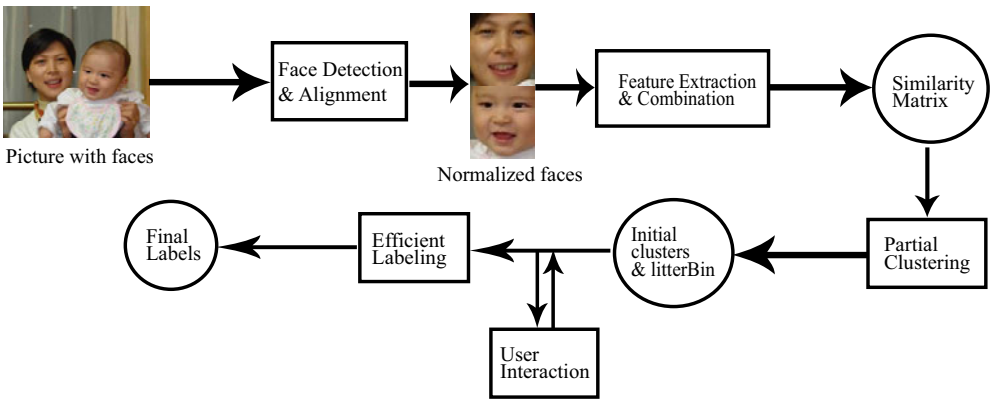

Figure 1. The workflow for face annotation framework

Using Eq. 2 - Eq. 4, Eq. 1 can be rewritten as

$$P(\Omega_I|d^F, d^C, d^T) = \frac{P(d^F|\Omega_I)P(d^C|\Omega_I, d^T)P(\Omega_I)}{P(d^F|d^T)P(d^C|d^T)}. \tag{5}$$

We use a similarity matrix A to store all pair-wise similarities, with $a_{ij} = P(\Omega_I|d^F_{ij}, d^C_{ij}, d^T_{ij})$.

The probabilities $P(d^F|\Omega_I)$, $P(d^C|\Omega_I, d^T)$, $P(\Omega_I)$, $P(d^F|d^T)$, and $P(d^C|d^T)$ can be estimated in a training set using a similar method described in [18].

In addition to the time prior, we also propose another prior, called *Cannot Link Prior* to further improve the discriminant power of face similarity. The *Cannot Link Prior* comes from the fact that two faces appearing in the same photo belong to different people. This prior is simply modeled as a hard constraint on face similarity.

2.2. Partial Clustering algorithm

Once pair-wise similarity is defined, many methods can be used for unsupervised clustering. Spectral clustering[11], is one of the algorithms that has been proven to be effective and stable.

Spectral clustering procedure can be decomposed into two parts, spectral embedding and clustering. Spectral embedding finds representations $\{\phi_i\}_{i=1\ldots N}$ for faces $\{x_i\}_{i=1\ldots N}$ in a metric-equipped compact manifold C for graph-structured data, where data are much more easily clustered. This compact manifold C is actually the surface of a d-dimensional unit hyper-sphere. Then classic K-means is used to cluster them in C.

However, for our specific problem, due to difficulties encountered in face recognition, pair-wise similarity does not work very well even if contextual information is added. In this situation, after spectral embedding, many faces are mapped into messy data points and simple K-Means only produces very noisy results.

In this section, we proposed a partial clustering algorithm to address this problem. Different from traditional clustering algorithms, the partial clustering algorithm will not group all samples into clusters. The basic assumption made in this algorithm is that the noisy samples which are difficult for clustering will be distributed uniformly after the spectral embedding.

Therefore, as shown in Eq. 6, we modified the traditional Gaussian Mixture Model by adding a uniform background noise distribution. We also propose to use an exponential prior to control the variances of component Gaussian distributions:

$$\begin{array}{rcc} \sigma_j^2 & \sim & Exp(\theta) \quad j = 1\ldots K \\ Y & \sim & P(Y=j) = \alpha_j \quad j = 0\ldots K \\ \Phi & \sim & P(\phi|y=j,\sigma_j) = N(\phi;\mu_j,\sigma_j^2) \quad j = 1\ldots K \\ & \sim & P(\phi|y=0) = p_b, \end{array} \tag{6}$$

where α_j is the cluster prior, μ_j is the mean of the cluster, σ_j is the variance of the cluster, p_b is the density of the background distribution, y_i is the label of face x_i, and ϕ_i is the spectral representation of face x_i.

