Joint Space Learning for Video-based Face Recognition

Dong Cao[1,2] Ran He[1,2,3] Zhenan Sun[1,2,3] Tieniu Tan [1,2,3]
[1]Center for Research on Intelligent Perception and Computing, CASIA
[2]National Laboratory of Pattern Recognition, CASIA
[3]Center for Excellence in Brain Science and Intelligence Technology, CAS
{dong.cao,rhe,znsun,tnt}@nlpr.ia.ac.cn

Abstract

Popularity of surveillance and mobile cameras provides great opportunities to video-based face recognition (VFR) in less-controlled conditions. This paper proposes a joint space learning method to simultaneously identify the most representative samples and discriminative features from facial videos for reliable face recognition. Specifically, we use a mixture modal by learning multiple feature spaces to capture the data variations where the representative samples in each subspace are learned. Actually, this procedure is a chick to egg problem and an alternate algorithm is developed to monotonically optimize the joint task. In addition, randomized techniques are applied to kernel approximations for capturing the nonlinear structure in data, so that both accuracy and efficiency of our method can be improved. The proposed method performs better than the state-of-the-art video based face recognition methods on Honda, Mobo and YouTube Celebrities databases.

1. Introduction

In the past decade, video-based face recognition (VFR) has attracted much attention, where each sample is a video instead of single image. Intuitively, with more images, more information can be used to identify. However, VFR is still challenging because face videos are usually captured in uncontrolled environments with low qualities, such as pose, illumination, expression and resolution. Generally, there are two major problems: how to alleviate large noise within a video and how to learn a robust feature space where the similarity between two videos can be accurately measured.

To address the two problems, some recent methods try to exploit information in the sample space[19, 3]. These methods often first select a subset of representative samples and then sequentially feed into recognition system which are only designed for small variations of pose and illumination. An improved version of these methods [4] is to first align an image set and then compare the aligned local models rather than the global models. In contrast to the sample space, some recent methods focus on learning on the feature space [12, 13]. They assume that the observed data are generated from a system that is driven by hidden variables in a low dimensional latent space. Hence, they aim to learn discriminative features in a latent space to suppress noise. Since a linear subspace cannot handle the complex intra-class variations very well, some improved methods [13, 17] learn nonlinear model or multiple models. But these methods have high computational complexity.

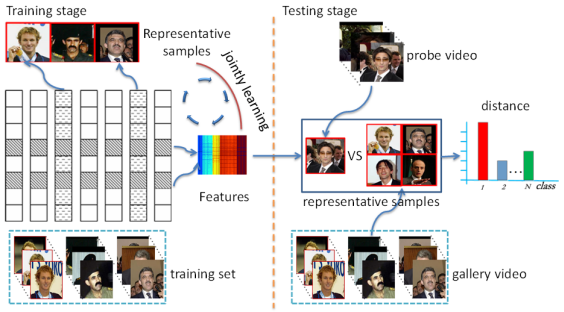


Figure 1. An illustration of the training (left) and testing (right) procedure of our joint space learning method.

Although much progress has been made on sample space learning and feature space learning, to the best of our knowledge, few of the previous works have studied these two problems simultaneously. Intuitively, on one hand, representative samples can can effectively represent intra-class variations and thus facilitate more discriminative features; on the other hand, discriminative features can guide the direction to select more meaningful samples in a video.

To this end, we propose a joint space learning algorithm as shown in Fig 1. An EM-liked alternate optimization algorithm to monotonically decrease the loss function. In each iteration, multiple discriminative feature subspaces are learned by maximizing inter-class distance and minimizing the intra-class distance and pair-wised representa-

tive samples are extracted in each subspace by minimizing the mutual-expression errors of samples. In addition, large intra-class variations often make facial images be a nonlinear distribution [13, 17], kernel technique is often used to map the data vectors into an higher dimensional Hilbert space. To save computational costs, we apply randomized technique for kernel approximation [7] and develop a randomized nonlinear extension of SFL, called RSFL. Extensive experiments demonstrate that our proposed method performs better than state-of-the-art video based face recognition methods.

2. Joint Sample and Feature Space Learning

2.1. Sample Space Learning

In [5], Elhamifar et al. introduced a self-expressiveness method to find representative samples, which assumed that each sample in a set could be described as a linear combination of a few representative samples. However, this method can only find a subset to represent itself. It is not applicable to find samples to represent each other. Inspired by [5], we propose a mutual-expressiveness model to find representative samples for each pair of videos. The basic premise behind is that the distance between images under the same condition is smaller than that under different condition. Suppose $X \in \mathbb{R}^{d\times m}$ and $Y \in \mathbb{R}^{d\times n}$ are two videos. Then we minimize the following problem,

$$\min_{A,B} ||XA - YB||_F^2 \tag{1}$$

with respect to the corresponding coefficient matrix $A = [a_1 \dots a_k] \in \mathbb{R}^{m\times k}$ and $B = [b_1 \dots b_k] \in \mathbb{R}^{n\times k}$, and k indicates the number of representative samples. In other words, we find the representative samples by minimizing the expressiveness error of each data point as a linear combination of all data in other set.

