Signal Processing 169 (2020) 107370

Contents lists available at ScienceDirect

Signal Processing

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

Low-resolution facial expression recognition: A filter learning perspective

Yan Yan[a], Zizhao Zhang[b], Si Chen[c], Hanzi Wang[a,*]

[a] *Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Fujian, China*
[b] *Department of Computer Information and Science Engineering, University of Florida, Florida, USA*
[c] *School of Computer and Information Engineering, Xiamen University of Technology, Fujian, China*

ARTICLE INFO

Article history:
Received 23 April 2019
Revised 28 October 2019
Accepted 9 November 2019
Available online 12 November 2019

Keywords:
Filter design
Subspace learning
Image representation
Low-resolution
Facial expression recognition

ABSTRACT

Automatic facial expression recognition has attracted increasing attention for a variety of applications. However, the problem of low-resolution generally causes the performance degradation of facial expression recognition methods under real-life environments. In this paper, we propose to perform low-resolution facial expression recognition from the filter learning perspective. More specifically, a novel image filter based subspace learning (IFSL) method is developed to derive an effective facial image representation. The proposed IFSL method mainly includes three steps: Firstly, we embed the image filter learning into the optimization process of linear discriminant analysis (LDA). By optimizing the cost function of LDA, a set of discriminative image filters (DIFs) corresponding to different facial expressions is learned. Secondly, the images filtered by the learned DIFs are added together to generate the combined images. Finally, a regression learning technique is leveraged for subspace learning, where an expression-aware transformation matrix is obtained using the combined images. Based on the transformation matrix, IFSL effectively removes irrelevant information while preserving useful information in the facial images. Experimental results on several facial expression datasets, including CK+, MMI, JAFFE, SFEW and RAF-DB, show the superior performance of the proposed IFSL method for low-resolution facial expression recognition, compared with several state-of-the-art methods.

1. Introduction

During the past few decades, automatic facial expression recognition has attracted extensive attention in computer vision and pattern recognition. It plays an important role in a variety of applications, including human computer interaction (HCI), data-driven animation [1–5]. Despite significant progress, facial expression recognition is still a difficult task, due to the variations caused by pose, illumination, and so on. One critical problem that is not well solved is low-resolution (LR). In real-life environments, the resolution of facial images captured by an ordinary camera may be low. The LR facial images usually lack sufficient visual information to extract informative features, thus leading to the performance degradation of facial expression recognition methods [6]. Therefore, effectively distinguishing different facial expressions based on the LR facial images is very challenging but meaningful for practical tasks.

To deal with the problem of LR facial expression recognition, existing methods mainly focus on two aspects: (1) face super-resolution (SR) [7,8], and (2) facial image representation [6,9,10]. The first aspect usually adopts two criteria to perform SR: pixel-level visual fidelity and image-level face identity preservation. The second aspect aims to extract the compact and discriminative feature representation. In this paper, we mainly study the second aspect.

SR methods aim to construct a high-resolution (HR) image from the corresponding LR image [11]. Theoretically, by applying the SR methods on the LR facial images to construct HR images, the reconstructed images can be used for facial expression recognition. However, the computational complexity of existing SR methods is usually high and these SR methods cannot guarantee that the resulting HR facial images are optimal for recognition [12].

Facial image representation also plays a critical role for LR facial expression recognition. The methods of representing facial expression images can be roughly categorized into geometric feature-based methods [13] and appearance-based methods [14]. Geometric feature-based methods detect salient facial feature points and characterize the variations of these points, which can achieve good

[*] Corresponding author.
E-mail addresses: yanyan@xmu.edu.cn (Y. Yan), zizhao@cise.ufl.edu (Z. Zhang), chensi@xmut.edu.cn (S. Chen), hanzi.wang@xmu.edu.cn (H. Wang).

https://doi.org/10.1016/j.sigpro.2019.107370
0165-1684/© 2019 Elsevier B.V. All rights reserved.

performance on action unit recognition [15]. However, precise localization of facial feature points is not a trivial task for LR facial images. Appearance-based methods represent the variations of facial appearance based on the whole face or specific regions in a facial image. This kind of methods usually attempts to extract discriminative features in facial images to distinguish different facial expressions.