The uniform probability density p_b is thus evaluated to be the inverse of a finite "area" $S_d(1)$ of C, with

$$S_d(1) = \frac{2\pi^{d/2}}{\Gamma(d/2)} \tag{7}$$

where $\Gamma(\cdot)$ is the gamma function $\Gamma(s) = \int_0^\infty x^{s-1}e^{-x}\mathrm{d}x$.

Then our goal is to find the MAP estimation of the following posterior likelihood function:

$$P(\phi, \sigma|\mu, \alpha) = \sum_y P(\phi, y, \sigma|\mu, \alpha), \tag{8}$$

which can be solved by the Expectation-Maximization (EM) algorithm.

Compared with the traditional K-Means algorithm, the partial clustering algorithm only focuses on finding "evident" clusters which contain samples with high intra-

cluster similarity. Samples in a disperse distribution will be grouped into a litter-bin cluster.

3. Interactive Labeling for Face Annotation

The partial clustering algorithm automatically groups similar faces into several evident clusters, and groups dissimilar faces into a background cluster, called the litter-bin. After the partial clustering stage, we use an "Initial labeling" procedure to annotate these evident clusters. Since faces in an evident cluster most likely belong to a single individual, user annotation interactions on these clusters can be significantly reduced. However, the workload of face annotation in the litter-bin is still huge.

In this section, we propose a parameter-free, iterative labeling procedure to address this problem. In each step, the system uses the information from the labeled faces to automatically infer an optimal subset of unlabeled faces for user annotation. This annotation step will be iteratively used until all faces are labeled. Using this strategy, the overall user interactions can be reduced by finding an optimal subset of unlabeled faces in each annotation step.

Suppose there are K labeled groups of identities $\mathcal{G} = \{G_1, \ldots, G_K\}$, with $G_j = \{x_i | y_i = j\}$ for $j = 1 \ldots K$, and an unlabeled face set G_0, which define the beginning state $s_0 = \{G_0, \mathcal{G}\}$. Each time we choose to label a subset $Q \subseteq G_0$, and then go to the next state $s' = \{G_0 \backslash Q, \mathcal{G} + Q\}$ with

$$\mathcal{G} + Q \equiv \bigcup_j \left(G_j + \bigcup_{x_k \in Q, y_k = j} \{x_k\} \right). \tag{9}$$

The transition weight between two states is defined as the information efficiency, the ratio r of expected *information gain* to estimated user operations in labeling Q:

$$r \equiv \frac{\mathbf{E}_{\mathcal{G}} \left[\mathrm{Gain}(Q; \mathcal{G}) \right]}{\mathrm{Operations}(Q)}. \tag{10}$$

We thus search for a path $P \equiv \{Q_1, \ldots, Q_m\}$ from s_0 to the common final state $s_F = \{\emptyset, \mathcal{G}_F\}$ that maximizes the sum of weights over transitions as the following:

$$\max_P \sum_{k=1}^{m} r_k, \tag{11}$$

and r_k is defined as:

$$r_k = \frac{\mathbf{E}_{\mathcal{G}_k} \left[\mathrm{Gain}(Q_k; \mathcal{G}_k) \right]}{\mathrm{Operations}(Q_k)}, \tag{12}$$

with $\mathcal{G}_k \equiv \mathcal{G} + \bigcup_{j=1}^{k-1} Q_j$.

To solve this problem, one has to enumerate all the possibilities to find the optimal solution, which results in an NP-hard problem. So we resort to a greedy approach. In each iteration, we find an optimal set of unlabeled faces $Q \subseteq G_0$ that maximizes the ratio r,

$$Q = \arg\max_Q \frac{\mathbf{E}_{\mathcal{G}} \left[\mathrm{Gain}(Q; \mathcal{G}) \right]}{\mathrm{Operations}(Q)}. \tag{13}$$

In the following subsections, we model $\mathrm{Gain}(Q; \mathcal{G})$ as the decrement of global entropy of the system conditioned on $\mathcal{G}$, and $\mathrm{Operations}(Q)$ as *subset-saliency entropy* (SSE), which represents estimated number of user operations.

