AN EFFICIENT VIDEO RETRIEVAL SCHEME BASED ON FACIAL SIGNATURES

Pengyi HAO, Sei-ichiro KAMATA

Waseda University, Japan
Graduate School of Information, Production and Systems

ABSTRACT

The topic of retrieving videos containing a desired person by just using facial content has many applications like video surveillance, social network, etc. In this paper, we propose a compact, discriminative and low-dimensional signature to describe an person with a set of high-dimensional features. The signature is generated by linear discriminant analysis with maximum correntropy criterion that is robust to outliers and noises. Based on the proposed signatures, a new video retrieval scheme is given for fast finding the desired videos by measuring the similarities between the signature of a query and the ones in the dataset. Evaluations on a large dataset of videos show that the proposed video retrieval scheme has the potential to substantially reduce the response time and slightly increase the mean average precision of retrieval.

Index Terms— Video retrieval, Signature, Linear discriminant analysis

1. INTRODUCTION

Recently, video retrieval has become a popular area in the following problem: given a face image of a desired person or a short video clip mainly containing the desired person as a query, all the videos containing the same person with the query should be found from a dataset. There are many applications of such a capability, for example, it would be helpful if all the shots containing the criminal suspect could be found from thousands of video sequences captured by CCTV cameras, or if movies on a website containing an actor of interest could be searched for.

In these applications, faces extracted from videos are usually implied as the most important object comparing with other objects like cloth. So for a retrieval system, a simple way is to match the face(s) in a query against every face extracted from the videos in the dataset. Since the feature vector of a face is usually high-dimensional and the number of faces is large for a large-scale video dataset, it is clear that exhaustive face matching is time complexity.

In order to get a fast video retrieval, a lot of approaches have been presented in this field. Some approaches focus on using face-tracks [1] instead of single face images. Then the number of feature vectors in each face-track is tried to be decreased by using some technologies [1, 2, 3]. Thus, exhaustive face matching is transformed to do matching among face-tracks. Although face-track based approaches exhibit better performance than those using single faces, they still need a long time to do comparison because a large number of face-tracks can be generated from a large-scale video dataset due to the sensitivity of most face trackers on the changes of illumination, occlusions, false face detection, etc. Some approaches focus on narrowing the searching scope based on clustering. Under this scheme, faces extracted from a video dataset are first partitioned into several clusters, and then the query will be introduced into one cluster. Agglomerative clustering and K-Means clustering are the two most used methods [4, 5]. Another example came from [6] applied affinity propagation to generate clusters. In this way, the matching is only limited in a small set. However, over-clustering may cause a low accuracy of retrieval.

In this paper, we propose a novel method for representing a person with a set of faces to be a compact, discriminative and reduced dimensional vector that is called ‘signature’. Fig.1 geometrically interprets the proposed signature, where p_i is a set of high-dimensional vectors. Different with the common signature used in video search and video copy detection where a signature was defined as a small set of key frames extracted from the original video [7], the proposed signature is fixed-length and generated by using a a matcher that is a set of basic vectors. Based on signatures, a efficient video retrieval scheme is given in this paper, where the desired videos are retrieved based on the similarities between the signature of a query and those in the dataset. Because the generated signatures are compact and their similarities can be computed rapidly, our retrieval scheme has low storage requirement and can achieve a fast search. In the experiments, it can achieve an average response time of 0.64s from 790 videos with about one million faces for one query.

The rest of this paper is organized as follows. Section 2 presents the generation of proposed signature, section 3 introduces our video retrieval scheme based on signatures. Experiments are given in section 4. Section 5 concludes this paper.

[1] A face-track is a sequence of faces that are collected from frames by some tracking method and depict the same person.