The appearance-based methods can be further classified into handcrafted feature-based methods [16–18] and feature learning-based methods [19–21]. Representative handcrafted feature-based methods include local binary patterns (LBP) [16], Haar-like features [17] and Gabor-wavelet features [18]. However, these manually designed features may not effectively handle the challenges caused by the non-linear facial appearance variations due to different poses, occlusions and etc. More recently, feature learning-based methods, such as auto-encoder [19] and convolutional neural networks (CNNs) [20,21], have attracted much attention. Zhang et al. [19] present a spatially coherent feature-learning method for pose-invariant facial expression recognition. They combine the learning-based features and the corresponding geometry features to construct robust features. Xie and Hu [20] propose a deep comprehensive multipatches aggregation CNNs method, which consists of two CNN branches to respectively extract the holistic features and local features, to solve the problem of facial expression recognition. Li et al. [21] present the CNN with an attention mechanism (ACNN) for facial expression recognition in the presence of occlusions. These methods show the excellent ability to extract the discriminative representation from the raw data.

Psychologists have shown that the crucial features for recognizing facial expressions are usually distributed around salient facial feature points, such as mouth, nose and eyes [1]. The variations of the salient facial feature points are the useful information for facial expression recognition. In contrast, the facial identities of different persons are the irrelevant information, which should be suppressed or removed (although such information is important for identifying a person) for the task of facial expression recognition. In particular, the information in the LR facial images is relatively limited. Therefore, the irrelevant information may significantly decrease the performance of LR facial expression recognition. Therefore, how to extract the discriminative facial image representation from the limited information is critical.

In this paper, we propose to perform LR facial expression recognition from the filter learning perspective, where a novel and effective facial image representation is developed for facial expression recognition. The process of constructing the facial image representation for LR facial expression recognition can be considered as the process of suppressing irrelevant information (e.g., facial identity differences), while enhancing the valuable information (e.g., wrinkled eyebrow, smiling mouth and other features that are critical for discriminating different expressions) in facial images.

More specifically, we propose a novel image filter based subspace learning (IFSL) method to achieve an effective image representation for LR facial expression recognition. In particular, we learn a discriminative image filter (DIF), based on the two-class linear discriminant analysis (LDA) technique [22], to discriminate a non-neural expression from the neutral expression. The learned DIF extracts discriminative information by mapping the facial images to a subspace where the intra-class variations are minimized and the inter-class variations are maximized. Therefore, the DIF is able to find subtle variations of facial expression among different LR facial images. When a set of learned DIFs (corresponding to different expressions) is applied to a multi-class classification task (e.g., facial expression recognition in this paper), we propose to use a regression learning technique (i.e., the linear ridge regression (LRR) technique [23]) to derive a new facial image representation with high discriminability, based on the filtered images rather than the original images. As a result, an expression-aware transformation matrix that encodes the expression information is obtained. This strategy extends the classification ability of the DIF from two-class to multi-class classification.

In summary, we present a novel image representation method by taking advantage of the discriminative image filter and the regression learning technique.

The preliminary versions of this work were reported in [24,25]. However, we have made several significant extensions compared with the previous versions. The new contributions include:

- We provide a general formulation of the image filter learning, where the image filter can take different forms (such as element-wise product, linear transform and convolution) as long as it is differentiable.
- We reformulate the original method and develop a more general framework for image filter based subspace learning. We also offer more mathematical details and motivations of the proposed method for facial expression recognition.
- We conduct extensive experiments on both in-the-lab datasets and in-the-wild datasets to demonstrate the effectiveness of the proposed method for LR facial expression recognition.

The remainder of this paper is organized as follows. In Section 2, we review related work. In Section 3, we introduce the details of the proposed IFSL method. In Section 4, we evaluate the performance of IFSL and compare IFSL with several state-of-the-art methods on various facial expression recognition datasets. Finally, the conclusion is drawn in Section 5.

2. Related work

This section reviews related work in LR facial expression recognition. Firstly, the recently developed methods on face super-resolution and LR recognition are introduced in Section 2.1. Then, some facial image representation methods are reviewed in Section 2.2. Finally, some filter learning methods are discussed in Section 2.3.

