Pattern Recognition 75 (2018) 33–40

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Collaborative discriminative multi-metric learning for facial expression recognition in video

Haibin Yan

School of Automation, Beijing University of Posts and Telecommunications, Beijing, 100876, China

ARTICLE INFO

Article history:
Received 9 September 2016
Revised 11 January 2017
Accepted 27 February 2017
Available online 2 March 2017

Keywords:
Facial expression recognition
Multi-metric learning
Collaborative learning
Video-based
Multi-view learning

ABSTRACT

Facial expression recognition in video has been an important and relatively new topic in human face analysis and attracted growing interests in recent years. Unlike conventional image-based facial expression recognition methods which recognize facial expression category from still images, facial expression recognition in video is more challenging because there are usually larger intra-class variations among facial frames within a video. This paper presents a collaborative discriminative multi-metric learning (CDMML) for facial expression recognition in video. We first compute multiple feature descriptors for each face video to describe facial appearance and motion information from different aspects. Then, we learn multiple distance metrics with these extracted multiple features collaboratively to exploit complementary and discriminative information for recognition. Experimental results on the Acted Facial Expression in Wild (AFEW) 4.0 and the extended Cohn–Kanada (CK+) datasets are presented to demonstrate the effectiveness of our proposed method.

1. Introduction

Automatic facial expression recognition [1–3] is an important technique to analyze and understand human facial behavior and has many potential applications such as human emotion perception, social advertisement, and human-robotic interaction. Over the past two decades, a variety of facial expression recognition methods have been proposed in the literature and some of them have achieved good performance in controlled environments. However, this problem is still challenging especially when human faces are captured in unconstrained environments as there are usually large variations of pose, illumination, expression and background.

Existing facial expression recognition methods can be mainly categorized into two classes [1–6]: geometric-based and appearance-based. For the first class, local facial features such as the shape and locations of facial components are extracted and their geometrical relationship are described as a feature vector to characterize. For the second category, each face image is represented as a holistical feature vector to represent the texture information. Since it is challenging to precisely localize and extract local geometrical features for face images in many real-world applications, appearance-based methods are more popular than geometric-based methods in facial expression recognition as it can usually achieve higher recognition performance.

Most previous studies of facial expression recognition focus on recognizing human expression from still facial images. In many real applications, it is more convenient to collect facial videos and video can provide more information than images for expression recognition. It is desirable to combine both visual and audio information and make better use of them to improve facial expression recognition. Therefore, the key issue in facial expression recognition in video is how to fuse both visual and audio information in an effective way so that the complementary information can be well exploited.

This paper presents a collaborative discriminative multi-metric learning (CDMML) for facial expression recognition in video. We first compute multiple feature descriptors for each face video to describe facial appearance and motion information from different aspects. Then, we learn multiple distance metrics with these extracted multiple features collaboratively to exploit complementary and discriminative information for recognition. Experimental results on the Acted Facial Expression in Wild (AFEW) 4.0 [7] and the extended Cohn–Kanada (CK+) [8] datasets are presented to demonstrate the effectiveness of our proposed method.

The rest of this paper is organized as follows. In Section 2, we briefly review some related work, and Section 3 presents our proposed CDMML method. Section 4 presents the experimental results and analysis. Section 5 concludes this paper finally.

E-mail address: eyanhaibin@bupt.edu.cn

http://dx.doi.org/10.1016/j.patcog.2017.02.031

2. Related work

In this section, we briefly review two related topics: 1) facial expression recognition, 2) metric learning.

2.1. Facial expression recognition

Conventional facial expression recognition methods first extract facial geometric and appearance information and then employ the classifier for recognition [1–6]. Among these methods, manifold-based methods have been widely considered in recent years because high-dimensional face samples can be considered as a set of geometrically related points lying on or nearby a smooth, low-dimensional manifold. Representative methods include locality preserving projections, orthogonal neighborhood preserving projections, and marginal fisher analysis. These methods have been successfully applied to various facial expression recognition systems. However, most these methods focus on recognizing human expression from still facial images. In many real applications, it is more convenient to collect facial videos in real applications and video can provide more information than images for expression recognition. Hence, it is desirable to combine both visual and audio information and make better use of them to improve facial expression recognition.

2.2. Metric learning

A variety of metric learning methods [9–35] have been widely used in numerous computer vision tasks [9–17]. These methods can be mainly classified into two classes: unsupervised and supervised. The first class of methods learn a low-dimensional manifold to preserve the geometrical information of samples, and the second class of methods seek an appropriate distance metric to exploit the discriminative information of samples. However, most of them are single-metric learning and are not suitable to multi-feature. In this work, we propose a collaborative discriminative multi-metric learning (CDMML) to exploit complementary information for facial expression recognition in video.