3.1. Information Gain

For $x_i \in G_0$, we assume that its label y_i has a probability distribution conditioned on $\mathcal{G}$:

$$P(y_i = j | \mathcal{G}) \propto \max_{x_k \in G_j} a_{ik}. \tag{14}$$

a_{ij} is the similarity measure between face i and j. We use the most similar criterion instead of average. Since the face distribution in the feature space is well known on a high dimensional manifold, using the similarity between the nearest-neighbor is more robust than using the average of similarities over all relevant samples.

The total uncertainty of all unlabeled faces in G_0 can be measured by entropy. Assuming that G_0 is an independent random variables set, its global (pseudo-)entropy is simply the addition of each independent part x_i:

$$H(G_0 | \mathcal{G}) = \sum_{x_i \in G_0} H(x_i | \mathcal{G}), \tag{15}$$

with each part $H(x_i | \mathcal{G})$ defined on the probability measure of Eq. 14.

Suppose the subset $Q \subseteq G_0$ is manually labeled, then the *information gain* can be defined as the decrement of $H(G_0 | \mathcal{G})$:

$$\mathrm{Gain}(Q; \mathcal{G}) \equiv -\Delta H(Q | \mathcal{G}) = H(G_0 | \mathcal{G}) - H(G_0 \backslash Q | \mathcal{G} + Q) \tag{16}$$

In general, $\mathrm{Gain}(Q; \mathcal{G})$ is not accessible since the true labels of Q are unknown. But we can instead evaluate the expectation of $\mathrm{Gain}(Q; \mathcal{G})$, conditioned on Eq. 14:

$$\mathbf{E}_{\mathcal{G}}(\mathrm{Gain}(Q; \mathcal{G})) = \sum_{l_Q \in L_Q} \mathrm{Gain}(l_Q; \mathcal{G}) P(l_Q | \mathcal{G}), \tag{17}$$

where l_Q is a label assignment of the set Q, and L_Q is the set of all possible label assignments. By the independence assumption of G_0, we can then actually evaluate it.

3.2. Subset-Saliency Entropy

Given subset $Q \subseteq G_0$, we can estimate the number of user operations via *Subset-Saliency Entropy* $H(Q)$:

$$H(Q) = -\sum_{l_Q \in L_Q} P(l_Q) \log P(l_Q), \tag{18}$$

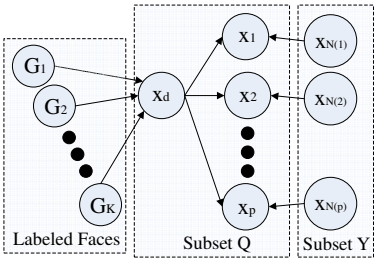

Figure 2. The graphical model for $H(Q)$

with $P(l_Q)$ evaluated by the following equation:

$$P(l_Q) = \sum_{l_{G\backslash Q} \in L_{G\backslash Q}} P(l_Q|l_{G\backslash Q})P(l_{G\backslash Q}), \quad (19)$$

where $G = \{x_i\}_{i=1...N}$ is set of all faces.

This entropy actually models a competition between Q itself and $G\backslash Q$, hence the name. As in Eq. 19, if $P(l_Q|l_{G\backslash Q})$ stays nearly constant when $l_{G\backslash Q}$ changes, then l_Q appears highly correlated and cohesive, which makes $H(Q)$ small. In short, Q are expected to share the same label; if $P(l_Q|l_{G\backslash Q})$ changes rapidly with $l_{G\backslash Q}$, then Q is heavily influenced by the external node from Q, which tends to make Q an independent set. In such a situation, intensive user operations are unavoidable to label Q.

3.3. The Algorithm to Solve $H(Q)$

In general, directly computing $H(Q)$ is NP-hard. Additionally, even optimizing Eq. 13 instead of Eq. 11 is intractable. We again adopt a greedy approach that solves both.

Indeed, we first pick one unlabeled face x_d as the seed of Q, and then do a local search over its neighbors, each time searching for $x_i = \arg\max_{x_i \in G_0\backslash Q} a_{id}$, and put it into Q, until Eq. 13 start to decrease.