2.1. Face super-resolution and low-resolution recognition

Traditional methods for handling the LR facial images aim to generate high-resolution (HR) facial images, based on which facial expression recognition can be performed. These methods can be roughly classified into two categories: generic super-resolution (SR) methods [7], and class-specific SR methods (also called face hallucination) [8].

Generic SR methods take advantage of the priors that ubiquitously exist in natural images without relying on the face class information. For instance, Gu et al. [26] develop a convolutional sparse coding method for image SR instead of the conventional patch-based methods. Dong et al. [7] use the CNNs to learn a non-linear mapping function between LR and HR images based on a large-scale image dataset.

On the other hand, face hallucination methods exploit the statistical information of faces and they usually achieve better results than generic SR methods for facial expression recognition or face recognition. For example, Ma et al. [27] use multiple local constraints learned from exemplar facial images to perform face hallucination. Wang et al. [28] propose to apply the global constraints between LR and HR facial images, and then hallucinate HR facial images based on the eigen-transformation. However, the output HR facial images may suffer from ghosting artifacts. Note that, generative adversarial networks (GANs) [29] can generate facial images with sharp details due to the discriminative networks. However, the generated images are only similar to one another in the class domain but they are different in the appearance domain [30]. In general, the computational complexity of these face hallucination methods is usually high.

Except for the above SR methods, some methods have been specifically developed for LR facial expression recognition/face recognition. These methods aim to extract resolution-insensitive features [6] or learn cross-resolution transformations [31–34].

For example, Khan et al. [6] propose a novel feature descriptor PLBP (Pyramid of LBP) for LR facial expression recognition. In fact, LR face recognition has achieved significant progress in the past few years. Ren et al. [31] propose a coupled kernel embedding (CKE) method for feature extraction with its application to LR face recognition. Jiang et al. [32] develop a coupled discriminant multi-manifold analysis (CDMMA) for LR face recognition. By exploiting the neighborhood information and local geometric structure of the manifold, CDMMA learns two mappings to project LR and HR images to a unified discriminative feature space. Xing and Wang [33] develop couple manifold discriminant analysis with bipartite graph embedding (CMDA_BGE) to solve the problem of LR face recognition. Chu et al. [34] propose a cluster-based regularized simultaneous discriminant analysis (C-RSDA) method for LR face recognition with single sample per person. Note that these LR face recognition methods usually match the LR probe faces against the HR gallery images. In this paper, we concentrate on the more general settings, where both the training and test images are LR.

2.2. Facial image representation

Automatic facial expression recognition usually consists of two main steps: feature extraction and facial expression classification [1]. Feature extraction extracts generative or discriminative representations from raw facial images to effectively represent the facial images. Generally, current methods for representing facial expression images can be divided into geometric feature-based methods [13] and appearance-based methods [9,16,17,35]. In this paper, we mainly review the appearance-based methods.

The representative appearance-based methods, including local binary patterns (LBP) [6,16,36,37], Haar-like features [17] and Gabor-wavelet features [18], have been successfully applied into facial expression analysis. Specifically, several LBP-based variants, such as m-LBP (representing salient micro-patterns in facial images) [36] and Boost-LBP (using a boosting algorithm to learn the most discriminative LBP histograms) [37], are proposed and achieve the state-of-the-art performance. Classical subspace learning methods, such as linear discriminant analysis (LDA) [38] and principle component analysis (PCA) [39], are also widely used for feature extraction.

The above methods consider the facial image as a whole without specifically emphasizing the important role of salient facial feature points. Actually, some local facial regions (e.g., eyes, eyebrows and mouths) contain critical information for expression recognition, since different expressions accompany the variations in different local facial regions. In recent years, some methods [9,35] have been proposed to analyze non-holistic facial images. For instance, Zhong et al. [35] propose a multi-task sparse learning framework to explore discriminative information in local facial regions for differentiating different expressions, and suggest that different local facial regions should be assigned with different weights. They use LBP to partition the facial images into isometric non-overlapping regions, where the relationship among different local facial regions is exploited. Experimental results in [35] show that the most important local facial regions for recognizing the expressions are the eyes, eyebrows, nose and mouths.

