21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

# Unsupervised People Organization and Its Application on Individual Retrieval From Videos

Pengyi HAO, Sei-ichiro KAMATA
*Waseda University, Japan*
*haopy@toki.waseda.jp, kam@waseda.jp*

## Abstract

*In this paper, a method named histogram intersection metric learning from scene tracks is proposed for automatic organizing people in videos. We make the following contributions: (i) learning histogram intersection distance instead of Mahalanobis distance for widely used face features; (ii) learning the metric from scene tracks without manually labeling any examples, which enables learning across large variations in pose, expression, occlusion and illumination with small number of face pairs and can distinguish different people powerfully. We firstly test face identification, track clustering, and people organization on a long film, then individual retrieval based on people organization from a large video dataset is evaluated, demonstrating significantly increased search quality with respect to previous approaches on this area.*

## 1. Introduction

Organizing people in videos according to their face features is so interesting that there are many applications, such as automatically naming or recognizing actors in a video, searching the videos of a desired person.

Owing to the large variations in pose, illumination, occlusion, and expression, face tracks are always used instead of faces in the research of videos [7, 6, 4]. A face track is a collection of detected faces automatically tracked by a straightforward visual tracker. For organizing the face tracks of people appeared in a video, most studies firstly performed agglomerative clustering or K-means to group face tracks as homogenous clusters as possible [8, 4], then the face tracks belonged to the same person can be recognized, named or retrieved as a whole. Recently some researchers began to explore the merits of a learning mechanism. For example, ref. [1] used manually labeled face images and unlabeled face tracks to learn facial attributes to train a classifier for recognizing these attributes in video data, and ref. [2] learned similarity metrics for the characters appearing in a specific video using a few labeled faces for face track identification.

In this paper, distinguished from previous metric learning methods [1, 2] which compute the Euclidean distance in the transformed space after learning a transformation matrix based on labeled faces, we learn histogram intersection similarity metrics adapted to the widely used histogram features like SIFT, LBP for faces without any manual labels. Since histogram intersection distance is commonly used in the research of face images, learning a similarity metric based on histogram intersection leads to an efficient method for face features matching.

Our another contribution focuses on how to automatically generate training examples. In traditional approaches positive pairs coming from the same face track show less appearance variations and negative pairs coming from the same frame can not provide rich information for distinguishing different people. In our approach, training examples are obtained from scene tracks which are collections of face tracks in a time period. Each scene track depicts a person with more variations in pose, expression, and illumination than a single face track. If two faces come from the same scene track, they can be a positive pair. If two faces appear in different scene tracks that exist in the same time period or have an overlap of time, they must not be the same identity, which can form a negative pair. In this way, more valuable examples can be used to learn a metric. At last the learned metric is used in an undirected graph to organize people in one video.

Several applications are evaluated, such as face identification, track clustering, people organization and individual retrieval. Experimental results show that the proposed histogram intersection metric learning performs better on our face features than logistic discriminant metric learning [3] and the people organization approach improves the quality of individual retrieval com-

paring with several state-of-the art approaches.

The rest of the paper is organized as follows: section 2 describes the proposed people organization method, and experiments are reported in section 3, while conclusions and future works are presented in section 4.

## 2. Unsupervised people organization

In this section, first, we present how to generate training examples automatically. Then, histogram intersection metric learning is described in details. Finally, we show how to use the undirected graph to organize people based on the learned metric.

### 2.1. Generating training examples from scene tracks

Firstly, faces detected in different frames of the same shot are associated into face tracks using Kanade-Lucas-Tomasi (KLT) tracker. Then, facial features are used to encode the appearance of the detected faces in each face track. Here, Local Binary Pattern (LBP) descriptors are extracted at five facial components (left and right eyes, tip of the nose, left and right corners of the mouth) at three different scales, which forms a face feature vector of 3840 dimensions.

Since some people may have more than one face track in one scene or a very short time period, to represent a person accurately and perform retrieval quickly, we connect face tracks based on scene information with a time restriction. The connected tracks are called 'scene tracks'. Each scene track depicts one person.