The greedy procedure also yields a plausible and efficient way of computing $H(Q)$. Letting $Q = \{x_d, x_1, \ldots, x_p\}$, we assume a graphical model as in Fig. 2. The loopy structure in Eq. 19 is then substantially simplified into tree-structure.

In this simplified model, let subset $Y \subseteq G\backslash Q$ hold competitors. Y can be any subset. Typically we choose two cases, $Y = G_0\backslash Q$ and $Y = \bigcup \mathcal{G}$, which correspond to pure unsupervised and pure supervised versions. Of course, any mixture version is allowed. Here we use an unsupervised version in the experiments.

For each $x_i \in Q\backslash\{x_d\}$, we choose $x_{N(i)}$ from Y via the most similar criterion:

$$x_{N(i)} = \arg\max_{x_k \in Y} a_{ik}, \quad (20)$$

and then define conditional probability in Fig. 2 as the following:

$$P(y_i|y_d, y_{N(i)}) \propto \begin{cases} a_{i,N(i)} & y_i = y_{N(i)} \\ a_{id} & y_i = y_d \\ 0 & otherwise. \end{cases} \quad (21)$$

For $P(y_{N(i)})$, if $x_{N(i)}$ is labeled, then it is a delta function peaked at $y_{N(i)}$, otherwise we assign a uniform probability over K possible labels. This is because in the unsupervised version, we want $H(Q)$ to truly reflect the saliency structure of Q in G_0, without any bias on labeled data. But the supervised version is equally reasonable.

Then for each x_i, by marginalization over $x_N(i)$, we get

$$P(y_i|y_d) = \sum_{y_{N(i)}} P(y_i|y_d, y_{N(i)})P(y_{N(i)}). \quad (22)$$

And $H(Q)$ is thus evaluated as the following:

$$H(Q) = H(y_d) + H(Q\backslash\{y_d\}|y_d) = H(y_d) + \sum_i H(y_i|y_d). \quad (23)$$

In essence, in the extreme case with strong intra-connection and weak interconnection of Q, $H(Q)$ will be exactly $H(y_d) \approx \log K$, which indicates only one operation is needed; whereas in the other extreme case, all y_i are mutual independent no matter whether y_d is given, which results in $H(y_d) + p \log K \approx (p+1)\log K$, and indicates $p+1$ operations is needed. This verifies the effectiveness of approximated $H(Q)$.

4. Experimental Results

We present three comparative experiments here to demonstrate the performance of our framework.

In the unsupervised stage, experimental comparisons are made between the prior-equipped partial clustering algorithm and the classic spectral clustering algorithm.

In the interactive stage, two related works are compared. One is interactive annotation based on (second-time) spectral clustering which labels one by one the newly-generated clusters from unlabeled faces. The cluster with lowest similarity variation will be annotated first. Another is the face annotation system developed by Riya.

Since the performance of K-mean and partial clustering will change with cluster initialization, we thus averaged experimental results on 200 runs of randomized initialization for all experiments so as to give a more convincing comparison.

4.1. Data preparation and experimental protocol

Four disjoint datasets are used in our experiment, all of which are extracted from typical photos in cluttered scenes.

Figure 3. Some example faces cropped from four datasets.

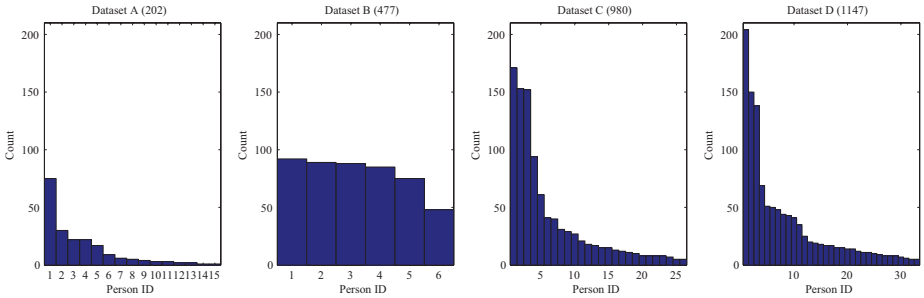

Figure 4. Identities histogram for three datasets

Table 1. The configuration of four datasets used for experiments.