978-1-4799-2341-0/13/$31.00 ©2013 IEEE

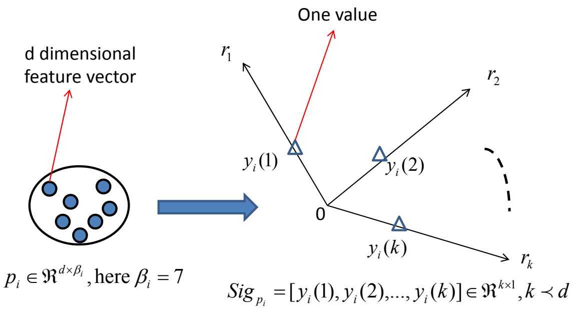
One value
d dimensional feature vector
r_1
r_2
$y_i(1)$
$y_i(2)$
0
$y_i(k)$
r_k
$p_i \in \Re^{d\times\beta_i}$, here $\beta_i = 7$
$Sig_{p_i} = [y_i(1), y_i(2), ..., y_i(k)] \in \Re^{k\times 1}, k \prec d$

Fig. 1. Geometric interpretation of the proposed signature. The signature can be viewed as the projection of an input high-dimensional feature matrix onto a set of basic vectors. The signature is compact, discriminative and fixed-length, which can be fast calculated and matched. k is much smaller than d.

2. SIGNATURE GENERATION

The general idea of the proposed signature generation is that a set of exemplars of an object (e.g. a set of faces of a person) is projected into a compact, discriminative and low-dimensional representation. For generating signatures, a discriminative matcher should be constructed, and a projection strategy is needed.

2.1. Discriminative Matcher Construction

In order to make signatures discriminative, basic vectors (named matcher) should have the ability to maximize the ratio of the between-class distance to the within-class distance. Linear Discriminant Analysis (LDA) is just the technology to achieve this goal. But the conventional LDA based on L2 norm (LDA-L2) [8, 9] is sensitive to the presence of outliers. Rotation invariant L1-norm based LDA (LDA-R1) [10] takes a lot of time to achieve convergence for a large dimensional input space and it can not effectively handle large outliers problem. Recently, we proposed maximum correntropy criterion based LDA (LDA-MCC) [11] that is more robust to outliers and noise. Therefore, LDA-MCC is used to construct the set of basic vectors in this research.

Suppose p_i is an object which is represented as a set of d-dimensional vectors $\{x_h^i\}_{h=1}^{\beta_i}$, $x_h^i \in \Re^d$, and the number of vectors is β_i. Let $X = \{\{\{x_h^i\}_{h=1}^{\beta_i}\}_{i=1}^{n_l}, \Pi_l\}_{l=1}^{C}$ are the samples in C classes. The vector x_h^i is a sample in l-th class Π_l. n_l is the number of individuals in Π_l. Let $\eta_l = \sum_{i=1}^{n_l} \beta_i$ be the sample size of Π_l, and $N = \sum_{l=1}^{C} \eta_l$ be the total number of samples. The mean vector of samples in Π_l is defined as $u_l = \frac{1}{\eta_l}\sum_{i=1}^{n_l}\sum_{h=1}^{\beta_i} x_h^i$. The mean vector of all samples is defined as $u = \frac{1}{N}\sum_{l=1}^{C}\eta_l u_l$. Let $S_b = \sum_{l=1}^{C}(u_l - u)(u_l - u)^T$ be the between-class scatter matrix, and $S_w = \sum_{l=1}^{C}\sum_{i=1}^{n_l}\sum_{h=1}^{\beta_i}(x_h^i - u_l)(x_h^i - u_l)^T$ be the within-class scatter matrix.

Based on LDA-MCC, generating a matcher is turned to find R that maximizes the following objective function:

$$\begin{array}{ll} \underset{R}{max} & J_{MCC} = \sum_{l=1}^{C} g(\sqrt{U_l^T U_l - U_l^T R R^T U_l}), \\ s.t. & R^T S_w R = I, \end{array}$$

where $U_l = u_l - u$ and $g(.)$ is Gaussian kernel.

By updating R according to $(S_w)^{-1} S_b W R = \lambda R$, a matcher $R = \{r_1, \cdots, r_k\} \in \Re^{d\times k}$ can be obtained. Here, W is a diagonal matrix whose diagonal entity $w(l,l) = -w_l$, and $w_l = -g(\sqrt{U_l^T U_l - U_l^T R R^T U_l})$.

2.2. Projection Strategy

After obtaining the matcher R, x_h^i in p_i can be projected into k values by calculating $y_h = R^T x_h^i$, $h \in \{1, \cdots, \beta_i\}$, resulting in

$$Y_i = \begin{bmatrix} y_1(1) & y_2(1) & \cdots & y_{\beta_i}(1) \\ \vdots & \vdots & & \vdots \\ y_1(k) & y_2(k) & \cdots & y_{\beta_i}(k) \end{bmatrix}.$$

Since the exemplar in the same set depict the same object with little variances, their projections in the basic vector space locate nearly. Therefore, we use the following two rules to further shorten Y_i. The shortened Y_i is call 'signature' for p_i. Geometric interpretation of signature is shown in Fig. 1.