3. Proposed method

Let $X = [x_1, x_2, \dots, x_M]$ be a training set of facial videos, where $x_i \in R^d$, $i = 1, 2, \dots, M$, M is the number of samples and d is the feature dimension of each sample. The facial expression class label of x_i is assumed to be $c_i \in \{1, 2, \dots, C\}$, where C is the number of classes. For the jth class, m_j denotes the number of its samples, where $j = 1, 2, \dots, C$. Hence, $M = \sum_{j=1}^{c} m_j$. For each face image, assume there are K different features extracted and $X^k = [x_1^k, x_2^k, \dots, x_M^k]$ is the kth feature representation. For these training samples, we generate a triplet training set $T = \{(x_i^k, y_i^k, z_i^k) | i = 1, 2, \dots, N\}$ which contains N sets of triplet of face videos, where x_i^k, y_i^k and z_i^k are the kth feature descriptor of the ith set of triplet of face videos. In this triplet, x_i^k and y_i^k are from the same expression class, and x_i^k and z_i^k are from different expression classes. Unlike most existing distance metric learning methods which usually directly optimizing the between-class and within-class variations [9,36–38], we employ the probability to measure the positive pairs and negative pairs in each triplet to learn the distance metrics. The key advantage of such a learning strategy is that the distance metrics to be learned will be dominated by some samples which have larger between-class and within-class variations, so that it will be more robust to facial variations because the possible over-fitting problem can be well alleviated. Specifically, in the ith triplet, we have a positive video pair (x_i^k, y_i^k) and a negative video pair (x_i^k, z_i^k) in the kth feature representation space. We learn a distance function $g^k(\cdot)$ to ensure that $g(x_i^k, y_i^k) < g(x_i^k, z_i^k)$, where $1 \le i \le N$. To achieve this goal, we measure the probability of the distance between a positive pair of face video being smaller than that of a negative pair of face video as follows:

$$P\Big(g(x_i^k, y_i^k) < g(x_i^k, z_i^k)\Big) = \Big(1 + \exp(g(x_i^k, y_i^k) - g(x_i^k, z_i^k))\Big)^{-1} \tag{1}$$

where

$$g(x_i^k, y_i^k) = (x_i^k - y_i^k)^T (M_0 + M_k)(x_i^k - y_i^k) \tag{2}$$

$$g(x_i^k, z_i^k) = (x_i^k - z_i^k)^T (M_0 + M_k)(x_i^k - z_i^k) \tag{3}$$

where M_k is a semi-definite matrix learned for the kth feature representation, and M_0 is a semi-definite matrix learned and shared by all feature representation, respectively.

In each triplet, the positive pair and negative pair are generated independently and randomly so that $g(x_i^k, y_i^k) < g(x_i^k, z_i^k)$ and $g(x_j^k, y_j^k) < g(x_j^k, z_j^k)$ are independent. According to the maximum likelihood principle, we formulate our CDMML method with the following optimization objective function:

$$\begin{aligned} \min_{M_0, M_1, \dots, M_K, \alpha} J &= \sum_{k=1}^{K} \alpha_k h_k(M_0, M_1, \dots, M_K) + \lambda l_k(M_0, M_1, \dots, M_K) \\ \text{subject to} \quad & \sum_{k=1}^{K} \alpha_k = 1, \alpha_k \ge 0. \end{aligned} \tag{4}$$

where

$$h_k(M_0, M_1, \dots, M_K) = -\log(\prod_{R^k} P\Big(g(x_i^k, y_i^k) < g(x_i^k, z_i^k)\Big)) \tag{5}$$

$$\begin{aligned} l_k(M_0, M_1, \dots, M_K) = & \sum_{\substack{k_1, k_2 = 1 \\ k_1 \ne k_2}}^{K} \sum_{i=1}^{N} (x_i^{k_1} - x_i^{k_2})^T (M_0 + M_k)(x_i^{k_1} - x_i^{k_2}) \\ & + \sum_{\substack{k_1, k_2 = 1 \\ k_1 \ne k_2}}^{K} \sum_{i=1}^{N} (y_i^{k_1} - y_i^{k_2})^T (M_0 + M_k)(y_i^{k_1} - y_i^{k_2}) \\ & + \sum_{\substack{k_1, k_2 = 1 \\ k_1 \ne k_2}}^{K} \sum_{i=1}^{N} (z_i^{k_1} - z_i^{k_2})^T (M_0 + M_k)(z_i^{k_1} - z_i^{k_2}) \end{aligned} \tag{6}$$

R^k is the triplet set of the kth feature representation, $\alpha = [\alpha_1, \dots, \alpha_K]$ is the weighting parameter, where α_k is the weight of the kth feature, $\lambda > 0$ is a parameter to control the different contributions of these two terms in our objective function.