Similarly, we enforce the affine constrain $1^T A = 1^T B$ to preserve the invariant property. That means if $Xa_i = Yb_i$, then $[x_1 - T \cdots x_m - T]a_i = [y_1 - T \cdots y_n - T]b_i$ is steady for any global translation $T \in \mathbb{R}^d$. For simplicity, we define $1^T A = 1^T B = 1^T$.

As a result, we solve

$$\begin{aligned} &\min_{A,B} ||XA - YB||_F^2 \\ &s.t.\ \ 1^T A = 1;\ \ 1^T B = 1^T \end{aligned} \tag{2}$$

The solution can be got by alternatively calculating A and B, i.e., fix A to solve B and vice versa. However, this alternate process is typically inapplicable in real-world due to high computational complexity. In this paper, we propose a regularized linear regression approach by forcing $k = 1$ (i.e., $A \in \mathbb{R}^m$ and $B \in \mathbb{R}^n$). Then XA and YB can be reformulated as

$$XA = [\widehat{x}_1 \widehat{x}_2 \cdots \widehat{x}_{m-1}]A + x_m \tag{3}$$

$$YB = [\widehat{y}_1 \widehat{y}_2 \cdots \widehat{y}_{n-1}]B + y_n \tag{4}$$

where $\widehat{x_i} = x_i - x_m$ and $\widehat{y_i} = y_i - y_n$ The optimal solution can be obtained by

$$\min_{\gamma} ||z - \widehat{XY}\gamma||_2^2 + \lambda||\gamma||_2^2 \tag{5}$$

where

$\widehat{XY}[-\widehat{x}_1 \ \ -\widehat{x}_2 \ \ \cdots \ \ -\widehat{x}_{m-1} \ \ \widehat{y}_1 \ \ \widehat{y}_2 \ \ \cdots \ \ \widehat{y}_{n-1}]$ and $z = x_m - y_n$. The parameter λ sets the trade-off between the two team. The solution $\gamma \in \mathbb{R}^{m+n-2}$ can be estimated by least squares. The desirable representative samples $Rx \in \mathbb{R}^d$ and $Ry \in \mathbb{R}^d$ of X and Y are

$$Rx = [-\widehat{x}_1 \ \ -\widehat{x}_2 \ \ \cdots \ \ -\widehat{x}_{m-1}](\gamma_1 \ \ \gamma_2 \ \ \cdots \gamma_{m-1})^T + x_m \tag{6}$$

$$Ry = [\widehat{y}_1 \ \ \widehat{y}_2 \ \ \cdots \ \ \widehat{y}_{n-1}](\gamma_m \ \ \gamma_{m+1} \ \ \cdots \gamma_{m+n-2})^T + y_n \tag{7}$$

Note that, the value in A (or B) can provide information about the ranking, i.e. relative importance, of the sample in X (Y) for describing another dataset Y (X).

2.2. Feature Space Learning

By sample space learning, the feature space learning can be manipulated using these representative samples instead of original data. Formally, write $V = [V_1, V_2 \cdots V_N]$ as the original video dataset, where $V_i = [v_{i1}, v_{i2}, \cdots, v_{in_i}] \in \mathbb{R}^{d\times n_i}$ is i-th video. Let $\widehat{x}_{i,j} = (x_i, x_j)$ denotes the representative pair learned from $\widehat{V}_{i,j} = (V_i, V_j)$. P and Q denote the inter-class and intra-class sets containing N_p and N_q pairs respectively. Then the training set $\{(\widehat{x}_{i,j}, l_{i,j})\}$ can be built where $l_{i,j} = 0$ if $\widehat{x}_{i,j} \in P$ and $l_{i,j} = 1$ if $\widehat{x}_{i,j} \in Q$. Let $W = [w_1, w_2, \cdots w_m] \in \mathbb{R}^{d\times m}$ be the feature projection matrix. The distance between x_i and x_j in $\widehat{x}_{i,j}$ is

$$d(x_i, x_j) = (x_i - x_j)^T WW^T (x_i - x_j) \tag{8}$$

where the projection W can be learned by maximizing the inter-class distance and minimizing the intra-class distance.