• *Average rule* Each row in Y_i is the projections of exemplars of the same object with little variances onto the same basic vector. Thus, the values in each row can be averaged into one value without losing the ability of discrimination. The signature based on *average* rule is obtained as:

$$Sig_{p_i}^{avg} = \frac{1}{\beta_i}\begin{bmatrix} \underbrace{\sum_{b=1}^{\beta_i} y_b(1)}_{1\times1} & \underbrace{\sum_{b=1}^{\beta_i} y_b(2)}_{1\times1} & \cdots & \underbrace{\sum_{b=1}^{\beta_i} y_b(k)}_{1\times1} \end{bmatrix}^T.$$

• *Minimum-mean rule* This rule selects the element in each row which has the minimum mean value by doing the following steps: 1) for each row, calculating the mean distance between each value with all the other values by

$$\begin{array}{l} D(e_a, a) = \frac{1}{\beta_i}\sum_{b=1, b\neq e_a}^{\beta_i} |y_{e_a}(a) - y_b(a)|, \\ \quad a \in \{1, \cdots, k\}, e_a \in \{1, \cdots, \beta_i\}; \end{array}$$

2) selecting the element who has the smallest value in each row according to

$$e_a^* = \arg\underset{e_a}{min}\{D(e_a, a)\}.$$

Finally, the signature based on *Minimum-mean* rule is obtained as

$$Sig_{p_i}^{min-mean} = \begin{bmatrix} \underbrace{y_{e_1^*}(1)}_{1\times1} & \underbrace{y_{e_2^*}(2)}_{1\times1} & \cdots & \underbrace{y_{e_k^*}(k)}_{1\times1} \end{bmatrix}^T.$$

No matter based on which rule, the original $d \times \beta_i$ feature matrix of p_i is transferred to a vector with k values, $k < d$.

3. VIDEO RETRIEVAL BASED ON SIGNATURES

Based on signatures, a new video retrieval scheme is given in this section, where desired videos will be found by measuring the similarities among signatures. In this video retrieval scheme, each video in a dataset is first transferred into a set of people, and then signatures of these people are generated and stored. When given a query, a signature will be first assigned to the query, and then it will be compared to the signatures stored in the dataset for finding a set of candidates. At last, desired videos can be returned based on candidate people.

3.1. Face Clustering

Because a large number of faces can be detected from a video, in the situation that we do not know which faces depict the same person and how many people are in a video, traditional exhaustive face matching not only takes time but also leads to a low accuracy of retrieval. Therefore, the first key step in our retrieval scheme is to cluster the faces extracted from each video so that each cluster depict one person and the faces of the same person belong to the same cluster as much as possible. This goal has been achieved in ref.[6], which is briefly summarized as follows. Firstly, faces in different frames in shots are associated into face-tracks using Kanade-Lucas-Tomasi (KLT) tracker [12]. Local Binary Pattern (LBP) descriptors [13] are extracted at five facial components at three different scales, which forms a feature vector of 3840 dimensions. Secondly, face-tracks of the same person appeared in a short time during (a scene) are connected. Then the connected tracks of the same person located in different parts of a video are grouped together by hierarchical clustering.

After face clustering, the J videos in a dataset are transferred into a set of people $\{P_1, \cdots, P_\rho\}$. Let P_i denote the i-th person in the dataset. P_i is a two-tuples $< p_i, v_i >$, where p_i is its feature set that is described as $p_i = \{FT_1^i, \cdots, FT_{\beta_i}^i\}$, and v_i is the ID of the video where P_i came from, $i \in [1, \rho]$, $v_i \in [1, J]$. β_i is the number of face-tracks in p_i. Note that since one face-track is formed from a shot, the variance between the faces in one face-track is very small. Thus, the mean vector of features in one face-track is used to describe this track, resulting in a $d \times \beta_i$ feature matrix for p_i. So FT_h^i is the h-th feature vector of p_i. Let $x_h^i = FT_h^i$, signature can be generated for P_i according to section 2.