Data Set	A	B	C	D
#Face	202	477	980	1147
#Individual	15	6	27	34

Some face samples cropped from these data sets are shown in Fig. 3.

We show the identity histogram in Fig. 4. Dataset A, C and D are regular family albums. All their distributions coincide with the fact that in typical albums, several core members comprise most samples, while other insignificant identities have only a small portion of the samples. Dataset B comes from an album taken within 5 days, travel snapshots.

For each image, face regions are aligned and cropped using a Haar-based Face Detector [14] and an eye detector [10]. Then for each face regions, Local Binary Pattern(LBP) features are extracted, and contextual Color Corregram[4] features are extracted on clothes areas. For distance measure, we choose chi-square for facial features and L1-distance for contextual features.

As in [8], we use *accuracy* (AC) to measure cluster performance as follows:

$$AC = \frac{\sum_{i=1}^{N} \delta\left(y_i, map(r_i)\right)}{N} \tag{24}$$

where N is the number of faces, $\delta(u, v)$ is the delta function that equals one if $u = v$ and otherwise equals zero, y_i and r_i are the groundtruth label and obtained cluster label respectively, and $map(r_i)$ is a function to map the cluster label to the ground truth label. Different from [8], in this paper, the map function is chosen to map the cluster label to the majority groundtruth label of the each estimated cluster.

To evaluate interactive and overall performance of our framework, a user interaction model is included to count user operations and efficiency.

User operations only occur at the interactive stage. In initial labeling, we assume that typically a user will only label those clusters with more than 70% accuracy. For each cluster, whether skipped or not, we count one *browse* operation. For each cluster to be annotated, we count one *tag* operation for each misplaced face.

During the interactive labeling procedure, the user is asked to label Q in each annotation step. The user interaction count can be estimated by using

$$N(Q) = \min_j [1 + \sum_{x_i \in Q} 1(y_i \neq y_j)], \tag{25}$$

where Q is a suggested subset of unlabel faces for user annotation, function $1(x)$ is the indicator function, and y_j is any possible label.

Labeling efficiency is thus defined as the ratio of the number of faces labeled to the number of user operations.

4.2. Evaluation on unsupervised stage

In this part, we compare the partial clustering algorithm with classic spectral clustering algorithm (K-Means in embedding space), as proposed in [11]. Since we claimed that this algorithm does not aim at overall performance, but on the leading clusters for subsequent initial labeling, the performance measure is then made on the first 80%, 85%, 90% and 95% clusters, with N in Eq. 24 modified accordingly. Fig. 5 shows the result, with both curves using prior-equipped similarity, defined in Eq. 5.

Obviously, in all datasets, partial clustering algorithm outperformed classical spectral clustering in the leading clusters. When more and more clusters are involved in cal-

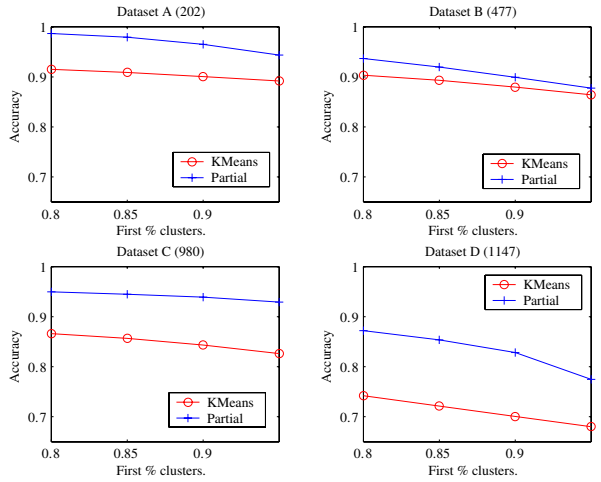

Figure 5. Performance comparison between partial clustering and spectral clustering on the first 80%, 85%, 90% and 95% clusters. This result is averaged over $K \in [20, 100)$.

culation, the result of two algorithms come closer. This shows that our algorithm does not help to produce overall better clusters, but aims to make clusters better, as we expect.