For each face track, it will be measured with its nearby face tracks in a fixed time $t$ to judge whether they belong to the same person or not. Here, because differences between faces of the same person are very small but differences between the corresponding frames of different people are much larger in a short time period, so discrepancies between frames are employed in scene tracks generation. For efficiency, we use key faces to represent each face track and their corresponding key frames are used at here. The distance between face track $FT_i$ and its nearby face tracks eg.$FT_j$ is defined as $dis$ = $d(FT_i, FT_j)+d(F_i, F_j)$, where, $FT_i$ denote the $i$-th face track in a video, and $F_i$ is the set of corresponding key frames of $FT_i$, $d(,)$ is Euclidean distance, $t = 2s$ in our experiment. Once several face tracks in a fixed time are connected by choosing the minimum $dis$, they will be treated as a new face track and the previous process continues until no face track can be connected with each other.

Then we can generate a set of positive examples by collecting the face pairs within each face track of each scene track and the face pairs between every two face tracks in each scene track. Similarly, we can generate a set of negative pairs by collecting all the faces between scene tracks which overlap in time. In this way, the positive pairs generating from scene tracks occur in a longer time than those only taken from face tracks, so the appearance variations among them are larger than within a face track. The number of negative pairs that are collected between scene tracks are larger comparing with the case of only considering the face pairs appearing in the same frame. Thus when some characters occur more frequently than others, they will be easier to be distinguished after metric learning.

### 2.2. Histogram intersection metric learning

Let $ST^i$=$\{FT_1^i, FT_2^i, \ldots, FT_{\omega_i}^i\}$ denote the $i$-th scene track with length $\omega_i$. $FT_u^i$=$\{f_{u1}^i, f_{u2}^i, \ldots, f_{u\kappa_u}^i\}$ denote the $u$-th face track with length $\kappa_u$ in $ST^i$. A distance between scene tracks $ST^i$ and $ST^j$ is defined:

$$D(ST^i, ST^j) = \frac{1}{\omega_i \times \omega_j} \sum_{u,v} \sum_{x,y} \frac{d(f_{ux}^i, f_{vy}^j)}{\kappa_u \times \kappa_v},$$

here, $d(f_{ux}^i, f_{vy}^j)$ is defined by the histogram intersection distance. In the transformed space, it is wrote as:

$$d(f_{ux}^i, f_{vy}^j) = min(\widehat{f_{ux_m}^i}, \widehat{f_{vy_n}^j}) = min(Af_{ux_m}^i, Af_{vy_n}^j),$$

where $A$ is a $d \times D$ transformation matrix, $A \in \Re^{d \times D}$. $D$ denotes the dimensions of feature space, $d \leq D$.

For learning the metric, firstly the histogram intersection distance is modeled using the probability $\rho_{ij}$:

$$\rho_{ij} = \sigma(b - d(f_{ux}^i, f_{vy}^j)) = \frac{1}{1 + exp(d(f_{ux}^i, f_{vy}^j) - b)},$$

where $b$ is a bias term and will be learned together with the metric parameter $A$. According to ref. [5], in order to suppress noise, the transformation matrix A should be regularized, $A_{ij} \in [0, 1]$. We use mixed (2, 1)-norm to force the sparsity of A. Then our objective function can be wrote as:

$$\max_{A,b} \Gamma(A, b) = \sum_{i,j} t_{ij} log\rho_{ij} + (1 - t_{ij}) log(1 - \rho_{ij}) + \lambda \parallel A^T A \parallel_{(2,1)},$$

where $t_{ij}$ denotes the pair label, $\lambda$ is a positive tradeoff parameter. Let $M = A^T A$, the above object function can be transformed as the following equation based on ref. [5],

$$\max_{A,b} \Gamma(A, b) = \sum_{i,j} t_{ij} log\rho_{ij} + (1 - t_{ij}) log(1 - \rho_{ij}) + \lambda tr(M),$$

then it can be solved using gradient descent.

After learning the metric, the learned metric can be used to identify whether the two scene tracks depict the same person or not. The distances between scene tracks with negative pairs will be larger than the distances corresponding to positive pairs, so we can easily find a threshold based on experiments for identification.

