Efficient Large-Scale Video Retrieval via Discriminative Signatures

Pengyi HAO (a), Student Member and Sei-ichiro KAMATA (b), Member

SUMMARY  The topic of retrieving videos containing a desired person from a dataset just using the content of faces without any help of textual information has many interesting applications like video surveillance, social network, video mining, etc. However, traditional face matching against a huge number of detected faces leads to an unacceptable response time and may also reduce the accuracy due to the large variations in facial expressions, poses, lighting, etc. Therefore, in this paper we propose a novel method to generate discriminative “signatures” for efficiently retrieving the videos containing the same person with a query. In this research, the signature is defined as a compact, discriminative and reduced dimensional-representation, which is generated from a set of high-dimensional feature vectors of an individual. The desired videos are retrieved based on the similarities between the signature of the query and those of individuals in the database. In particular, we make the following contributions. Firstly, we give an algorithm of two directional linear discriminant analysis with maximum correntropy criterion (2DLDA-MCC) as an extension to our recently proposed maximum correntropy criterion based linear discriminant analysis (LDA-MCC). Both algorithms are robust to outliers and noise. Secondly, we present an approach for transferring a set of exemplars to a fixed-length signature using LDA-MCC and 2DLDA-MCC, resulting in two kinds of signatures that are called 1D signature and 2D signature. Finally, a novel video retrieval scheme is given based on the signatures, which has low storage requirement and can achieve a fast search. Evaluations on a large dataset of videos show reliable measurement of similarities using the proposed signatures to represent the identities generated from videos. Experimental results also demonstrate that the proposed video retrieval scheme has the potential to substantially reduce the response time and slightly increase the mean average precision of retrieval.

key words: individual retrieval, signature, linear discriminant analysis, video dataset

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 desired criminal suspect could be found from thousands of video sequences captured by CCTV cameras. It would be also interesting if movies on a website containing an actor of interest could be searched by using some tracking methods, resulting in face-tracks. The face-track is defined as a collection of faces that should depict the same person. Then the number of feature vectors in each face-track is tried to be decreased by using some technologies. For example, Ref. [1] modeled each face-track as a histogram of facial part appearance. In Ref. [2], a face-track was considered as a cluster of faces. In Ref. [3], K faces selected from each face-track were used to compute the similarity. Although the face-track based approaches showed better performance than those using single faces, they will still need a long time to do matching 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, and so on.

• Reducing the number of feature vectors. This type of approach first collects the faces detected from frames by using some tracking methods, resulting in face-tracks. The face-track is defined as a collection of faces that should depict the same person. Then the number of feature vectors in each face-track is tried to be decreased by using some technologies. For example, Ref. [1] modeled each face-track as a histogram of facial part appearance. In Ref. [2], a face-track was considered as a cluster of faces. In Ref. [3], K faces selected from each face-track were used to compute the similarity. Although the face-track based approaches showed better performance than those using single faces, they will still need a long time to do matching 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, and so on.

Copyright © 2013 The Institute of Electronics, Information and Communication Engineers
proposed signature. 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]–[10], the proposed signature is fixed-length and generated by using a set of basic vectors. In order to make signatures discriminative, basic vectors 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) [11], [12] is sensitive to the presence of outliers. Rotation invariant L1-norm based LDA (LDA-R1) [13] 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) [14] that is more robust to outliers and noise. Figure 2 shows the projections of LDA-MCC and LDA-L2 in the case that outliers are in the between-class and the case that outliers exist in the within-class. No matter which situation, the proposed LDA-MCC can get correct projection space to eliminate the effect of outliers, while LDA-L2 can not separate the classes correctly. Therefore, LDA-MCC is used to construct the set of basic vectors in this research. In particular, we make the following contributions:

1. We give an algorithm of two directional linear discriminant analysis with maximum correntropy criterion (2DLDA-MCC) as an extension to our recently proposed LDA-MCC. Both algorithms are robust to outliers and facial occlusion. And 2DLDA-MCC considers more spacial structure information.

2. We propose an approach for transferring a set of samples to a fixed-length signature using LDA-MCC and 2DLDA-MCC, resulting in two kinds of signatures that are called 1D signature and 2D signature. The proposed signature generation can be extended to other applications, such as generating video signatures for the copy detection of videos, assigning a unique code for an object that contains a set of samples in the field of object recognition.

3. Based on the proposed signature, a new retrieval scheme is proposed, where the desired videos are retrieved based on the similarities between the signature of a query and those of individuals in the database. Because the generated signatures are compact and their similarities can be computed rapidly, our proposed scheme has low storage requirement and can achieve a fast search. In the experiments, it can achieve an average response time of 0.64 s from 790 videos with about one million faces for one query.

The rest of this paper is organized as follows. Section 2 describes some related schemes for getting a fast video retrieval. Section 3 presents the proposed approach of signature generation and the similarity measurements among signatures. Section 4 gives our video retrieval scheme. Experiments are given in Sects. 5 and 6 concludes this paper.

2. Related Works

The study of individual retrieval from videos is a topic of video retrieval. But it has some differences with common video retrieval. Generally, key frames should be detected and used to measure whether two videos similar or not in video retrieval, and color or texture features are usually extracted and matched. In the research of individual retrieval from videos, faces extracted from frames are the important clues instead of key frames detection. The analysis of faces and the measurement among faces are significant tasks.

In order to get a fast video retrieval, video signatures were introduced in some recent works. A video signature usually is a small collection of frames that can be selected by many methods like randomized summarization [7], [8]. Video signatures can also be generated by the analysis of the movement of luminosity [9], or by using semantic information like frame interval given in Ref. [10]. Besides video signatures, fingerprint [15] or index codes [18] were also researched for fast matching.