4.3. Evaluation on interactive stage

We also conduct experiments on interactive labeling procedure. This procedure is compared with labeling on (second-time) spectral clustering. After initial labeling by partial clustering, we use the proposed interactive labeling algorithm and spectral clustering respectively to label the rest of the faces. For the latter, we first re-cluster unlabeled faces, then simply label the cluster, whose similarity variation is the lowest.

Results in Fig. 6 show that our proposed labeling algorithm achieves higher efficiency at the beginning of labeling, which results in labeling more faces in fewer steps than labeling on spectral clustering.

4.4. Overall performance evaluation

Since there are few related works in the literature, we compared our framework with two approaches: the face annotation based on pure spectral clustering, and the one-by-one annotation, which is similar to the strategy used by many commercial systems like Photoshop Elements.

The experimental results in Fig. 7 illustrate that our framework outperforms other approaches in four data sets. An obvious turning point in each figure shows the two-stage nature of proposed framework.

We also compare our result with Riya's [1] on data set A. In every iteration, the user was asked to manually label some faces, then the system used this labeled information to recognize faces that belong to the same person, and proposed them for user confirmation. We thus count one labeling as one user operation. As shown in Fig. 8, in 200 repeated experiments, our system needed 42.3 user interactions to annotate the whole data set on average. Compared with 80 user interactions used by Riya, our system outperforms Riya by about 46%.

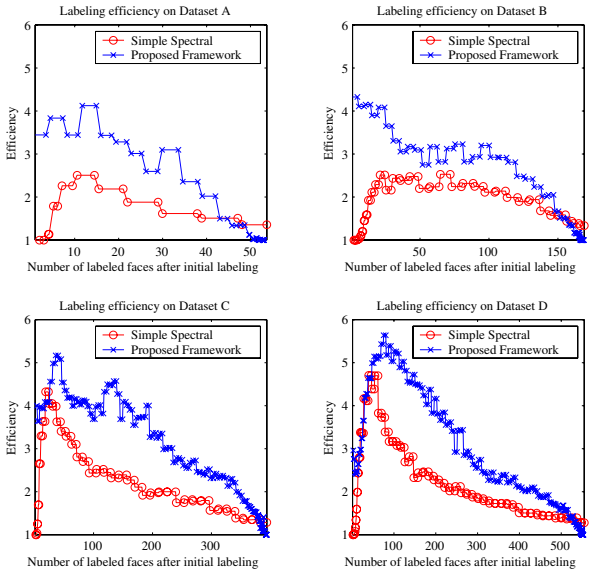

Figure 6. Performance comparison on interactive stage in four datasets.

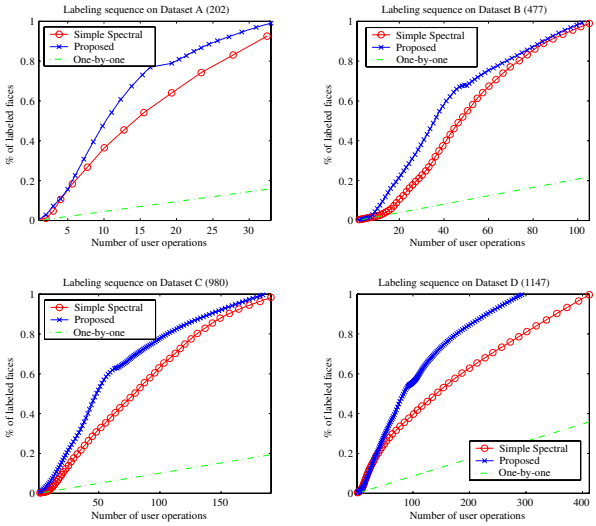

Figure 7. Overall performance comparsion in four datasets.

5. Conclusion and Future Work

In this paper, we propose a novel face annotation framework based on the partial clustering algorithm and the interactive labeling procedure.

There are two main contributions. The first contribution

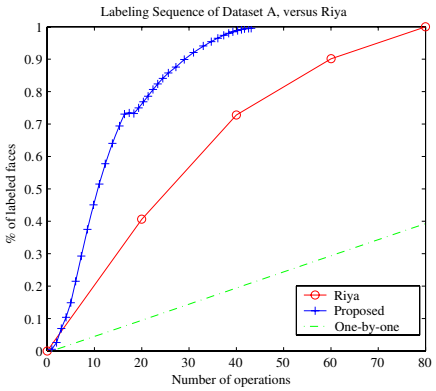

Figure 8. Performance comparison with Riya on dataset A.

is the formulation of the partial clustering algorithm, which aims to reduce user labor rather than improve overall accuracy. The second contribution is the interactive labeling strategy, which maximizes the information gain of each user interaction.