Recently, deep learning has achieved outstanding performance in a variety of computer vision tasks, including facial expression recognition [20,21,40–44]. For example, Xie et al. [40] propose the deep attentive multi-path CNN (DAM-CNN) method, which not only automatically locates the expression-related regions, but also generates an effective facial expression representation. Li and Deng [41] develop a deep locality-preserving CNN (DLP-CNN) method for unconstrained facial expression recognition, which uses a new locality-preserving loss layer for deep learning. Moreover, they introduce a new real-world facial expression dataset (i.e., RAF-DB) captured from the Internet. In [42], a deep emotion-conditional adaption network (ECAN) method for unsupervised cross-dataset facial expression recognition is developed.

Recently, the video-based facial expression recognition has received great interest. Compared with a static image, a video sequence not only contains the spatial appearance, but also provides facial motions. Gupta et al. [45] develop a scale invariant architecture for generating illumination invariant deep motion features for video-based facial expression recognition. Alam et al. [46] propose a biologically inspired sparse-deep simultaneous current network (S-DSRN) for robust facial expression recognition. S-DSRN makes use of the weight sharing technique in the hidden recurrent layers to reduce the number of network parameters, where the simultaneous recurrency offers efficient control over the depth of the model. Chen et al. [47] combine a new feature descriptor called histogram of oriented gradients from three orthogonal planes (HOG-TOP) and a new geometric feature descriptor to respectively extract dynamic textures and facial configuration changes for video-based facial expression recognition. Moreover, the audio modality is also considered for recognition.

2.3. Filter learning

Filter learning-based methods are widely applied to many computer vision tasks, such as face recognition [48,49], visual tracking [50], object detection [51], due to their high generalization ability and robustness. For example, Yan et al. [49] propose an effective correlation filter bank method to extract features for face recognition. Henriques et al. [50] propose to use the kernel correlation filter method for fast visual tracking. Generally speaking, the filter is designed to suppress noisy signals and amplify useful signals so that the discriminability of filtered signals can be enhanced.

It is worthy pointing out that Gabor-wavelet [18] and LBP [16,36] can also be considered as the special forms of filter. However, these filters are manually designed without using the learning technique. In contrast, our proposed discriminative image filter is learned via the objective function of LDA with maximizing discriminability. In addition, the recent popular CNNs-based methods [3,52] use the convolutional filters to obtain rich representations for accurate facial expression recognition and these methods have achieved superior performance. The parameters of the convolutional filters can be effectively learned based on the back-propagation technique. Nevertheless, the CNNs-based methods require the relatively HR images as the input to extract hierarchical representations, which can make these methods difficult to handle the LR facial expression recognition problem [52]. Note that these CNNs-based methods usually require a large amount of training data. But current facial expression datasets typically contain a small number of labelled samples. Therefore, cross-corpus training [42] or transfer learning techniques [53] (which take advantage of the extra available training data) can be used to effectively deal with facial expression recognition with limited training data. In this paper, we aim to solve the problem of limited training data from the image filtering perspective. The proposed method is a good alternative for dealing with LR facial expression recognition using limited training data. Experimental results on multiple public facial expression datasets verify the excellent performance of the proposed method for LR facial expression recognition.