In most of these traditional video signature approaches, one video has one signature that is used to replace the original data for transporting or matching. In our method, one video is summarized to be a set of individuals. If there are more than one individual summarized from the video, it will have more than one signature. Each signature depicts one
individual.

Another difference with the common video signature is the way of signature generation. Traditional video signature is constructed by collecting a small set of key frames, so the elements of common video signature are features or descriptions of those key frames. In our approach, signature is generated by projecting the input data into a matcher, so the elements in our signature is a set of projections, where the matcher is defined as a set of basic vectors.

Similar concept of matcher can also be found in speaker recognition [16], [17]. In speaker recognition, an enrolled speaker was described by a set of distance (or similarity) scores against a fixed set of reference speakers. In Ref. [18], a set of distance scores (called index code in Ref. [18]) was generated by matching an input image against a set of reference images, and during identification, the index code was compared to the index codes of the enrolled identities. This technique led to a major improvement in computational cost, but the identification rate was usually decreased. And how to select a good reference set has been an important issue of this kind of approaches, because using a larger reference set can improve the identification rate but it will also increase the computational cost, while using a small reference set may not have the discriminability.

In our research, different with the above distance scores based approaches that use the raw data of a dataset as a reference set, we will use our recently proposed LDA-MCC which is robust to outliers and noise to construct a fixed set of basic vectors, and then by projecting the exemplars of an individual to the set of basic vectors, a discriminative, compact and fixed-length signature can be generated. In particular, we present a method of dimensionality reduction. There were also some works in the field of machine learning [19]–[21], which learned some distance metric and at the same time the dimension of each original feature was reduced by the learned metric. Different with these metric learning approaches, our approach can reduce a set of high-dimensional matrices in a single low-dimensional and small-size matrix.

3. Discriminative Signature Generation

The general idea of the proposed signature generation is that a set of exemplars of an individual is projected into a compact, discriminative and low-dimensional representation. Two cases are considered in signature generation. (i) In the case that \( p_i \) is represented as a set of \( d \)-dimensional vectors and the number of vectors is \( \beta_i \), LDA-MCC will be employed to project the \( d \times \beta_i \) matrix to be a vector that has \( k_1 \) compact values. \( k_1 \) is much smaller than \( d \). We call the \( k_1 \times 1 \) vector “1D signature”. (ii) Supposing that each exemplar of \( p_i \) is a matrix whose size can be fixed to \( d \times m \) and there are \( \delta_i \) matrices in \( p_i \), LDA-MCC can not work well in this case since it does not consider the spatial structure information of the matrix. Therefore, we first extend LDA-MCC to be two directional one that is denoted as 2DLDA-MCC, then \( p_i \) will be transferred to be a \( k_2 \times n \) matrix, where \( k_2 \) is much smaller than \( d \) and \( n \) is smaller than \( m \). We call the \( k_2 \times n \) matrix “2D signature”. For easy reference, main symbols used in the paper are listed in Table 1.

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>( J )</td>
<td>the number of videos in a dataset</td>
</tr>
<tr>
<td>( \rho )</td>
<td>the number of individuals extracted from ( J ) videos</td>
</tr>
<tr>
<td>( P_i )</td>
<td>the ( i )-th person in the dataset, ( P_i ) is a two tuples (&lt;p_i, \nu_i&gt;), ( i \in [1, \gamma], \nu_i \in [1, \beta] )</td>
</tr>
<tr>
<td>( p_i )</td>
<td>( p_i ) is a two directional one that is denoted as 2DLDA-MCC</td>
</tr>
<tr>
<td>( r_i )</td>
<td>the ID of video where ( P_i ) came from</td>
</tr>
<tr>
<td>( n_i )</td>
<td>the number of vectors in ( p_i )</td>
</tr>
<tr>
<td>( m_i )</td>
<td>the number of matrices in ( p_i )</td>
</tr>
<tr>
<td>( C )</td>
<td>the number of classes in the training</td>
</tr>
<tr>
<td>( d )</td>
<td>the dimension of feature</td>
</tr>
<tr>
<td>( \Pi_l )</td>
<td>the ( l )-th class</td>
</tr>
<tr>
<td>( n_l )</td>
<td>the number of individuals in ( \Pi_l )</td>
</tr>
<tr>
<td>( X )</td>
<td>( X = { [(x_1^1, \cdots, x_1^m) , \cdots, (x_n^1, \cdots, x_n^m) ] } ) samples in ( C ) classes where ( p_i = { x_1^i, \cdots, x_m^i } )</td>
</tr>
<tr>
<td>( \hat{X} )</td>
<td>( \hat{X} = { [(x_1^1, \cdots, x_1^m) , \cdots, (x_n^1, \cdots, x_n^m) ] } ) samples in ( C ) classes where ( p_i = { \hat{x}_1^i, \cdots, \hat{x}_m^i } )</td>
</tr>
<tr>
<td>( \eta_i )</td>
<td>( \sum_{\delta_i} \beta_i ) the sample size of in ( \Pi_l ) for ( X )</td>
</tr>
<tr>
<td>( \eta_i )</td>
<td>( \sum_{\delta_i} \beta_i ) the sample size of in ( \Pi_l ) for ( \hat{X} )</td>
</tr>
<tr>
<td>( n_i )</td>
<td>( \sum_{\delta_i} \beta_i ) the total numbers of samples in ( X )</td>
</tr>
<tr>
<td>( n_i )</td>
<td>( \sum_{\delta_i} \beta_i ) the total numbers of samples in ( \hat{X} )</td>
</tr>
<tr>
<td>( u_i )</td>
<td>( \frac{1}{n_i} \sum_{\delta_i} x_h^i ) the mean vector of class ( \Pi_l )</td>
</tr>
<tr>
<td>( \hat{u}_i )</td>
<td>the mean matrix of class ( \Pi_l )</td>
</tr>
<tr>
<td>( \eta_i )</td>
<td>( \frac{1}{n_i} \sum_{\delta_i} x_h^i ) the global mean of all the samples</td>
</tr>
<tr>
<td>( \hat{u}_i )</td>
<td>( \frac{1}{n_i} \sum_{\delta_i} x_h^i ) the global mean matrix</td>
</tr>
<tr>
<td>( k_1 )</td>
<td>the length of 1D signature</td>
</tr>
<tr>
<td>( k_2 )</td>
<td>the length of 2D signature</td>
</tr>
</tbody>
</table>