The physical meaning of (4) is as follows: the first term is to optimize the probability of the distance between the positive pair of sample being smaller than that of the distance between the negative pair of sample in each triplet is as large as possible, and the second term is to maximize the similarity of different feature descriptors of each sample from different feature spaces. We assume that there is a special part and a shared part in the extracted feature so that there are two distance metrics M_0 and M_k employed to compute the similarity of different features of the same sample from different feature spaces, where M_0 is sued to compute similarity of the shared part and M_k is used to compute the similarity of individual part.

Unlike most previous distance metric learning methods which usually directly optimizing the between-class and within-class variations [9,12,36–41], we employ the probability to measure the positive pairs and negative pairs in each triplet to jointly learn multiple distance metrics, which can be more robust to facial appearance variations and alleviate the over-fitting problem.

Since M_i is a semi-define positive matrix, it can be decomposed as follows:

$$M_i = W_i^T W_i \tag{7}$$

where W_k is low-dimensional projection which is decomposed from M_k, $M_k = W_k W_k^T$, and $0 \le i \le K$.

Then, we can rewrite (4) as follows:

$$\min_{W_0, W_1, \ldots, W_K, \alpha} J = \sum_{k=1}^{K} \alpha_k h_k(W_0, W_1, \ldots, W_K) + \lambda l_k(W_0, W_1, \ldots, W_K)$$
$$\text{subject to} \quad \sum_{k=1}^{K} \alpha_k = 1, \alpha_k \ge 0. \tag{8}$$

where

$$h_k(W_0, W_1, \ldots, W_K) = \prod_{R^k} \log(1 + \exp(\|W_k^T x_{ik}^y\|^2 - \|W_k^T x_{ik}^z\|^2)) + \prod_{R^k} \log(1+\exp(\|W_0^T x_{ik}^y\|^2 - \|W_0^T x_{ik}^z\|^2)) \tag{9}$$

$$l_k(W_0, W_1, \ldots, W_K) = \sum_{\substack{k_1,k_2=1\\k_1\ne k_2}}^{K} \sum_{i=1}^{N} \|(W_0 + W_K)^T (x_i^{k_1} - x_i^{k_2})\|_F^2 + \sum_{\substack{k_1,k_2=1\\k_1\ne k_2}}^{K} \sum_{i=1}^{N} \|(W_0 + W_K)^T (y_i^{k_1} - y_i^{k_2})\|_F^2 + \sum_{\substack{k_1,k_2=1\\k_1\ne k_2}}^{K} \sum_{i=1}^{N} \|(W_0 + W_K)^T (z_i^{k_1} - z_i^{k_2})\|_F^2 \tag{10}$$

and $x_{ik}^y = x_i^k - y_i^k$, $x_{ik}^z = x_i^k - z_i^k$.

There is no closed-form solution to the problem defined in (8) because multiple projection matrix and one weighting vector are learned simultaneously. In this work, we use an alternating optimization approach to obtain a local optimal solution. We first fix $W_0, W_1, \ldots, W_{k-1}, W_{k+1}, \ldots, W_K$ and α and solve W_k. Then, we solve α with fixed $W_0, W_1, \ldots, \ldots, W_K$.

When $W_0, W_1, \ldots, W_{k-1}, W_{k+1}, \ldots, W_K$ and α are fixed, (8) can be rewritten as

$$\min_{W_k} J(W_k) = \alpha_k h_k(W_k) + \lambda \sum_{l=1, l\ne k}^{K} L(W_k) \tag{11}$$

where

$$h_k(W_k) = \prod_{R^k} \log(1 + \exp(\|W_k^T x_{ik}^y\|^2 - \|W_k^T x_{ik}^z\|^2)) \tag{12}$$

$$L_k(W_k) = \sum_{\substack{k_1,k_2=1\\k_1\ne k_2}}^{K} \sum_{i=1}^{N} \|W_K^T (x_i^{k_1} - x_i^{k_2})\|_F^2 + \sum_{\substack{k_1,k_2=1\\k_1\ne k_2}}^{K} \sum_{i=1}^{N} \|W_K^T (y_i^{k_1} - y_i^{k_2})\|_F^2 + \sum_{\substack{k_1,k_2=1\\k_1\ne k_2}}^{K} \sum_{i=1}^{N} \|W_K^T (z_i^{k_1} - z_i^{k_2})\|_F^2 \tag{13}$$