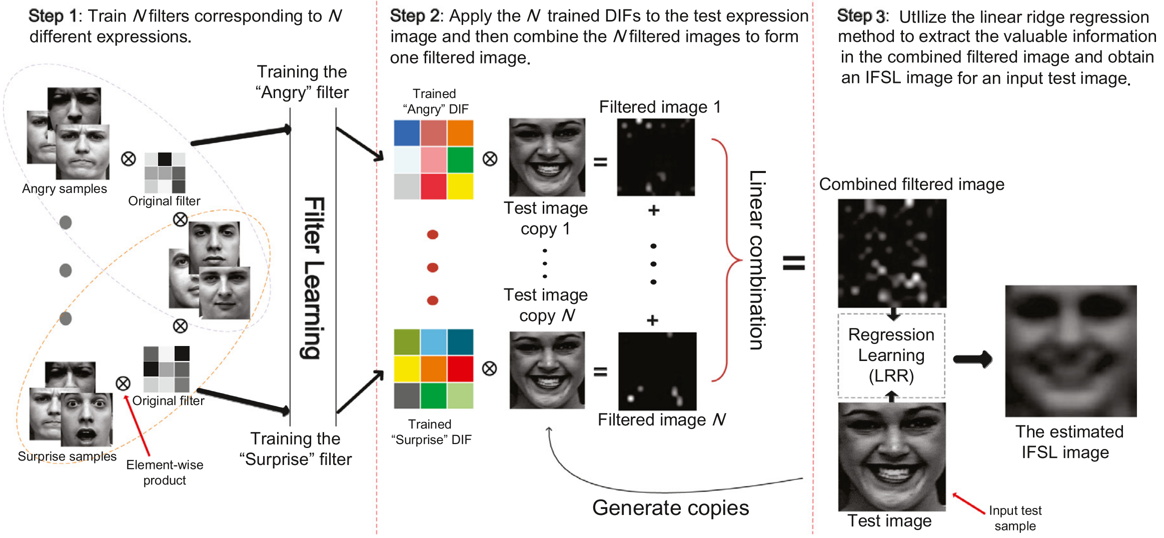
Step 1: Train N filters corresponding to N different expressions.
Step 2: Apply the N trained DIFs to the test expression image and then combine the N filtered images to form one filtered image.
Step 3: Utilize the linear ridge regression method to extract the valuable information in the combined filtered image and obtain an IFSL image for an input test image.
Training the "Angry" filter
Angry samples
Original filter
Filter Learning
Original filter
Surprise samples
Element-wise product
Training the "Surprise" filter
Trained "Angry" DIF
Test image copy 1
Filtered image 1
Test image copy N
Trained "Surprise" DIF
Filtered image N
Linear combination
Generate copies
Combined filtered image
Regression Learning (LRR)
Test image
Input test sample
The estimated IFSL image

Fig. 1. The framework of the proposed IFSL method.

3. Image filter based subspace learning

The proposed image filter based subspace learning (IFSL) method contains three steps. In the first step, we embed the process of discriminative image filter (DIF) learning into the optimization of two-class LDA. In the second step, we linearly combine the filtered images generated by the learned DIFs. In the third step, based on the combined results, we propose to use a regression learning approach to perform feature extraction. The overall framework of the proposed IFSL is shown in Fig. 1.

The discriminative image filter learning is introduced in Sections 3.1, and the details of the optimization procedure are described in Sections 3.2. The details of the second and the third steps are discussed in Section 3.3. The complete algorithm is given in Section 3.4.

3.1. Discriminative image filter learning

For the task of facial expression recognition, the significant discriminative information mainly lies in the local facial regions such as the eyes, eyebrows, nose and mouths. These local facial regions have different influence on recognizing different expressions (e.g., lips rise in a happy expression face; eyebrows wrinkle in an angry expression face; a mouth widely opens and eyebrows rise in a surprise expression face). The local facial regions like eyebrows, mouths and ajina contain more discriminative information than the regions like cheeks to identify the angry expression. In other words, these regions play an important role in discriminating different expressions. Therefore, the objective of an image filter is to simultaneously emphasize discriminative information in the crucial local facial regions while suppressing irrelevant information in the other facial regions.

A variety of filter functions can be used. For example, $\boldsymbol{t}$

- Element-wise product: $f(\boldsymbol{\lambda}_1, \boldsymbol{p}) = \boldsymbol{\lambda}_1 \otimes \boldsymbol{p}$, where $\otimes$ represents the dot product operator; $\boldsymbol{\lambda}_1$ is the filter function in $\mathbb{R}^d$; $\boldsymbol{p}$ is a facial image represented as a d-dimensional column vector. Therefore, each element λ_{1i} decides the intensity of the ith pixel p_i in a facial image that passes through.
- Linear transform: $f(\boldsymbol{\lambda}_2, \boldsymbol{p}) = \boldsymbol{\lambda}_2 \boldsymbol{p}$, where $\boldsymbol{\lambda}_2 \in \mathbb{R}^{d\times d}$ is the transformation matrix.
- Convolution: $f(\boldsymbol{\lambda}_3, \boldsymbol{p}) = \boldsymbol{\lambda}_3 * \boldsymbol{p}$, where $*$ denotes the convolution operator and $\boldsymbol{\lambda}_3 \in \mathbb{R}^d$ is the convolutional kernel.