3.1 1D Signatures

Firstly, we consider the case of \( p_i = \{ x_1^i, \cdots, x_m^i \} \), where \( x_h^i \in \mathbb{R}^d, h \in [1, \cdots, \beta_i] \). Define \( S_h = \sum_{i=1}^{C} (u_i - u)(u_i - u)^T \) as the between-class scatter matrix, and \( S_{\infty} = \sum_{h=1}^{\infty} \sum_{i=1}^{C} \sum_{i=1}^{\beta_i} (x_h^i - u)(x_h^i - u)^T \) as the within-class scatter matrix.

By LDA-MCC, generating a matcher is turned to find \( W \) that maximizes the following objective function:

\[
\max_W \ J_{MCC} = \sum_{l=1}^{C} g \left( \sqrt{U_l^T U_l - U_l^T W W^T U_l} \right),
\]

s.t. \( W^T S_w W = I \),

where \( U_l = u_i - u \) and \( g(.) \) is Gaussian kernel, \( g(\sqrt{U_l^T U_l - U_l^T W W^T U_l}) = \exp((U_l^T W W^T U_l - U_l^T U_l)/2\sigma^2) \). \( \sigma \) is the kernel size, which is a free parameter that must be chosen by the user. In the experiments, we follow the lines of correntropy given in Ref. [22] and estimate the bandwidth of kernel using Silverman’s rule given in Ref. [23].

By updating \( W \) according to \((S_w)^{-1} S_b RW = \lambda W \), a matcher \( W = [w_1, \cdots, w_h] \in \mathbb{R}^{d \times k_h} \) can be obtained. Here, \( R \) is a diagonal matrix whose diagonal entity \( r(l, l) = -r_l \),
Then each $x_i^e$ in $p_i$ can be projected into $k_1$ values by calculating $y_{h} = W^T x_i^e$, $h \in \{1, \cdots, \beta_i\}$, resulting in

$$Y_i = \begin{bmatrix} y_{1}(1) & y_{2}(1) & \cdots & y_{\beta_i}(1) \\ \vdots & \vdots & \ddots & \vdots \\ y_{1}(k_1) & y_{2}(k_1) & \cdots & y_{\beta_i}(k_1) \end{bmatrix}.$$ 

Since that face-tracks in the set of one individual depict the same person with very little variances, their projections in the basic vector space locate nearly. Therefore, the following two rules are given to further shorten $Y_i$. The shortened $Y_i$ is called "1D signature" for $P_i$. An example is shown in Fig. 1.

- **Average rule** Each row in $Y_i$ is the projections of faces 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 1D signature based on average rule is obtained by the following equation:

$$S_i^\text{avg}_{p_i} = \frac{1}{\beta_i} \left[ \sum_{h=1}^{\beta_i} y_{h}(1) \sum_{h=1}^{\beta_i} y_{h}(2) \cdots \sum_{h=1}^{\beta_i} y_{h}(k_1) \right]^T.$$ 

- **Minimum-mean rule** This rule selects the element in each row who 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 $D(e_a, a)$; 2) selecting the element who has the smallest value in each row according to $e_a^*$. The algorithm is given in Algorithm 1. The 1D signature based on Minimum-mean rule is obtained as:

$$S_i^\text{min-mean}_{p_i} = \left[ y_{1}^*(1) \ y_{1}^*(2) \cdots y_{1}^*(k_1) \right]^T.$$ 

**Algorithm 1** Algorithm for obtaining 1D signature based on Minimum-mean rule

Require: $Y_i$

for $a = 1$ to $\beta_i$ do

for $e_a = 1$ to $\beta_i$ do

$D(e_a, a) = \frac{1}{\beta_i} \sum_{h=1}^{\beta_i} |y_{h}(a) - y_{h}(a)|$; 

end for

end for

for $e_a \in \{1, \cdots, \beta_i\}$, get $e_a^* = \arg \min_{e_a} D(e_a, a)$; 

$S_i^\text{min-mean}_{p_i}(a) = y_{e_a^*}(a)$;

end for

return $S_i^\text{min-mean}_{p_i}$

No matter based on which rule, 1D signature transfers the original $d \times \beta_i$ feature matrix to be a vector with $k_1$ values, and $k_1 < d$.

### 3.2 2D Signatures

Secondly, we consider the case of $[\vec{x}_1, \cdots, \vec{x}_h]$, where $\vec{x}_h \in \mathbb{R}^{d \times h}$, $h \in \{1, \cdots, \delta_i\}$. Since each exemplar in the input $\vec{X}$ is a two dimensional matrix, we first extend LDA-MCC to 2DLDA-MCC to consider the spacial structure information of matrix. The algorithm of 2DLDA-MCC is given in Algorithm 2, where $\tilde{U}_i = \tilde{u}_i - \bar{u}$. $\tilde{R}$ and $\tilde{\tilde{R}}$ in Algorithm 2 are diagonal matrices whose diagonal entity $\tilde{R}(l, l) = -\tilde{r}_l$ and $\tilde{\tilde{R}}(l, l) = -\tilde{\tilde{r}}_l$ respectively.