3.2. Similarity Measurement between Signatures

Given a query q, a signature Sig_q will be first generated, and then Sig_q is measured against the ones stored in the dataset. Manhattan distance is employed to measure the similarity between two signatures. If two signatures depict the same person, most of the points of them should be close with each other on the basic vectors, consequently, having a small Manhattan distance. The Manhattan distance between two signatures is calculated by

$$d(Sig_q, Sig_{p_i}) = \sum_{j=1}^{k} |Sig_q(j) - Sig_{p_i}(j))|,$$

where $Sig_q(j)$, $Sig_{p_i}(j)$ are the j-th values in Sig_q and Sig_{p_i} respectively.

After getting the similarities between Sig_q and Sig_{p_i}, $i = 1, \cdots, \rho$, a set of candidate people $\{P_i\}$ can be obtained by using a threshold. Then the videos containing P_i can be got based on the video'ID stored in the structure of P_i.

4. EXPERIMENTS

4.1. Dataset, Evaluation and Training

The whole dataset of ref. [6] is used to evaluate the performance of our video retrieval scheme by using signatures. There are six types of videos in the dataset: films, TV shows, educational videos, interviews, press conferences and domestic activities. Because each film is 90 minutes, each TV show and educational video are also longer than 20 minutes, it is not easy to calculate the accuracy when retrieving a person appeared few times in a long video. Thus we clipped each film into 45 videos, and split each TV show and each educational video to 10 videos respectively, resulting in 790 videos in the dataset. Then 6287 people were extracted from this dataset using the face clustering given in section 3.1.

The mean Average Precision (mAP) is used for evaluations. In the experiments, two kinds of queries are used. One is a face image, the other is a short video clip that mainly contains a person. The query set has 30 people such as "Jennifer Aniston", which covers the six types of videos. Each person in the query set has a face image and a short video clip.

For training a matcher, we manually counted the people extracted from the dataset, resulting in 381 classes. For each class, we randomly selected 2 people, totally 22304 face-tracks, which formed a training set for training the basic vectors (matcher). We selected some vectors from the training set to initialize the matcher according to the following steps: for each vector, first computing its mean value; then sorting the vectors in descending order of their mean values; finally the top k vectors were used for initialization.

4.2. Results

Firstly, which rule performs better and how the size of matcher effects the mean average precision of retrieval are investigated. In this evaluation, the query type is face image.

Fig. 2 shows the mAPs obtained by using average rule and minimum-mean rule with the size of signatures. It can be seen that higher mAPs can be achieved when the number of basic vectors locates from 50 to 70, and the minimum-mean rule consistently preforms better than the average rule. Thus, in the following experiments, $k = 60$ and minimum-mean rule are used.

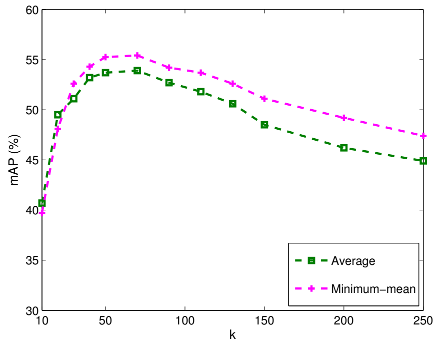

Fig. 2. Comparison of the average rule and the minimum-mean rule. For each k, mAPs (%) are obtained by running the whole query set.

Secondly, the matcher constructed by LDA-MCC is compared to other two matchers: LDA-L2 based and LDA-R1 based. Fig. 3 shows the mAPs achieved by measuring signatures generated by the three matchers respectively, where x-axis corresponds to the number of basic vectors and y-axis is associated with the mAPs. It can be seen that LDA-MCC based matcher can achieve almost 13% higher mAP than LDA-L2 based matcher and about 6% higher than LDA-R1 based matcher. The reason is that LDA-MCC can better characterize the separability of different people and reduce the facial variations among the same person.

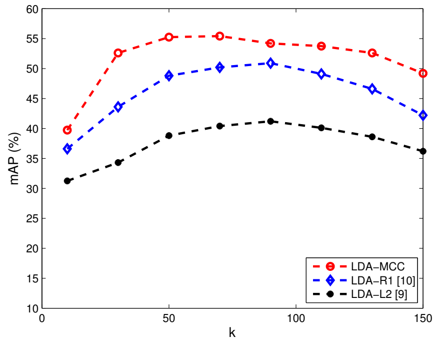

Fig. 3. Comparison of the matcher constructed by our LDA-MCC and the ones based on LDA-L2 and LDA-R1.