Given a filter function $\boldsymbol{\lambda}$ (defined as one of the above-mentioned three functions) and an input matrix $\mathbf{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n] \in \mathbb{R}^{d\times n}$ consisting of n facial images, we can obtain the output matrix as,

$$\begin{aligned} f(\boldsymbol{\lambda}, \mathbf{P}) &= [f(\boldsymbol{\lambda}, \boldsymbol{p}_1), \ldots, f(\boldsymbol{\lambda}, \boldsymbol{p}_n)] \\ &= [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n], \end{aligned} \tag{1}$$

and we define $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$. Here, $\mathbf{X} \in \mathbb{R}^{d\times n}$ contains n filtered facial images, and each filtered facial image is a d-dimensional column vector.

Generally speaking, we expect that the learned image filter has the discriminative ability to extract the useful information for facial expression recognition. In other words, the filtered images corresponding to different expressions should be more separable for subsequent classification. Therefore, in order to learn a discriminative image filter (DIF), we propose to take advantage of linear discriminant analysis (LDA) during the training process.

LDA [22] is a popular subspace analysis method which projects high-dimensional samples to an optimal discriminative subspace, where the projected samples are well-separated. It can effectively extract the information from samples and compress the dimensionality of samples through a supervised learning strategy. LDA is originally proposed to handle two-class classification problems. In fact, LDA can also be extended to handle multi-class classification problems (where the inter-class matrix is the sum of the pairwise scatter matrix of any two different classes). However, multi-class LDA suffers from the problem of unbalanced pairwise distances (i.e., the distance of two different classes may be much smaller or larger than that between another two different classes), which may significantly degrade the performance in facial expression recognition [54,55]. Therefore, we mainly focus on two-class LDA in this paper.

Next we give the detailed steps of embedding the DIF learning into the optimization process of LDA. LDA attempts to seek for an optimal linear transformation to minimize the intra-class variance (characterized by an intra-class covariance $\mathbf{S}_W$) as well as to maximize the inter-class variance (characterized by an inter-class covariance $\mathbf{S}_B$). In our method, we use the facial images with a neutral expression, denoted as $\mathbf{P}_1$, and those with a non-neutral expression (e.g., angry, disgust, fear, happy, surprise, or sad), denoted as $\mathbf{P}_2$, as the inputs for training the two-class LDA model.

The cost function of two-class LDA is defined as,

$$L(\mathbf{X}_1, \mathbf{X}_2) = \frac{\boldsymbol{\omega}^T \mathbf{S}_B(\mathbf{X}_1, \mathbf{X}_2)\boldsymbol{\omega}}{\boldsymbol{\omega}^T \mathbf{S}_W(\mathbf{X}_1, \mathbf{X}_2)\boldsymbol{\omega}}, \tag{2}$$

where $\mathbf{X}_i = f(\boldsymbol{\lambda}, \mathbf{P}_i),\ i = 1,\ 2.\ \boldsymbol{\omega} \in \mathbb{R}^d$ is the linear transformation vector. The inter-class covariance $\mathbf{S}_B$ is defined as:

$$\mathbf{S}_B(\mathbf{X}_1, \mathbf{X}_2) = (\boldsymbol{m}_1 - \boldsymbol{m}_2)(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T, \tag{3}$$

and the intra-class covariance $\mathbf{S}_W$ is defined as:

$$\mathbf{S}_W(\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{X}_1 - \mathbf{M}_1)(\mathbf{X}_1 - \mathbf{M}_1)^T + (\mathbf{X}_2 - \mathbf{M}_2)(\mathbf{X}_2 - \mathbf{M}_2)^T, \tag{4}$$

where the column vector $\boldsymbol{m}_i$ is the mean of $\mathbf{X}_i$ ($i = 1,\ 2$) in $\mathbb{R}^d$. The matrix $\mathbf{M}_i$ includes n copies of $\boldsymbol{m}_i$.