**Algorithm 2** 2DLDA-MCC

Require: $\vec{X} = \{([x_{1}^{(1)}, \cdots , x_{1}^{(\delta)}], [h_{1}^{(1)}, \cdots , h_{1}^{(\delta)}]), [x_{2}^{(1)}, \cdots , x_{2}^{(\delta)}], \cdots \}$, $x_i^{(h)} \in \mathbb{R}^{d \times h}$, $k_2 \leq d$, $\alpha \leq m$;

Initialization: $W_L = [w_{1}^{(1)}, w_{2}^{(1)}, \cdots, w_{\delta}^{(1)}] \in \mathbb{R}^{d \times k_2}$, $W_L^T W_L = I$; 

and $W_R = [w_{1}^{(2)}, w_{2}^{(2)}, \cdots, w_{\delta}^{(2)}] \in \mathbb{R}^{d \times m}$, $W_R^T W_R = I$;

while not converge do

1. Calculate $\tilde{r}_l = -g(\sqrt{U_l^T U_l - U_l^T W_L W_L^T U_l})$;

2. Update $W_L$ according to $(S_{r\alpha})^{-1}S_{r\alpha}^\text{WR} = \lambda W_L$;

end while

Update $\vec{X} = (W_L^T \vec{X})^T$;

while not converge do

1. Calculate $\tilde{\tilde{r}}_l = -g(\sqrt{U_l^T U_l - U_l^T W_R W_R^T U_l})$;

2. Update $W_R$ according to $(S_{r\alpha})^{-1}S_{r\alpha}^\text{WR} = \lambda W_R$;

end while

return $W_L \in \mathbb{R}^{d \times k_2}$ and $W_R \in \mathbb{R}^{d \times m}$.

Based on 2DLDA-MCC, two matchers $W_L \in \mathbb{R}^{d \times k_2}$ and $W_R \in \mathbb{R}^{m \times m}$ can be obtained. Then each $\vec{X}$ of $p_i$ is projected to be a $k_2 \times m$ matrix by calculating $\tilde{y}_h = W_L^T \vec{X}_h W_R$, resulting in the following $\tilde{Y}_i$:

$$\tilde{Y}_i = \begin{bmatrix} \tilde{y}_1(1) & \tilde{y}_2(1) & \cdots & \tilde{y}_\beta(1) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{y}_1(k_1) & \tilde{y}_2(k_1) & \cdots & \tilde{y}_\beta(k_1) \end{bmatrix}.$$ 

Similarly, the average rule and minimum-mean rule are applied to further shorten $\tilde{Y}_i$. The shortened $\tilde{Y}_i$ is called "2D signature".

- **Average rule** The 2D signature based on average rule is obtained by the following equation:

$$S_i^\text{avg}_{p_i} = \frac{1}{\delta_i} \left[ \sum_{h=1}^{\delta_i} \tilde{y}_h(1) \sum_{h=1}^{\delta_i} \tilde{y}_h(2) \cdots \sum_{h=1}^{\delta_i} \tilde{y}_h(k_2) \right]^T.$$ 

- **Minimum-mean rule** Based on Minimum-mean rule, elements are selected by doing the following steps: 1) for each row, calculating the mean distance between each vector with all the other vectors $D(E_A, A) = \frac{1}{\delta_i} \sum_{h=1}^{\delta_i} ||\tilde{y}_h(A) - \tilde{y}_h(A)||_2$, $A \in \{1, \cdots, k_2\}$, $E_A \in \{1, \cdots, \delta_i\}$; 2) selecting the element who has the smallest value in each row according to $E_A = \arg \min_{E_A} D(E_A, A)$. The 2D signature with Minimum-mean rule is obtained by

$$S_i^\text{min-mean}_{p_i} = \begin{bmatrix} \tilde{y}_{E_1}(1) & \tilde{y}_{E_2}(2) & \cdots & \tilde{y}_{E_{\delta_i}}(k_2) \end{bmatrix}^T.$$ 

No matter based on which rule, 2D signature transfers
the original \(d \times m \times d_\delta\) features to be a \(k_2 \times n\) matrix. Because of \(k_2 < d\) and \(n < m\), thus \(k_2 \times n < d \times m \times d_\delta\).

### 3.3 Similarity Measurement between Signatures

Several ways are given to measure the distance between two signatures.

1. **Manhattan Distance** \((\ell_1)\). Two signatures are viewed as two sets of points distributed in the space of basic vectors. The distance between two signatures is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes. If two signatures depict the same individual, 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 1D (2D) signatures is shown in the following:

\[
\ell_1(Sig_q^{1D}, Sig_{p_i}^{1D}) = \sum_{j=1}^{k_1} |Sig_q^{1D}(j) - Sig_{p_i}^{1D}(j)|,
\]

where \(Sig_q^{1D}(j), Sig_{p_i}^{1D}(j)\) are the \(j\)-th values in \(Sig_q^{1D}, Sig_{p_i}^{1D}\) respectively.

\[
\ell_1(Sig_q^{2D}, Sig_{p_i}^{2D}) = \sum_{h=1}^{n} \sum_{j=1}^{k_2} |Sig_q^{2D}(h, j) - Sig_{p_i}^{2D}(h, j)|,
\]

where \(Sig_q^{2D}(h, j), Sig_{p_i}^{2D}(h, j)\) are the \(j\)-th value of the \(h\)-th vector in \(Sig_q^{2D}, Sig_{p_i}^{2D}\) respectively.