Thirdly, the proposed video retrieval scheme for finding the desired person is compared to two kinds of approaches: i) K-Faces [3], an approach of focusing on reducing the number of feature vectors, where K faces were selected from each face-track to compute a mean vector; ii) clubs [6], an approach of focusing on narrowing the retrieving scope, which used affinity propagation to group the face-tracks into clusters and then introduced a query into one or several clusters. The mAPs and retrieval times are given in Table 1. The retrieval time is calculated from submitting a query to obtaining a returned list, and is computed based on an C++ implementation on a 2.66GHz CPU with 8GB memory. As shown in Table 1, K-Faces doesn't perform well, although it had good performance for the news videos given in ref. [3]. The reason is that the six types of videos have larger variations than news videos, which causes the method to fail when the K faces of two face-tracks have different poses, illumination conditions, etc. In contrast, the method of clubs takes less time than K-faces. The proposed retrieval scheme by using signatures performs best both on mAP and retrieval time. If a video clip is taken as a query, it can achieve a little higher mAP than using a face image for retrieval, but using a video clip as a query takes longer time for getting a response.

Table 1. Comparison of our proposed method and two kinds of recently proposed approaches on mAP (%) and retrieval time (s).

	Face image		Video clip	
	mAP	time	mAP	time
K-faces [3]	42.52	116.98	46.41	129.15
Clubs [6]	51.05	8.74	54.13	11.92
Signature	55.46	0.64	58.39	2.73

5. CONCLUSIONS

We proposed a novel retrieval scheme for fast searching videos containing the same person with a query. Under this scheme, a set of faces for a person are projected into only one compact, discriminative and low-dimensional signature by using linear discriminant analysis with maximum correntropy criterion optimization. The desired videos are retrieved by measuring the similarities between the signature of a query and the ones in the dataset. The proposed method significantly decreased the response time of retrieval comparing with two retrieval approaches. At the same time, the mean average precision is slightly improved.

6. REFERENCES

[1] J. Sivic, M. Everingham, and A. Zisserman, "Person spotting: Video shot retrieval for face sets," *Proc. of CIVR*, pp. 226–236, 2005.

[2] M. Everingham, J. Sivic, and A. Zisserman, “Automatic naming of characters in TV video,” *Proc. of BMVC*, 2006.

[3] T. Nguyen, T. Ngo, D.-D. Le, S. Satoh, B. Le, and D. Duong, “An efficient method for face retrieval from large video datasets,” *Proc. of CIVR*, pp. 382–389, 2010.

[4] M. Zhao, J. Yagnik, H. Adam, and D. Bau, “Large scale learning and recognition of faces in web videos,” *Proc. of AFGR*, pp. 1–7, 2008.

[5] 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 Trans. on Multimedia*, vol. 11, no. 7, pp. 1276–1288, 2009.

[6] P. Hao and S. Kamata, “Efficiently finding individuals from video dataset,” *IEICE Trans. on Information and Systems*, vol. E95-D, no. 5, pp. 1280–1287, 2012.

[7] C. Cotsaces, N. Nikolaidis, and I. Pitas, “Face-based digital signatures for video retrieval,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, no. 4, pp. 549–553, APRIL 2008.

[8] G. J. McLachlan, “Discriminant analysis and statistical pattern recognition,” *Wiley*, 1992.

[9] W. Zhao, R. Chellappa, and A. Krishnaswamy, “Discriminant analysis of principal components for face recognition,” *Proc. of AFGR*, pp. 336–341, 1998.

[10] X. Li, W. Hu, H. Wang, and Zh. Zhang, “Linear discriminant analysis using rotational invariant l1 norm,” *Neurocomputing*, vol. 73, no. 13, pp. 2571–2579, 2010.

[11] W. Zhou and S. Kamata, “Linear discriminant analysis with maximum correntropy criterion,” *Proc. of ACCV*, pp. 500–511, 2012.

[12] J. Shi and C. Tomasi, “Good features to track,” *Proc. of CVPR*, pp. 593–600, 1994.

[13] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *PAMI*, vol. 24, no. 7, pp. 971–987, 2002.