Considering the variations in the face appearance in real-world, we use a mixture modal to capture the data variations by learning multiple feature projections using a shared representation, where different component plays the complementary role to promote each other.

Specifically, in our model, K project matrixes indexed as $W_1, W_2 \cdots W_K$ are learned based on the conditional probability $p(W_k)$. The joint distribution is

$$p = \prod_{i,j=1}^{N} \sum_{k=1}^{K} p(l_{i,j} | \widehat{x}_{i,j}^k, W_k) p(W_k) \tag{9}$$

where

$$p(l_{i,j}|\widehat{x}^k_{i,j},W_k)=(\frac{1}{1+exp(\widehat{d}_k(\widehat{x}^k_{i,j}))})^{l_{i,j}}(\frac{exp(\widehat{d}_k(\widehat{x}^k_{i,j}))}{1+exp(\widehat{d}_k(\widehat{x}^k_{i,j}))})^{1-l_{i,j}} \tag{10}$$

$$\widehat{d}_k(x^k_{i,j})=||x_i-x_j||^2_2 \tag{11}$$

We use an EM algorithm to learn W_k.

E-step

We first project the video dataset V into the low-dimension space W_k, (i.e., $V^t=VW^t_k$). Then the pair-wised representative sample set $\{(\widehat{x}_{i,j},l_{i,j})\}$ can be built by Eqs.(5-7).

$$\begin{aligned} r^t_{i,j,k} &= p(W_k|l_i,\widehat{x}^t_{i,j}) \\ &= \frac{p(W^t_k)\prod p(l_{i,j}|\widehat{x}^t_{i,j},W^t_k)}{\sum_{k=1}^K p(W^t_k)\prod p(l_{i,j}|\widehat{x}^t_{i,j},W^t_k)} \end{aligned} \tag{12}$$

M-step

$$\xi^{t+1}_k=\frac{r^t_k}{N_p+N_q} \tag{13}$$

$$W^{t+1}_k=\arg\max_{W^{t+1}_k} tr((W^{t+1}_k)^T(\frac{S^{'}_q}{M}-\frac{S^{'}_P}{N})(W^{t+1}_k)) \tag{14}$$

where

$$\begin{aligned} S^t_p &= \sum_{i,j=1\ \widehat{x}^t_{i,j}\in P}^{N} r^t_{i,j,k}(x^t_i-x^t_j)(x^t_i-x^t_j)^T \\ S^t_q &= \sum_{i,j=1\ \widehat{x}^t_{i,j}\in Q}^{N} r^t_{i,j,k}(x^t_i-x^t_j)(x^t_i-x^t_j)^T \end{aligned} \tag{15}$$

Here t indicates the iteration number. In the E-step, the sample learning is operated in the learned feature space. The posterior $r^t_{i,j,k}$ indicates how well the low dimensional feature space W_k expresses the sample x_i. If W_k can not explain $\widehat{V}_{i,j}$ well, it would have less contribution to the learning of W_k since the each sample is multiplied by $r^t_{i,j,k}$ in Eq.(20).

Finally, given two videos V_i and V_j, the pair-wised representative samples $\{(Rv^1_i,Rv^1_j),\cdots,(Rv^K_i,Rv^K_j)\}$ can be obtained in the K feature spaces. The similarity between two videos can be estimated using Euclidean distance as

$$d(V_i,V_j)=\sum_{k=1}^K p(W_k)||Rv^k_i-Rv^k_j||^2_2 \tag{16}$$

Then, given a testing video V_{new}, the class c can be obtained by minimizing the distance between V_{new} and V_i in the gallery.

$$c=\arg\min_i d(V_{new},V_i) \tag{17}$$

The algorithm is summarized in Algorithm 1.

Algorithm 1 Joint Sample and Feature Space Learning (SFL)

Input:

The video set $V=[V_1,V_2\cdots V_N]$;

The number of feature projection matrixes K;

Output:

feature projection matrixes $\{W_k\}$ and conditional probability $\{p(W_k)\}$ $(k=1,2\cdots K)$;

Randomly divide V into K parts and calculate the corresponding W_k by Eq.(8) using fisher criterion.
for $t=1$ to T **do**
  for $k=1$ to K **do**
    Project the original data V into K feature spaces $\{W_k\}$ to get $\{V^1,\cdots V^K\}$ where $V^k=VW_k$.
  end for
  Sample Space Learning
  for $k=1$ to K **do**
    Update the training set $\{(\widehat{x}^k_{i,j},l_{i,j})\}$ using Eqs.(5-7);
    Update $r^t_{i,j,k}$ using Eq.(12)
  end for
  Feature Space Learning
  for $k=1$ to K **do**
    Update ξ^{t+1}_k using Eq.(13);
    Update W_k using Eqs.(14-15);
  end for
end for
return $\{W_k\}$ and $\{p(W_k)\}$