2. **Euclidean Distance** \((\ell_2)\). Signatures can also be viewed as vectors in a Euclidean space, and the similarity between them can be measured by their spatial proximity. The Euclidean distance between signatures can be calculated as following:

\[
\ell_2(Sig_q^{1D}, Sig_{p_i}^{1D}) = \sqrt{\sum_{j=1}^{k_1} (Sig_q^{1D}(j) - Sig_{p_i}^{1D}(j))^2},
\]

\[
\ell_2(Sig_q^{2D}, Sig_{p_i}^{2D}) = \sqrt{\sum_{h=1}^{n} \sum_{j=1}^{k_2} (Sig_q^{2D}(h, j) - Sig_{p_i}^{2D}(h, j))^2}.
\]

3. **Cosine Distance**. If signatures are viewed as vectors, the similarity between two signatures can be measured by the cosine of the angle between them.

\[
\cos(Sig_q^{1D}, Sig_{p_i}^{1D}) = \frac{<Sig_q^{1D}, Sig_{p_i}^{1D}>}{\|Sig_q^{1D}\| \cdot \|Sig_{p_i}^{1D}\|},
\]

\[
\cos(Sig_q^{2D}, Sig_{p_i}^{2D}) = \frac{<Sig_q^{2D}[h], Sig_{p_i}^{2D}[h]>}{\|Sig_q^{2D}[h]\| \cdot \|Sig_{p_i}^{2D}[h]\|},
\]

where \(Sig_q^{2D}[h], Sig_{p_i}^{2D}[h]\) are the \(h\)-th vector in \(Sig_q^{2D}, Sig_{p_i}^{2D}\) respectively.

4. **Correlation**. Signatures belonging to the same person are expected to have a strong positive correlation, while signatures belonging to different individuals are expected to be uncorrelated. So the correlation between two signatures can also be used to measure the similarity between them.

\[
\rho(Sig_q^{1D}, Sig_{p_i}^{1D}) = \frac{\text{Cov}(Sig_q^{1D}, Sig_{p_i}^{1D})}{\sigma_{Sig_q^{1D}} \cdot \sigma_{Sig_{p_i}^{1D}}},
\]

where \(\text{Cov}(Sig_q^{1D}, Sig_{p_i}^{1D})\) is the covariance between the signatures of \(q\) and \(p_i\), and \(\sigma_{Sig_q^{1D}}, \sigma_{Sig_{p_i}^{1D}}\) are respectively the standard deviation of \(Sig_q^{1D}\) and \(Sig_{p_i}^{1D}\).

\[
\rho(Sig_q^{2D}, Sig_{p_i}^{2D}) = \frac{\sum_{h=1}^{n} \text{Cov}(Sig_q^{2D}[h], Sig_{p_i}^{2D}[h])}{\sigma_{Sig_q^{2D}[h]} \cdot \sigma_{Sig_{p_i}^{2D}[h]}},
\]

where \(\sigma_{Sig_q^{2D}[h]}, \sigma_{Sig_{p_i}^{2D}[h]}\) are respectively the standard deviation of \(Sig_q^{2D}[h]\) and \(Sig_{p_i}^{2D}[h]\).

### 3.4 Extension to Other Types

The signatures and the similarity measurements that were described above are not limited to the case of faces. It is also possible to use them for other types of objects. For example, a video clip containing a set of frames, a person including several fingerprints taken in different periods, a set of samples that are variant in rotations and scales but describe the same object. Because the signatures are fixed-length and compact, so a fast matching can be achieved.

### 4. Our Video Retrieval Scheme Based on Signatures

In this section, we present our video retrieval scheme based on signatures. The whole frame is given in Fig. 3. In our scheme, there are two parts: off-line process and on-line process. In the off-line part, faces detected in each video are first summarized into individuals, whose details will be given in Sect. 4.1. Then signatures of these individuals are generated using the approach given in Sect. 3. In the on-line part, a signature will be first assigned to the query, and then it will be compared to the signatures stored in the database for finding a set of candidates. Based on the similarities of candidate individuals, desired videos will be returned. The querying process will be shown in Sect. 4.2.

4.1 Face Feature and Faces Summarization

Because a large number of faces can be detected in a video, the original \(d \times m \times d_\delta\) features to be a \(k_2 \times n\) matrix. Because of \(k_2 < d\) and \(n < m\), thus \(k_2 \times n < d \times m \times d_\delta\).
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. Therefore, the first key step of the proposed scheme is to summarize the faces extracted from each video to be individuals. One individual is a set of faces which should depict the same person. One video may contain several individuals. The approach given in our previous work [24] is used to achieve this goal, whose flow chart is shown in Fig. 4.

Firstly, faces in different frames are associated into face-tracks using Kanade-Lucas-Tomasi (KLT) tracker [25]. Local Binary Pattern (LBP) [26] 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 feature vector of 3840 dimensions. Secondly, since faces of the same person may disappear in particular frames and then reappear later due to occlusion or turning, the same person may have more than one face-track in one scene, thus face-tracks of the same person are connected to be scene-track based on ‘cannot link’ and ‘may link’ rules with a time restriction [24]. A scene-track is defined as a set of face-tracks came from a short time period and depict the same person. In the experiments, we set the short time period to 8 seconds. Figure 5 gives four scene-tracks and one face-track generated from the film “Along came polly”. Thirdly, unlike the commonly used clustering methods which need to predefined the number of clusters, undirected graph with histogram intersection metric learning [24] is employed to group the scene-tracks of the same person located in different parts of a video together, which forms a set of individuals for each video. The histogram intersection distance metric is learned based on a set of positive face pairs coming from the same scene-track and a set of negative face pairs coming from the two same scene-tracks that have a time overlapping like the four scene-tracks in Fig. 5.