### 2.3. Undirected graph

Given a collection of scene tracks, we now wish to divide the collection into several sets where each set depicts a person. We define an undirected graph $G$ for a video, $G = \langle V, E \rangle$, $V$ is the set of scene tracks, each scene track is a vertex. $E$ is the set of edges among vertices. If $ST^i$ and $ST^j$ are identified as the same person, there is an edge between $ST^i$ and $ST^j$, denoted as $(ST^i, ST^j)$. The algorithm of organizing people using undirected graph is listed as follows:

1. Compute the number of connected component, if the number is larger than 1, do steps 2, 3, 4; if no connected component, each vertex depicts a person;

2. For each connected component, if there are circles, firstly the circles that are subsets of some big circles will be incorporated e.g. if circle A is included in circle B, A will be ignored, then the redundant vertices connected with circles will be split;

3. For the rest of vertices that have more than one edge, the edge with the smallest weight will be reserved, and other edges will be deleted;

4. Split $V$ into several sets, each set depicts a person.

Here, two things need to be pointed out. (i) Judge whether a circle exists. Because some errors of identification are unavoidable, so if there exist $(ST^x, ST^y)$ and $(ST^y, ST^z)$, we can not say that $ST^x$, $ST^y$ and $ST^z$ are the same person, unless $ST^x$ and $ST^z$ are also identified as a same person. (ii) When a vertex $ST^x$ has more than one edge, e.g. there are two edges $(ST^x, ST^y)$ and $(ST^x, ST^z)$ while $(ST^y, ST^z)$ is not existed, it just needs to select one edge for $ST^x$ based on its similarities.

## 3. Experiments

The long film "Along Came Polly" came from ref. [4] was used to evaluate face identification, track clustering and people organization. The 90 minutes film gave us 7,332 face tracks, which formed 1,781 scene tracks. There were 93 people according to the manual statistics on these scene tracks. 41 people remained expect the people who had only one or two scene tracks. For the evaluation of individual retrieval, the whole dataset [4] was used, where there are six types of videos: films, TV shows, educational videos, interviews, press conferences and domestic activities.

**Face identification** Firstly, we evaluate face identification on 7,332 face tracks and 1,781 scene tracks using different methods: directly measuring the similarity between face tracks (FT) or scene tracks (ST), logistic discriminant metric learning from face tracks [2] (FT+LDML) or scene tracks (ST+LDML), histogram intersection metric learning from face tracks (FT+HIML) or scene tracks (ST+HIML). Figure. 1 shows the ROC curves which are plotted by computing the true positive rates (TPR) and false positive rates (FPR) on all distance thresholds. We see that scene tracks perform better than face tracks whenever LDML or HIML is used. This is because of that the positive examples obtained from face tracks are not powerful in identifying the same person with large variations on light, expression and pose. Also, the negative pairs obtained from face tracks can not distinguish all the different people who appeare in a very short time period.

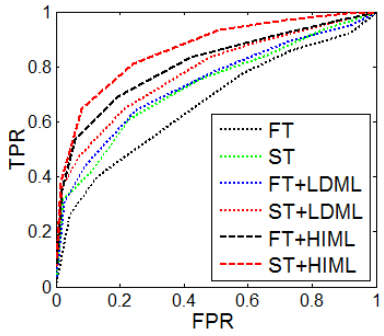
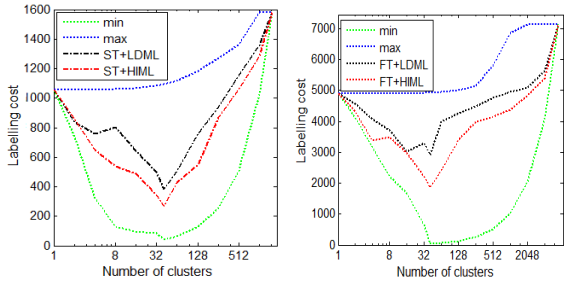



**Figure 1. The ROC curves**

**Track clustering** Secondly, we compare HIML with several methods when agglomerative clustering is used to cluster the 1583 tracks of 41 main people. We use the labeling cost [3] to evaluate. For a cluster of $p$ people with $M$ tracks and the person $i$ has $m_i$ tracks in the cluster, the cost for the cluster is $1 + (M - max\{m_i, i = 1, ..., p\})$. For the film "Along Came Polly", the minimum cost is 41, and when there is only one cluster, the maximum cost is 1,054, because that we have 529 scene tracks of the most frequent character "Reuben Feffer". When using face tracks to perform clustering, the maximum cost is 4885. Figure. 2(a) gives the labeling costs when using scene tracks to do clustering, and figure. 2(b) shows the case of using face tracks. It can be seen that histogram intersection metric learning significantly improves the performance of clutering.