2.3. Kernel SFL via Randomized Nonlinear

The feature learning method above can not efficiently deal with nonlinear data due to its inherent linearity. Inspired by these recent advances in kernel approximations [7] , we produce an m-dimensional random feature $z(x)$ for x by randomly sample ω from a data-independent distribution $p(\omega)$.

$$z(x)=[cos(\omega^T_1\cdot x+b_1),\cdots,cos(\omega^T_m\cdot x+b_m)]\in\mathbb{R}^m \tag{18}$$

where $\omega_1,\cdots,\omega_m\sim p(\omega)$ and $b\sim Unif(0,2\pi)$.

For video data, each frame is firstly randomized by Eq.(22), and then fed to SFL, which lead to a randomized version called RSFL. It is easy to verify that the computational complexity of RSFL is $O(m^2n)$, which will be helpful to deal with large scale data.

3. Experiments

Three benchmark video face databases including Honda [11], Mobo [6] and YouTube Celebrities (YTC) [9] are used for evaluation. The Honda/UCSD dataset contains 59 videos of 20 persons, each has at least 2 video. The CMU Mobo dataset contains 96 video sequences of 24 persons.

The YTC is the most challenge dataset. There are 1910 video sequences of 47 celebrities collected from YouTube. For Honda and YTC datasets, we first detect face automatically by the face detector proposed in [14], and then resize the face images to 30×30. In the image pre-processing procedure, only simple Histogram equalization is used to suppress the illumination noise. For Mobo dataset, we directly use the LBP feature as input mentioned in [1].

To make a fair comparison, the same protocol mentioned in the state-of-the-art works [17, 15, 1, 13, 12] is used in our experiments. In our experiments, the feature dimension of W_k and the number of feature projection matrix K are empirically specified as 300 and 3, respectively.

3.1. Results and Analysis

Comparison with other algorithms : We investigate the performance of RSFL in multiple experiments against state-of-the-art methods [18, 10, 17, 15, 1, 8, 16, 2, 13, 12]. The standard implementations of all the compared methods were provided in the original literatures. We follow the same parameter settings. Since there is a single video from each class in the Honda and MoBo datasets for training, for those need within-class sets, we randomly and equally divided each video clip into several partitions to model the within-class variation.

Table 1. The average recognition rates % of methods on the three datasets.

Method	Honda	MoBo	YTC
MSM [18]	92.5	85.5	61.5
DCC [10]	94.9	88.1	64.8
MMD [17]	94.9	91.7	66.7
MDA [15]	97.4	94.4	68.1
AHISD [1]	89.5	94.1	66.5
CHISD [1]	92.5	95.8	67.4
SANP [8]	93.6	96.1	68.3
CDL [16]	97.4	87.5	69.7
DFRV [2]	97.4	94.4	74.5
LMKML [13]	98.5	96.3	78.2
SFDL [12]	100	96.7	76.7
RSFL	**100**	**98.8**	**79.1**

All the recognition results are summarized in Table 1. It is easy to see that, our RSFL gets better performance than the other state-of-the-art algorithms. This is because most other methods only consider one learning task, either in sample space or in feature space, which is not powerful enough to deal with these challenge datasets. However, our joint learning strategy effectively promotes each other and hence outperforms others.

Joint vs. Individual : To analyse the individual effect of the two space learning and the superiority of combination, we compare our RSFL method with the individual sample space learning (SL) and feature space learning (FL) method. More specially, SL directly calculates the pairwised representative samples without learning W. While FL simply uses the mean of the video sequence as representative samples (i.e., $XA = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $YB = \frac{1}{n}\sum_{i=1}^{n} y_i$). Table 2 demonstrates the recognition rates of these three methods. We can find that our proposed joint method outperforms the individual methods because these two tasks are complementary to each other and the iterative procedure can correct the possible error in each task. Also SL gets better performance than FL, illustrating that learning in sample space is more important when dealing with data with large intra-class variations.

Table 2. The average recognition rates % of different sample and feature learning strategy in the three datasets .

Method	Honda	MoBo	YTC
SL	94.9	94.4	75.5
FL	94.9	91.7	66.3
RSFL	100	98.8	79.1

Randomize vs. Linear : To show the advantage of the proposed randomized property, we compare with the linear version called LSFL which directly feeds the original data into the learning algorithm. Table 3 illustrates the performance comparison. We can find that RSFL outperforms LSFL in all the three datasets, especially in the most challenging YTC, illustrating that RSFL can more effectively explain the real-world data than LSFL.