After faces summarization on each video, the faces in a dataset with J videos are transferred into a set of individuals \( \{P_1, \cdots, P_J\} \). Let \( p_i \) denote the \( i \)-th person in a video dataset. \( P_i \) is a two-tuples \( \langle p_i, v_i \rangle \), where \( p_i \) is its feature set and \( v_i \in [1, J] \) is the ID of the video where \( P_i \) came from, \( i \in [1, \rho] \). Two strategies are used to describe \( p_i \): i) by face-tracks, \( p_i = \{FT_1^i, \cdots, FT_{\beta_i}^i\} \); ii) by scene-tracks, \( p_i = \{ST_1^i, \cdots, ST_{\delta_i}^i\} \). \( \beta_i \) and \( \delta_i \) are the numbers of face-tracks and scene-tracks in \( p_i \) respectively. Note that since one face-track is formed from a video in a short time during, 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 \( d \)-dimensional feature vector. And a scene-track is represented by a set of \( d \)-dimensional feature vectors. Let \( \bar{x}_h^i = FT_{h}^i \), 1D signature can be generated for \( P_i \) by Sect. 3.1. Let \( \bar{x}_h^i = ST_{h}^i \), 2D signature can be generated for \( P_i \) by Sect. 3.2.

### 4.2 Querying

According to the process given in Sects 3 and 4.1, each video in a dataset is transformed as a set of individuals and each individual is assigned a 1D signature or 2D signature. All of the generated signatures in a dataset are stored for retrieval. When given a query \( q \), a signature \( S\bar{i}q \) will be first generated, and then the desired videos will be retrieved by calculating the similarities between \( S\bar{i}q \) with the ones stored in the database.

There are two types of queries. (i) When a face image is given, 1D signature will be generated based on matcher \( W \). (ii) When a short video clip mainly including a desired person is given, face-tracks will be firstly extracted. If each face-track is treated as a mean vector, then 1D signature based on matcher \( W \) can be constructed; if using the structure of scene-tracks, 2D signature will be assigned based on matchers \( W_L \) and \( W_R \).

After getting the similarities between \( S\bar{i}q \) and \( S\bar{i}p_m \), \( i = 1, \cdots, \rho \), a set of individuals \( \{P_i\} \) can be obtained by using a threshold. Then the videos containing \( P_i \) can be returned based on the video ID stored in the structure of \( P_i \). An example of querying is given in Fig. 6.
5. Experiments

5.1 Dataset, Evaluation, Query Set, Training and Initialization

Dataset The dataset of Ref. [6] is used to evaluate the performance of the proposed signatures and the video retrieval scheme. 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 divided each long film into 45 videos, and split each TV show to 10 videos, and each educational video was segmented into 10 videos, resulting in totally 790 videos in the dataset. Then 6287 individuals were extracted from this dataset using the faces summarization given in Sect. 4.1.

Evaluation We use precision, recall and mean average precision (mAP), commonly used in content-based image retrieval, to evaluate our methods with several other approaches. The precision is the percentage of the videos which contains the same individual with the query and retrieved by the searching mechanism in the returned list. The recall is the percentage between the number of returned videos containing the same individual with the query and the number of videos including the desired individual in the dataset. The mAP is the mean of average precision scores for a set of queries.

Query set In the process of face detection from videos, a large-size face image of each detection was also automatically obtained. The width of a large-size face image is 3 times greater than that of its corresponding detection and the height of a large-size face image is 4 times greater that of its corresponding detection. We first selected 30 people (2 people for Film, TV shows and Educational videos, respectively; 8 people for Interview, Press and Domestic videos, respectively), and each of them has more than 6 videos in the dataset. Then, we randomly selected 30 large-size face images of them as query images, some examples of query images are given in Fig. 7. The 30 video clips that contain these query images were selected and split to be shorter-length video clips (5 seconds~11 seconds), which are used for querying.

Training and Initialization We manually counted the individuals extracted from the dataset, resulting in 381 classes. For each class, we randomly selected 2 individuals, which formed a training set for generating the basic vectors (matcher). In the training set, there are two collections, one is the collection of face-tracks (22304), and the other is the collection of scene-tracks (7851). The collection of face-tracks was used to train a 1D matcher, and the collection of scene-tracks was used to train a 2D 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 \( k_1 \) \((k_2)\) vectors were used for initialization. For the case of 1D, \( k_1 \) vectors were selected from the training set to initialize \( W \). For the case of 2D, \( k_2 \) vectors were first selected from the training set to initialize \( W_L \), and then after updating \( \tilde{X} \), \( n \) vectors were selected in the same way to initialize \( W_R \). After obtaining a matcher, all the individuals extracted from the video dataset were assigned 1D and 2D signatures that are used in the following evaluations.

5.2 Distance Measurements and Parameters

Firstly, which kind of distance measurement can accurately measure the similarity between signatures will be investigated. We record the first rank of each search contains the same individual with the query, and then use the accuracy of the first ranks of all the queries to evaluate the four similarity measurements given in Sect. 3.3. Figure 8 shows the accuracy obtained by using different similarity measurements with the increase of the dimensions of signatures. Here, queries are images and 1D signatures are used. From Fig. 8, we can see that Manhattan distance and Euclidean distance can get higher accuracy when the dimensions of signatures are 50 to 70. When using cosine or correlation to measure the similarity between signatures, higher accuracy can be achieved when \( k_1 \) locates from 70 to 110. Comparing with

![Fig. 7](image1.png) Some examples of query images which are automatically obtained from videos.