We employ the gradient decent method to learn W_k as follows:

$$W_k^{t+1} = W_k^t - \eta \frac{\partial J(W_k)}{\partial W_k} \tag{14}$$

where

$$\frac{\partial J(W_k)}{\partial W_k} = \alpha_k \prod_{R^k} \frac{2 + \exp(\|W_k^T x_{ik}^p\|^2 - \|W_k^T x_{ik}^n\|^2)}{1+\exp(\|W_k^T x_{ik}^p\|^2 - \|W_k^T x_{ik}^n\|^2)} (x_{ik}^p x_{ik}^{p_T} - x_{ik}^n x_{ik}^{n_T}) W_k$$
$$+ 2\lambda(K-1) W_k \sum_{i=1}^{N} (x_i^k)^T x_i^k - 2\lambda W_k \sum_{\substack{l=1\\l\ne k}}^{K} \sum_{i=1}^{N} (x_i^l)^T x_i^l$$
$$+ 2\lambda(K-1) W_k \sum_{i=1}^{N} (y_i^k)^T y_i^k - 2\lambda W_k \sum_{\substack{l=1\\l\ne k}}^{K} \sum_{i=1}^{N} (y_i^l)^T y_i^l$$
$$+ 2\lambda(K-1) W_k \sum_{i=1}^{N} (z_i^k)^T z_i^k - 2\lambda W_k \sum_{\substack{l=1\\l\ne k}}^{K} \sum_{i=1}^{N} (z_i^l)^T z_i^l \tag{15}$$

When $W_0, W_1, \ldots, W_{k-1}, W_k, W_{k+1}, \ldots, W_K$ are fixed, (8) can be rewritten as

$$\min_{\alpha} J(\alpha) = \sum_{k=1}^{K} \alpha_k h_k(W_0, W_1, \ldots, W_K)$$
$$\text{subject to} \quad \sum_{k=1}^{K} \alpha_k = 1, \quad \alpha_k > 0. \tag{16}$$

It seems that the best feature which yields the best performance will be selected from (16). To address this, we modify (16) as follows

$$\min_{\alpha} J(\alpha) = \sum_{k=1}^{K} \alpha_k^p h_k(W_0, W_1, \ldots, W_K)$$
$$\text{subject to} \quad \sum_{k=1}^{K} \alpha_k = 1, \quad \alpha_k > 0. \tag{17}$$

We construct the Lagrange function as follows:

$$S(\alpha, \zeta) = \sum_{k=1}^{K} \alpha_k^r h_k(W_0, W_1, \ldots, W_K) - \zeta(\sum_{k=1}^{K} \alpha_k - 1) \tag{18}$$

Let $\frac{\partial S(\alpha,\zeta)}{\partial \alpha_k} = 0$ and $\frac{\partial S(\alpha,\zeta)}{\partial \zeta} = 0$, we have

$$p\alpha_k^{p-1} h_k(W_0, W_1, \ldots, W_K) - \zeta = 0 \tag{19}$$

$$\sum_{k=1}^{K} \alpha_k - 1 = 0 \tag{20}$$

We solve α_k as follows

$$\alpha_k = \frac{(1/h_k(W_0, W_1, \ldots, W_K))^{1/(p-1)}}{\sum_{k=1}^{K} (1/h_k(W_0, W_1, \ldots, W_K))^{1/(p-1)}} \tag{21}$$

where p is a parameter and $p > 1$.

4. Experiments

To evaluate the performance of the proposed CDMML method for facial expression recognition in video, we conducted experiments on the Acted Facial Expression in Wild (AFEW) 4.0 [7] and the extended Cohn–Kanada (CK+) [8] datasets to show the effectiveness of the proposed method.

4.1. Datasets

The Acted Facial Expression in Wild (AFEW) 4.0 dataset contains facial videos captured in different movies in real world environments. There are three subsets in this dataset: a training set, a validation set, and a testing set, which contains 578, 383, and 307

Fig. 1. Some example image frames on the AFEW 4.0 and the CK+ datasets. From top to bottom are samples from the AFEW 4.0 and the CK+ datasets, respectively.

facial videos, respectively. For each face video in different datasets, one of seven expression labels (anger, disgust, fear, happiness, neutral, sadness, and surprise) is assigned. The original and aligned face videos were provided in the dataset, where the pre-processing method in [42] was employed to align and crop each face from each frame in these videos. Unlike most previous facial expression datasets, facial variations in the AFEW 4.0 dataset are much larger due to the more natural and spontaneous environments.

The Extended Cohn–Kanade (CK+) dataset contains 593 facial videos of 123 persons. Unlike the AFEW dataset, facial images in the CK+ dataset were captured in the lab with controlled conditions. Among these 593 facial videos, 327 of them were labelled and each was classified into one of the following several categories: anger, disgust, fear, happiness, neutral, sadness, and surprise. The number of frames per video varies from 10 to 60, where the expression change for each video was the neutral frame to the expression frame progressively. Unlike the AFEW 4.0 dataset, 68 landmark positions for each image frame were also provided in the CK+ dataset. The positions of these landmarks in key frames were manually labelled, and those for other frames were automatically detected. Fig. 1 shows some example image frames on the AFEW 4.0 and the CK+ datasets.