**People organization** Thirdly, we evaluate the ability of organizing people by the following six methods: undirected graph based on HIML from scene tracks




(a) Using scene tracks (b) Using face tracks

**Figure 2. Labeling cost of clustering**

(ST+HIML+UG), undirected graph based on LDML from scene tracks (ST+LDML+UG), LDML from face tracks [2], the hierarchical clustering method given in ref. [4], agglomerative clustering on face tracks (AC on FTs) or on scene tracks (AC on STs). We do not fix the number of final sets for each method. The labeling cost and accuracy are shown in table 1. Clearly, ST+HIML+UG performs much better than others. For ST+HIML+UG, 84.7% of the scene tracks are correctly organized if we label the tracks in each cluster by the identity of the most frequent person in that cluster. AC on FTs gets the worst result. Again, it can be seen that the advantage of HIML and scene tracks not only because of the decreasing errors arose by the variation between faces but also due to the more useful examples.

**Table 1. Labeling cost and accuracy**

| | #people | Cost | Accuracy |
|---|---|---|---|
| ST+HIML+UG | 43 | 286 | 0.847 |
| ST+LDML+UG | 36 | 384 | 0.780 |
| Cinbis2011 [2] | 74 | 2911 | 0.613 |
| Hao2012 [4] | 49 | 674 | 0.605 |
| AC on STs | 51 | 809 | 0.521 |
| AC on FTs | 97 | 4452 | 0.406 |

**Individual retrieval** Finally, mean average precision (mAP) and query time are used to evaluate the performance of searching for videos containing a desired person by performing searches on a 2.66GHz CPU with 8GB memory. Table 2 gives the results of four methods: Sivic's method [7], modeling each face track as a histogram of facial part appearance; k-Faces [6], using k faces selected from face tracks to perform matching; Hao's method [4], organizing each video to a set of people by hierarchical clustering, and the searching based on the proposed people organization method. From table 2 we see that k-Faces gives the lowest mAP. Although it has better performance on news videos, it fails when the k faces of two face tracks have large appearance variations. In contrast, Sivic's method performs slightly better. The proposed method obtains the highest precision than others in respect that distance metrics are learned from powerful positive and negative pairs and most of the scene tracks are correctly grouped into people, but the query time is slightly longer than ref. [4].

**Table 2. Search quality and query time**

| Method | mAP (%) | Query Time (s) |
|---|---|---|
| Sivic's method [7] | 49.87 | 84.5 |
| Nguyen's method [6] | 46.39 | 116.7 |
| Hao's method [4] | 49.25 | 21.9 |
| Proposed method | 57.14 | 25.3 |

## 4. Conclusions and future work

We have demonstrated that learning histogram intersection metrics is effective for matching face features and have shown that learning from scene tracks can improve the accuracy of automatic organizing people in a video. Another conclusion is that the proposed people organization approach can improve the search quality comparing with several state-of-the art approaches. In addition, extending the proposed approach to automatically labeling characters in movies is our future work.

## References

[1] N. Cherniavsky, I. Laptev, J. Sivic, and A. Zisserman. Semi-supervised learning of facial attributes in video. *ECCV*, 2010.

[2] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in tv video. *ICCV*, 2011.

[3] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. *ICCV*, 2009.

[4] P. Y. Hao and S. Kamata. Efficiently finding individuals from video dataset. *IEICE Transactions on Information and Systems*, E95-D(5), May 2012.

[5] K. Huang, Y. Ying, and C. Campbell. Gsml: A unified framework for sparse metric learning. *ICDM*, 2009.

[6] T. Nguyen, T. Ngo, D.-D. Le, S. Satoh, B. Le, and D. Duong. An efficient method for face retrieval from large video datasets. *CIVR*, 2010.

[7] J. Sivic, M. Everingham, and A. Zisserman. Person spotting: Video shot retrieval for face sets. *CIVR*, 2005.

[8] Y. F. Zhang, C. S. Xu, H. Q. Lu, and Y. M. Huang. Character identification in feature-length films using global face-name matching. *IEEE Transactions on Multimedia*, 11(7):1276–1288, November 2009.