![Fig. 8](image2.png) Comparison with different similarity measurements.
the highest accuracy obtained by these four measurements, using Manhattan distance achieves the highest one 79.3%. In the following experiments, Manhattan distance is used.

Then, which rule is better for generating signature and how many basic vectors can ensure good diversity are tested. Figure 9 draws the mean average precisions of retrieval using 1D signatures and 2D signatures with respect to different numbers of basic vectors obtained by LDA-MCC and 2DLDA-MCC respectively. In this experiment, video clips are used as queries. Average rule and minimum rule are used respectively. From Fig. 9, we can see that 1D signatures can obtain higher mAPs when the dimensions of signatures locate from 50 to 70 and 2D signatures obtain higher mAPs when their dimensions belong to 40 to 60, so in the following experiments we fix $k_1 = 60$ and $k_2 = 45$. It also can be seen that the minimum rule preforms better than average rule. Table 2 shows the results obtained with the change of $n$. From Table 2, we can see clearly that the mAP is a little higher when $n$ equals to 4, so we select $n = 4$ in the following experiments.

### 5.3 Comparison with Different Matcher Generations

In this experiment, the matcher constructed by LDA-MCC not only will be compared with LDA-L2 [12] based matcher and LDA-R1 [13] based matcher but also will be compared with the reference set based matcher. In the reference set based approach, an input image is matched against these reference images to generate a set of distance scores. During identification, the distance scores of the input image is compared to the distance scores of the enrolled identities. In this evaluation, three ways will be used to select a reference set.

1. **Max-mean rule** given in Ref. [18] is used to select a reference set by doing the following steps: i) the whole individual dataset is thought as a candidate pool of reference individuals; ii) for each $p_i$, $i = 1, \cdots, \rho$, compute $\text{mean}_i = \text{Mean} [s(p_i, p_j)]_{j=1,\cdots,\rho}$, where $s(p_i, p_j)$ is the distance score between $p_i$ and $p_j$; iii) sort the individuals in descending order of their $\text{mean}_i$ and select the top $N$ individuals as a reference set.

2. **Do clustering** on the dataset, the number of clusters is $N$. Then the clustering center of each cluster is taken as a reference individual. In this experiment, k-means clustering [27] is used.

3. **Randomly selecting $N$ individuals** from the dataset as a reference set.

Note that there are two differences between our matcher and the reference set based matcher. (i) The elements in the reference set based matcher are raw data came from the dataset, while the elements in our matcher are basic vectors generating from the training set. (ii) The distance scores is a set of distances against the reference set, while our signature is a set of projections.

In order to choose an optimized $N$ for reference set based matcher, we tried several $N$ (50, 100, 200, 300) when using Max-mean rule. The larger value of $N$ resulted in higher mAP, but needed longer matching time. Since the selection of reference set has been experimentally explained in Ref. [18], we will not discuss the selection at here. For comparison, the highest mAP ($N = 300$) is used. Because that both Clustering and Random belong to the same type of approach with Max-mean, so we simply fixed $N = 300$ when using Clustering and Random rules.

Because that the error of scene-tracks generation and individuals formation is unavoidable, there exist some false faces in some sets of individuals, which are outliers for the construction of matcher. And there are also some face occlusions in detected faces. Figure 10 shows some examples of outliers and face occlusions. In the training set and the set of individuals, outliers and face occlusions are included. Since the robustness of our LDA-MCC to outliers and facial occlusion have been experimentally evaluated in our previous work [14], here we just evaluate the performance of LDA-MCC based matcher. For evaluation, the precision and recall of retrieval by searching among distance scores or 1D signatures are computed. We set thresholds from 0.3 to 0.8 to obtain the returned list. For each threshold, the
Table 3  The mAPs obtained by using different matchers.

<table>
<thead>
<tr>
<th>Methods</th>
<th>N or k</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>N=300</td>
<td>36.48</td>
</tr>
<tr>
<td>Clustering [27]</td>
<td>N=300</td>
<td>42.96</td>
</tr>
<tr>
<td>Max-mean [18]</td>
<td>N=300</td>
<td>45.11</td>
</tr>
<tr>
<td>LDA-L2 [12]</td>
<td>k1=60</td>
<td>45.75</td>
</tr>
<tr>
<td>LDA-R1 [13]</td>
<td>k1=60</td>
<td>52.43</td>
</tr>
<tr>
<td>LDA-MCC</td>
<td>k1=60</td>
<td>58.39</td>
</tr>
</tbody>
</table>

Table 4  The time of generating a signature and distance scores for an individual by using different matchers.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Generation time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance scores</td>
<td>228</td>
</tr>
<tr>
<td>1D signature</td>
<td>6</td>
</tr>
<tr>
<td>2D signature</td>
<td>23</td>
</tr>
</tbody>
</table>

precision and recall is calculated for each query, and then the averages of the precision and recall of all the queries are reported in Fig.11. Table 3 shows the mAPs when fix the threshold to 0.6. From Fig.11 and Table 3, it can be seen that LDA based matchers perform better than reference set based approaches on the whole. And LDA-MCC based matcher performs best. It 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 individuals and reduce the facial variations among the same person. Table 4 shows the time of generating distance scores for \( p_i \) and the time of generating a 1D (2D) signature for \( p_i \).