4.2. Experimental settings

For each face video, we extracted two types of features by following the same settings in [42]: 1) visual feature and 2) audio feature. For visual feature representation, we extracted two different feature descriptors: 1) 3D-HOG and 2) Geometric warp feature. The 3D-HOG is an extension of the conventional 2D HOG [43]. Given a face video, we first obtain three orthogonal planes and then extract HOG features on each plane, respectively. Finally, these histogram features are concatenated into a longer feature vector. Specifically, we first divide each plane into several blocks and extract a HOG feature for each block and then the HOG features in each frame are combined for each frame. We combine those HOG features over the whole video as the final representation of the face video. For each frame in facial videos, we first cropped and resize it into 128×128 and partitioned it into 8×8 blocks, where each block size is 16×16. Each block was represented as one 9-dimensional HOG feature and the whole face video was represented as a $3 \times 9 \times 8 \times 8 = 1728$ dimensional feature vector.

For the geometric warp feature, we first obtain some landmarks for each face image [42]. For each face image frame with expression, there are some facial motion among neighboring frames so that facial image can be considered as the displacements of facial landmarks. Generally, each face image can be considered as many sub-regions and these sub-regions can generate many triangles with the corresponding vertexes located at facial landmarks. Then, the displacements of facial landmarks can be considered as shape feature for facial expression representation. Specifically, each face image was annotated as 68 facial landmarks, and these landmarks divided each face many non-overlapped sub-regions. In this work, we took 109 pair of triangles and used 6 parameters to measure facial expression of the transformation. Therefore, the whole face video was represented by these warp transform coefficients, which was a feature vector of $6 \times 109 = 654$ dimensions.

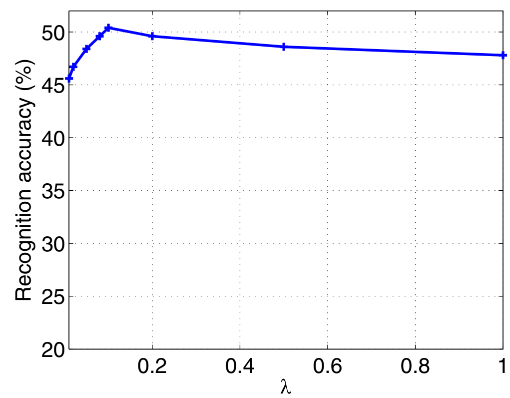

Fig. 2. The recognition accuracy of CDMML with 5-fold cross validation on the training set of the AFEW dataset versus different values of λ.

For the audio feature, we computed the acoustic features and employed 21 functionals and removed 16 zero-information features [42]. Therefore, a total of 1582 acoustic features were extracted from for each video. In our work, we used the open-source Emotion Affect Recognition (openEAR) toolkit to extract the audio features.

Having obtained these three features, we applied PCA to project each feature into 150 dimensions for multi-metric learning with samples in the training set. The PCA projection matrices are also used in for the testing samples before using the learned distance metrics to compute the similarity of samples.

4.3. Results and analysis

This subsection presents the results and analysis of our method for facial expression recognition in video.

4.3.1. Parameter determination

We first determined the parameter of λ on the training set of the AFEW 4.0 dataset. Specifically, we employed the 5-fold cross validation strategy to select the parameter of λ. Fig. 2 shows the recognition rate of our approach versus different values of λ on

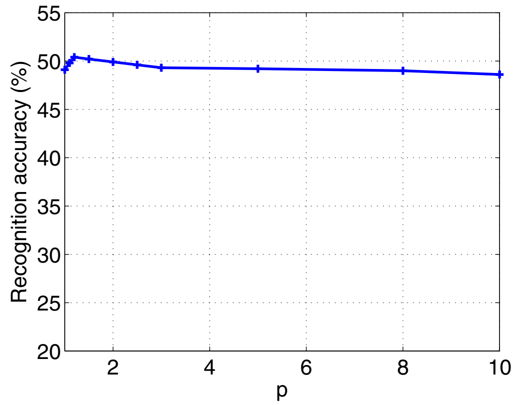

Fig. 3. The recognition accuracy of CDMML with 5-fold cross validation on the training set of the AFEW dataset versus different values of λ.

Table 1
Comparison of the recognition accuracies (%) of different methods on the AFEW and CK+ datasets.

Method	Feature	AFEW 4.0	CK+	Mean
Single-metric learning	HOG-TOP feature	35.8	92.6	64.2
Single-metric learning	Geometric feature	29.8	91.3	60.6
Single-metric learning	Audio feature	32.8	91.5	62.2
CDMML	All feature	**46.8**	**96.6**	**71.7**

the training set of the AFEW 4.0 dataset. We see that the optimal λ was determined as 0.1.