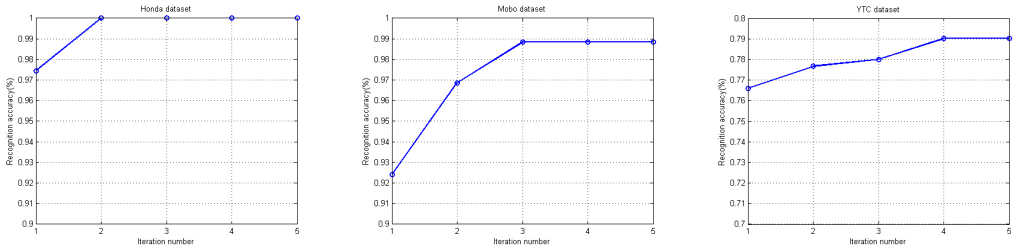
Table 3. The average recognition rates % of linear and randomized nonlinear methods on three datasets.

Method	Honda	MoBo	YTC
LSFL	97.44	97.39	76.7
RSFL	100	98.8	79.1

Table 4. The average recognition rates % of Single-RSFL and RSFL on three datasets.

Method	Honda	MoBo	YTC
Single-RSFL	100	97.2	78.0
RSFL	100	98.8	79.1

Parameter Analysis : We first investigate the effect of the number of components in the mixture modal. Table 4



(a) Honda dataset (b) Mobo dataset (c) YTC dataset

Figure 2. Average recognition rates (%) versus different number of iterations on Honda, Mobo and YTC datasets.

shows the results of Single-RSFL($K = 1$) and RSFL($K = 3$). We can find that RSFL can explain the data better with higher recognition rate.

Another important parameter is the the number of iterations of our RSFL. Fig.2 illustrates the recognition rate versus different number of iterations on Honda, Mobo and YTC. We can find that the performance of the proposed RSFL is stable with small number of iterations.

4. Conclusion

We have proposed a joint space learning method for video-based face recognition, which simultaneously identifies the most representative samples and discriminative features from facial videos. We have developed an alternate minimization algorithm to monotonically optimize the joint learning problem. In each iteration, pair-wised representative samples are extracted by minimizing the mutual expression errors of samples, and discriminative features are learned by maximizing inter-class distance and minimizing the intra-class distance. To further capture the nonlinear structure in data, we have applied randomized techniques to approximate kernels. Experiments demonstrate that our proposed joint space learning method surpasses state-of-the-art results on three commonly used video datasets.

Acknowledgements

This work is funded by the Youth Innovation Promotion Association, CAS (Grant No. 2015190) and the National Natural Science Foundation of China (Grant No. 61135002 and 61473289).

References

[1] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, 2010. 4

[2] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *ECCV*. 2012. 4

[3] Y.-C. Chen, V. M. Patel, S. Shekhar, R. Chellappa, and P. J. Phillips. Video-based face recognition via joint sparse representation. In *FGR*, pages 1–8, 2013. 1

[4] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen. Image sets alignment for video-based face recognition. In *CVPR*, pages 2626–2633, 2012. 1

[5] E. Elhamifar, G. Sapiro, and R. Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *CVPR*, pages 1600–1607, 2012. 2

[6] R. Gross and J. Shi. The cmu motion of body (mobo) database. 2001. 3

[7] R. Hamid, Y. Xiao, A. Gittens, and D. DeCoste. Compact random feature maps. *ICML*, 2014. 2, 3

[8] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, pages 121–128, 2011. 4

[9] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, pages 1–8, 2008. 3

[10] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *PAMI*, pages 1005–1018, 2007. 4

[11] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, pages I–313, 2003. 3

[12] J. Lu, G. Wang, W. Deng, and P. Moulin. Simultaneous feature and dictionary learning for image set based face recognition. In *ECCV*, pages 265–280. 2014. 1, 4

[13] J. Lu, G. Wang, and P. Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, pages 329–336, 2013. 1, 2, 4

[14] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, pages 137–154, 2004. 4

[15] R. Wang and X. Chen. Manifold discriminant analysis. In *CVPR*, pages 429–436, 2009. 4

[16] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, pages 2496–2503, 2012. 4

[17] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*, 2008. 1, 2, 4

[18] O. Yamaguchi, K. Fukui, and K.-i. Maeda. Face recognition using temporal image sequence. In *FGR*, pages 318–323, 1998. 4

[19] M. Yang, P. Zhu, L. Van Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. In *FGR*, pages 1–7, 2013. 1