5.4 Comparing with Different Individual Retrieval Approaches

We compare our retrieval scheme to two kinds of individual retrieval approaches: i) K-Faces [3], an approach of focusing on reducing the number of feature vectors, where K faces were selected from the face-track for matching; ii) clubs [6], an approach of focusing on narrowing the scope of retrieval, which used a structure of clubs to introduce a query into one or several clusters. The mAPs and retrieval times are given in Table 5. The retrieval time is calculated from submitting a query to obtaining a returned list, and computed based on an optimized C++ implementation on a 2.66 GHz CPU with 8 GB memory. As shown in Table 5, K-Faces doesn’t perform well in terms of mAP, although it had better performance for the news videos used in Ref. [3]. The reason is that the six types of videos in our dataset 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 and so forth. In contrast, the method of clubs takes less time than K-faces because the videos in the dataset were grouped into several clubs which could limit the range of retrieval. The proposed retrieval scheme based on 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 video clip as a query takes longer time for getting a response. In addition, 2D signatures can achieve higher mAP than using 1D signatures. But it takes a little longer time for retrieving among 2D facial signatures than retrieving among 1D signatures. The reason is that the size of 2D signature is \( k_2 \times n \) which is larger than the size of 1D signature, so it takes time for calculating the distances among 2D signature. In particular, when a face image is used as a query, the retrieval time by proposed scheme can reduce about 99.4% and 92.7% compared with those of K-Faces and the clubs methods respectively.

Figure 12 shows an example of retrieval by using the proposed approach. In this example, the query video clip came from the film “Along Came Polly”. For this query, the number of relevant videos in the dataset is 23. All the first 12 videos in the returned list are relevant videos. In the first 23 retrieved videos, there are 19 relevant ones, resulting in a precision of 82.6%. Figure 13 shows an example of retrieval by using different approaches. In this example, there are six videos relevant to the query in the dataset. The ranks of these relevant videos in the retrieved list achieved by different approaches are compared in Fig. 13. By using the proposed approach, there are four relevant videos in the first six ranks, while other two approaches achieve two relevant videos respectively.

Finally, retrieval time is simulated with the increase of the number of individuals in a dataset. The query type is video clip at here. Let \( n \) denote the current number of face-tracks in a dataset and \( r \) denote the current number of clubs by using the method of [6], \( p' \) denotes the number of individuals after increasing. Assuming that the increase of face-tracks is linear like \( O(\rho \times t) \) and the increase of clubs is like \( O(\rho' \times r) \). The retrieval time of the five methods can be plot-

Table 5  Comparison with two kinds of individual retrieval approaches in terms of mAP (%) and retrieval time (s).

<table>
<thead>
<tr>
<th>Methods</th>
<th>Face image</th>
<th>Video clip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mAP</td>
<td>retrieval time</td>
</tr>
<tr>
<td>K-Faces [3]</td>
<td>42.52</td>
<td>116.98</td>
</tr>
<tr>
<td>Clubs [6]</td>
<td>51.05</td>
<td>8.74</td>
</tr>
<tr>
<td>1D signature</td>
<td>55.46</td>
<td>0.64</td>
</tr>
<tr>
<td>2D signature</td>
<td>( _ )</td>
<td>( _ )</td>
</tr>
</tbody>
</table>
Fig. 12 Example retrieval of the proposed approach. (a) Query video clip, which totally contains 253 frames, and 22 face-tracks are extracted. (b) Face-tracks of the query shown by the first face detection of each face-track. (c) The first 12 retrieved video clips shown by key frames.

Fig. 13 The ranks in the retrieved list obtained by using different approaches. (a) Query video clip, which totally contains 230 frames, and 11 face-tracks are extracted. (b) Face-tracks of the query shown by the first face detection of each face-track. (c) The relevant videos with the query in the dataset. The ranks of the six relevant videos obtained by using different approaches are shown in the bottom of figure.

As expected, when the number of individuals increases with the number of videos, the methods which find desired videos by matching the distances between faces or face-tracks, will fail. The proposed retrieval scheme based on signatures can reduce the retrieval time significantly. And the growth of the retrieval time by using 1D signatures is slowest, which will achieve a response of about 6 seconds from one million individuals.

6. Conclusions

Since conventional methods, which match faces or face-tracks, will fail when the number of faces or face-tracks is very large, in this paper we proposed a novel retrieval scheme for quickly finding 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 response time of retrieval achieved by the proposed scheme is significantly decreased compared with two kinds of state-of-the-art individual retrieval approaches. At the same time, the mean average precision of retrieval is slightly improved.

Acknowledgments

We would like to thank all the people for providing their data for testing and all the observers for contributing to this study.

References


Pengyi Hao is a Ph.D. student of the Graduate School of Information, Production and Systems, Waseda University, Japan. She received her first M.E. degree in computer application and technology from Shanghai University, China, in March 2010, and her second M.E. degree in computer science from Waseda University, Japan, in July 2010. Her current research interests are multimedia retrieval and pattern recognition. She is a student member of the IEEE.

Sci-ichiro Kamata received his M.S. degree in computer science from Kyushu University, Japan, in 1985, and his doctor of engineering degree from the department of Computer Science, Kyushu Institute of Technology, Japan, in 1995. From 1985 to 1988, he was with NEC Ltd., Kawasaki, Japan. In 1988, he joined the faculty at Kyushu Institute of Technology. From 1996 to 2001, he was an associate professor in the Department of Intelligent Systems, Graduate School of Information Science and Electrical Engineering, Kyushu University. Since 2003, he has been a professor at Graduate School of Information, Production and Systems, Waseda University. In 1990 and 1994, he was a visiting researcher at the University of Maine, Orono. His research interests are image processing, pattern recognition, image compression, remotely sensed image analysis, space-filling curves and fractals. Prof. Kamata is a member of the IEEE and the ITE in Japan.