We also determined the parameter of p on the training set of the AFEW 4.0 dataset. Specifically, we used the 5-fold cross validation strategy to select the parameter of p. Fig. 3 shows the recognition rate of our approach versus different values of p on the training set of the AFEW 4.0 dataset. We see that the optimal λ was determined as 1.2.

4.3.2. Multi-metric learning vs. single-metric learning

We first compared our method with single metric learning to show the advantages of the proposed method. Specifically, we learn a single distance metric with a single feature descriptor. The recognition accuracies of different methods are shown in Table 1. We see that our multi-metric learning method achieves better performance than the single-metric learning method because more feature information can be utilized.

4.3.3. Comparisons of different multi-metric learning methods

We compared our method with existing multi-metric learning methods for facial expression recognition in video. Specifically, we compared our CDMML method Multi-feature Canonical Correlation Analysis (MCCA) [44], Multi-feature Marginal Fisher Analysis (MMFA) [44], Discriminative Multi-Manifold Analysis (DMMA) [18], Multi-view Neighborhood Repulsed Metric Learning (MNRML) [45], and Discriminative Multi-Metric Learning (DMML) [46]. The parameters of these methods are set based on the recommendations of these papers. Table 2 shows the recognition accuracies of different multi-metric learning methods. As can be seen, our CDMML outperforms all other compared multi-metric learning methods in terms of the mean recognition accuracy.

4.3.4. Comparisons of the state-of-the-arts

We also compared our method with the state-of-the-art method for facial expression recognition in video in [42], where multiple feature descriptors were also employed for recognition. Table 3 shows the recognition accuracies of different multi-metric learning methods. We see that our CDMML outperforms all other compared multi-metric learning methods in terms of the mean recognition accuracy.

Table 2
Comparison of the recognition accuracies (%) of different methods on the AFEW 4.0 and CK+ datasets.

Method	AFEW 4.0	CK+	Mean
MCCA	37.8	92.6	65.2
MMFA	38.6	93.5	66.1
DMMA	40.6	94.7	67.7
MMNRML	42.6	94.8	68.7
DMML	44.5	95.3	69.9
CDMML	**46.8**	**96.6**	**71.7**

Table 3
Comparison of the recognition accuracies (%) of different methods on the AFEW 4.0 and CK+ datasets.

Method	AFEW 4.0	CK+	Mean
Method in [42]	45.2	95.7	70.4
CDMML	**46.8**	**96.6**	**71.7**

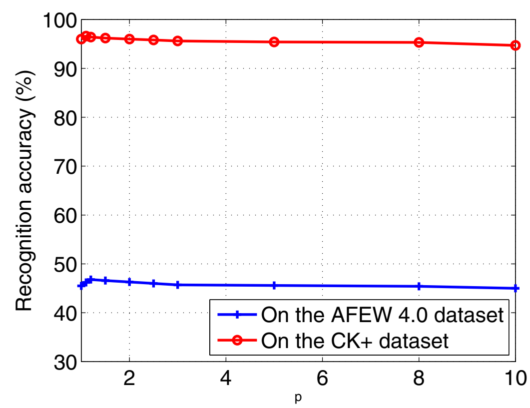

Fig. 4. The recognition accuracy of CDMML versus different values of p.

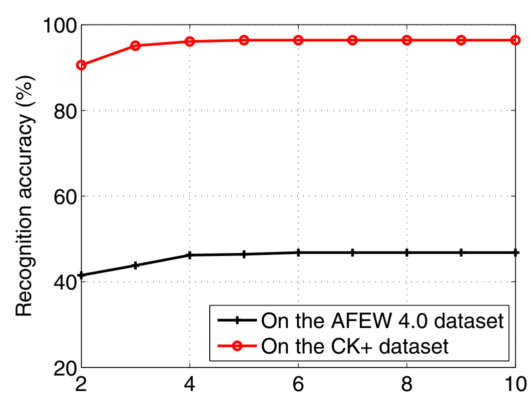

Fig. 5. The recognition accuracy of CDMML versus different number of iterations.

4.3.5. Parameter analysis

We investigated the importance of the parameter of p in our CDMML. Fig. 4 shows the recognition accuracy of CDMML versus p on the AFEW 4.0 and CK+ datasets. We see that our CDMML achieves stable performance across a large range of p.

Fig. 5 shows the recognition accuracy of CDMML versus different number of iterations on different datasets. We see that our CDMML achieve stable recognition rate within a few number of iterations.