The experimental results show that: 1) using the unsupervised clustering algorithm can significantly reduce the face annotation workload; 2) the partial clustering can group most similar faces into evident clusters to improve the performance of initial labeling; 3) the interactive labeling procedure provides an efficient way to carry out face annotation in the interactive stage. Results compared with Riya also show that the proposed framework is superior for the face annotation task.

There is still much work to further improve the framework. First, pairwise face similarity is important for overall performance improvement. This depends on discriminative facial features and stronger inferring from contextual information. Second, our system adopts a two-stage framework including a clustering stage and an interactive stage. It would be better to integrate the whole system in a compact way, thus eliminating user operations as much as possible. Finally, we plan to integrate the new algorithms here into the EasyAlbum, a photo annotation system we recently developed [5].

References

[1] http://www.riya.com.

[2] T. Ahonen, A. Hadid, and M. Pietikinen. Face recognition with local binary patterns. *Proc. European Conf. on Computer Vision*, pages 469–481, 2004.

[3] V. Belhumeur, J. Hespanda, and D. Kiregeman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. on Pattern Analysis and Machine Intelligence.*, pages 711–720, 1997.

[4] L. Chen, B. Hu, L. Zhang, M. Li, and H. Zhang. Face annotation for family photo album management. *International Journal of Image and Graphics*, pages 1–14, 2003.

[5] J. Cui, F. Wen, R. Xiao, Y. Tian, and X. Tang. EasyAlbum: An interactive photo annotation system based on face clustering and re-ranking. *Proc. CHI 2007. ACM Press*, 2007.

[6] M. Davis, M. Smith, F. Stentiford, A. Bambidele, J. Canny, N. Good, S. King, and R. Janakiraman. Using context and similarity for face and location identification. *Proc. IS&T/SPIE 18th Annual Symposium on Electronic Imaging Science and Technology Internet Imaging*, 2006.

[7] A. Girgensohn, J. Adcock, and L. Wilcox. Leveraging face recognition technology to find and organize photos. *Proc. of the 6th ACM SIGMM international workshop on Multimedia information*, pages 99–106, 2004.

[8] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. *Advances in Neural Information Processing Systems 18*, 2005.

[9] D. Lin and X. Tang. Recognize high resolution faces: From macrocosm to microcosm. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1355–1362, 2006.

[10] Y. Ma, X. Ding, Z. Wang, and N. Wang. Robust precise eye location under probabilistic framework. *Proc. of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 339–344, 2004.

[11] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 14*, 2001.

[12] B. Suh and B. B. Bederson. Semi-automatic image annotation using event and torso identification. *Tech Report HCIL-2004-15, Computer Science Department, University of Maryland, College Park, MD*, 2004.

[13] M. Turk and A. Pentland. Eigenfaces for recognition. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[14] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[15] X. Wang and X. Tang. Random sampling LDA for face recognition. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 259–265, 2004.

[16] X. Wang and X. Tang. A unified framework for subspace face recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pages 1222–1228, 2004.

[17] X. Wang and X. Tang. Random sampling for subspace face recognition. *International Journal of Computer Vision*, pages 91–104, 2006.

[18] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated annotation of human faces in family albums. *Proc. ACM Multimedia*, pages 355–358, 2003.

[19] L. Zhang, Y. Hu, M. Li, W.-Y. Ma, and H. Zhang. Efficient propagation for face annotation in family albums. *Proc. ACM Multimedia*, pages 716–723, 2004.

[20] M. Zhao, Y. Teo, S. Liu, T.-S. Chua, and R. Jain. Automatic person annotation of family photo album. *Proc. International Conf. on Image and Video Retrieval*, pages 163–172, 2006.

[21] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face recognition: A literature survey. *UMD CfAR Technical Report CAR-TR-948*, 2000.