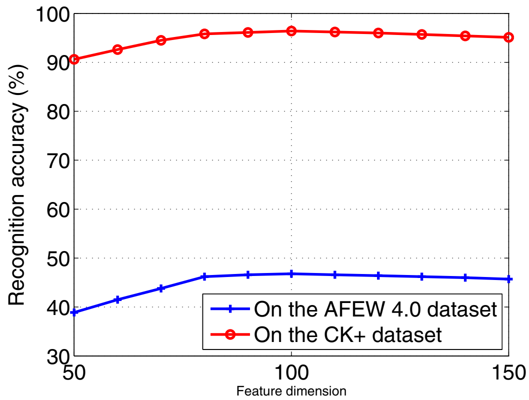

Fig. 6. The recognition accuracy of CDMML versus different feature dimensions.

Fig. 6 shows the recognition accuracy of CDMML versus different number of feature dimension. We see that our CDMML achieves stable recognition accuracy when the feature dimension is larger than 80.

4.4. Discussions

We make the following observations from experimental results listed in Tables 1, 2 and 3 and Figs. 2, 3, 4, 5 and 6:

- Our CDMML achieves better performance than single-metric learning because more feature information can be utilized.
- Our CDMML outperforms all other compared multi-metric learning methods in terms of the mean recognition accuracy.
- Our CDMML consistently outperforms the state-of-the-art video-based facial expression recognition methods.

5. Conclusion

In this paper, we have proposed a collaborative discriminative multi-metric learning (CDMML) for facial expression recognition in video. Experimental results on the AFEW 4.0 and CK+ datasets are presented to demonstrate the effectiveness of our proposed method.

In our future work, we plan to design more efficient feature learning methods and combine them with our CDMML to further improve the performance of facial expression recognition in video.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61603048, the Beijing Natural Science Foundation under Grant 4174101, and the Fundamental Research Funds for the Central Universities.

References

[1] R.A. Calvo, S. D'Mello, Affect detection: an interdisciplinary review of models, methods, and their applications, IEEE Trans. Affect. Comput. 1 (1) (2010) 18–37.
[2] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 39–58.
[3] H. Yan, M.H. Ang Jr, A.N. Poo, A survey on perception methods for human–robot interaction in social robots, Int. J. Soc. Robot. 6 (1) (2014) 85–119.
[4] H. Yan, J. Lu, X. Zhou, Prototype-based discriminative feature learning for kinship verification, IEEE Trans. Cybern. 45 (11) (2015) 2535–2545.
[5] H. Yan, Transfer subspace learning for cross-dataset facial expression recognition, Neurocomputing 208 (2016) 165–173.
[6] H. Yan, Biased subspace learning for misalignment-robust facial expression recognition, Neurocomputing 208 (2016) 202–209.
[7] A. Dhall, R. Goecke, J. Joshi, M. Wagner, T. Gedeon, Emotion recognition in the wild challenge 2013, in: ACM on International Conference on Multimodal Interaction, 2013, pp. 509–516.
[8] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression, in: 2010 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2010, pp. 94–101.
[9] M. Guillaumin, J. Verbeek, C. Schmid, Is that you? metric learning approaches for face identification, in: IEEE International Conference on Computer Vision, 2009, pp. 498–505.
[10] J. Lu, Y.-P. Tan, Regularized locality preserving projections and its extensions for face recognition, IEEE Trans. Syst. Man Cybern. Part B 40 (3) (2010) 958–963.
[11] M. Köstinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.
[12] D. Tran, A. Sorokin, Human activity recognition with metric learning, in: European Conference on Computer Vision, 2008, pp. 548–561.
[13] B. Xiao, X. Yang, Y. Xu, H. Zha, Learning distance metric for regression by semidefinite programming with application to human age estimation, in: ACM International Conference on Multimedia, 2009, pp. 451–460.
[14] W.-S. Zheng, S. Gong, T. Xiang, Person re-identification by probabilistic relative distance comparison, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 649–656.
[15] A. Mignon, F. Jurie, Pcca: A new approach for distance learning from sparse pairwise constraints, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2666–2672.
[16] J. Hu, J. Lu, Y.-P. Tan, Discriminative deep metric learning for face verification in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1875–1882.
[17] J. Lu, G. Wang, P. Moulin, Human identity and gender recognition from gait sequences with arbitrary walking directions, IEEE Trans. Inf. Forensics Secur. 9 (1) (2014) 51–61.
[18] J. Lu, Y.-P. Tan, G. Wang, Discriminative multimanifold analysis for face recognition from a single training sample per person, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 39–51.
[19] J. Lu, Y.-P. Tan, Ordinary preserving manifold analysis for human age and head pose estimation, IEEE Trans. Hum. Mach. Syst. 43 (2) (2013) 249–258.
[20] J. Lu, Y.-P. Tan, Uncorrelated discriminant nearest feature line analysis for face recognition, IEEE Sig. Process. Lett. 17 (2) (2010) 185–188.
[21] J. Lu, Y.-P. Tan, A doubly weighted approach for appearance-based subspace learning methods, IEEE Trans. Inf. Forensics Secur. 5 (1) (2010) 71–81.
[22] J. Lu, Y.-P. Tan, Cost-sensitive subspace learning for face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2661–2666.
[23] J. Lu, Y.-P. Tan, Nearest feature space analysis for classification, IEEE Sig. Process. Lett. 18 (1) (2011) 55–58.
[24] J. Lu, E. Zhang, Gait recognition for human identification based on ica and fuzzy svm through multiple views fusion, Pattern Recognit. Lett. 28 (16) (2007) 2401–2411.
[25] J. Lu, V.E. Liong, X. Zhou, J. Zhou, Learning compact binary face descriptor for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2041–2056.
[26] J. Lu, Y.-P. Tan, G. Wang, Discriminative multimanifold analysis for face recognition from a single training sample per person, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 39–51.
[27] J. Lu, V.E. Liong, J. Zhou, Cost-sensitive local binary feature learning for facial age estimation, IEEE Trans. Image Process. 24 (12) (2015) 5356–5368.
[28] J. Lu, V.E. Liong, G. Wang, P. Moulin, Joint feature learning for face recognition, IEEE Trans. Inf. Forens. Secur. 10 (7) (2015) 1371–1383.
[29] J. Lu, G. Wang, W. Deng, K. Jia, Reconstruction-based metric learning for unconstrained face verification, IEEE Trans. Inf. Forens. Secur. 10 (1) (2015) 79–89.
[30] J. Lu, Y.-P. Tan, Cost-sensitive subspace analysis and extensions for face recognition, IEEE Trans. Inf. Forens. Secur. 8 (3) (2013) 510–519.
[31] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, J. Zhou, Cost-sensitive semi-supervised discriminant analysis for face recognition, IEEE Trans. Inf. Forens. Secur. 7 (3) (2012) 944–953.
[32] J. Lu, V.E. Liong, J. Zhou, Simultaneous local binary feature learning and encoding for face recognition, in: 2015 IEEE International Conference on Computer Vision, 2015, pp. 3721–3729.
[33] J. Lu, G. Wang, W. Deng, P. Moulin, J. Zhou, Multi-manifold deep metric learning for image set classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1137–1145.
[34] V.E. Liong, J. Lu, G. Wang, P. Moulin, J. Zhou, Deep hashing for compact binary codes learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2475–2483.
[35] J. Lu, G. Wang, P. Moulin, Localized multifeature metric learning for image-set-based face recognition, IEEE Trans. Circ. Syst. Video Technol. 26 (3) (2016) 529–540.
[36] K. Weinberger, J. Blitzer, L. Saul, Distance metric learning for large margin nearest neighbor classification, Advances in Neural Information Processing Systems, 2005.
[37] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: International Conference on Machine Learning, 2007, pp. 209–216.
[38] H.V. Nguyen, L. Bai, Cosine similarity metric learning for face verification, in: Asian Conference on Computer Vision, 2010, pp. 709–720.

[39] E. Xing, A. Ng, M. Jordan, S. Russell, Distance metric learning, with application to clustering with side-information, in: Advances in Neural Information Processing Systems, 2002, pp. 505–512.
[40] J. Goldberger, S. Roweis, G. Hinton, R. Salakhutdinov, Neighborhood component analysis, in: Advances in Neural Information Processing Systems, 2004, pp. 2539–2544.
[41] R. Cinbis, J. Verbeek, C. Schmid, Unsupervised metric learning for face identification in tv video, in: IEEE Interntaional Conference on Computer Vision, 2011, pp. 1559–1566.
[42] J. Chen, Z. Chen, Z. Chi, H. Fu, Facial expression recognition in video with multiple feature fusion, IEEE Trans. Affect Comput. (2016).
[43] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[44] A. Sharma, A. Kumar, H. Daume III, D. Jacobs, Generalized multiview analysis: a discriminative latent space, in: IEEE International Conference Computer Vision and Pattern Recognition, 2012, pp. 1867–1875.
[45] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, J. Zhou, Neighborhood repulsed metric learning for kinship verification, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2) (2014) 331–345.
[46] H. Yan, J. Lu, W. Deng, X. Zhou, Discriminative multimetric learning for kinship verification, IEEE Trans. Inf. Forens. Secur. 9 (7) (2014) 1169–1178.

Haibin Yan received the B.Eng. and M.Eng. degrees from the Xi'an University of Technology, Xi'an, China, in 2004 and 2007, and the Ph.D. degree from the National University of Singapore, Singapore, in 2013, all in mechanical engineering. Now, she is an Assistant Professor in the School of Automation, Beijing University of Posts and Telecommunications, Beijing, China. From October 2013 to July 2015, she was a research fellow at the Department of Mechanical Engineering, National University of Singapore, Singapore. Her research interests include robotics and computer vision.