The optimal $\boldsymbol{\omega}^*$ can then be computed as [23],

$$\begin{aligned}\boldsymbol{\omega}^* &= \underset{\boldsymbol{\omega}}{\operatorname{argmax}} \frac{\boldsymbol{\omega}^T \mathbf{S}_B(\mathbf{X}_1, \mathbf{X}_2)\boldsymbol{\omega}}{\boldsymbol{\omega}^T \mathbf{S}_W(\mathbf{X}_1, \mathbf{X}_2)\boldsymbol{\omega}} \\ &= \mathbf{S}_W(\mathbf{X}_1, \mathbf{X}_2)^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1).\end{aligned} \tag{5}$$

In two-class LDA, the linear transformation vector $\boldsymbol{\omega}$ has a closed-form formulation.

During the training process, the objective of the DIF learning is to obtain an optimal DIF $\boldsymbol{\lambda}$ embedded in the cost function, and this problem can be solved based on the gradient descent (which will be discussed in the following subsection). More specifically, by incorporating a DIF into the cost function of LDA, we can obtain the following objective function, i.e.,

$$\begin{aligned}O(\boldsymbol{\lambda}) &= -\ln\, L(\mathbf{X}_1, \mathbf{X}_2) + \frac{1}{2}Ctr(\boldsymbol{\lambda}^T\boldsymbol{\lambda}) \\ &= -\ln\, L\big(f(\boldsymbol{\lambda}, \mathbf{P}_1), f(\boldsymbol{\lambda}, \mathbf{P}_2)\big) + \frac{1}{2}Ctr(\boldsymbol{\lambda}^T\boldsymbol{\lambda}),\end{aligned} \tag{6}$$

where $tr(\boldsymbol{\lambda}^T\boldsymbol{\lambda})$ is a regularization term which enhances the generalization capability and robustness of the learned filter. $C\ (\geq 0)$ is a scalar parameter. Therefore, we aim to obtain the optimal DIF, such that

$$\boldsymbol{\lambda}^* = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}}\, O(\boldsymbol{\lambda}). \tag{7}$$

3.2. Discriminative image filter optimization

The minimization problem in Eq. (7) can be solved via the gradient descent technique [56], since both f and L in $O(\boldsymbol{\lambda})$ are differentiable. It is worthy pointing out that in each iteration, the calculation of $\boldsymbol{\omega}^*$ shown in Eq. (5) is dynamically updated for computing $\partial\boldsymbol{\omega}^*/\partial\boldsymbol{\lambda}$ in L. The derivation details for optimizing $O(\boldsymbol{\lambda})$ are shown as follows.

The partial derivative of $O(\boldsymbol{\lambda})$ with respect to λ_j (or $\lambda_{i,j}$ if $\boldsymbol{\lambda}$ is the transformation matrix) is computed. In the following, we use λ_j for simplification without loss of generality. Thus, the partial derivative can be written as:

$$\frac{\partial O(\boldsymbol{\lambda})}{\partial \lambda_j} = -\frac{\frac{\partial}{\partial \lambda_j} L(f(\boldsymbol{\lambda}, \mathbf{P}_1), f(\boldsymbol{\lambda}, \mathbf{P}_2))}{L\big(F(\boldsymbol{\lambda}, \mathbf{P}_1), f(\boldsymbol{\lambda}, \mathbf{P}_2)\big)} + C\lambda_j. \tag{8}$$

The cost function of LDA is differentiable, and we have:

$$\frac{\partial L}{\partial \lambda_j} = \frac{\frac{\partial}{\partial \lambda_j}(\boldsymbol{\omega}^{*T}\mathbf{S}_B\boldsymbol{\omega}^*)}{\boldsymbol{\omega}^{*T}\mathbf{S}_W\boldsymbol{\omega}^*} - \frac{\boldsymbol{\omega}^{*T}\mathbf{S}_B\hat{\boldsymbol{\omega}}^*}{(\boldsymbol{\omega}^{*T}\mathbf{S}_W\boldsymbol{\omega}^*)^2}\frac{\partial}{\partial \lambda_j}(\boldsymbol{\omega}^{*T}\mathbf{S}_W\boldsymbol{\omega}^*), \tag{9}$$

where

$$\frac{\partial}{\partial \lambda_j}(\boldsymbol{\omega}^{*T}\mathbf{S}_W\boldsymbol{\omega}) = \left(\frac{\partial \boldsymbol{\omega}^*}{\partial \lambda_j}\right)^T(\mathbf{S}_W\boldsymbol{\omega}^*) + \boldsymbol{\omega}^{*T}\left(\frac{\partial \mathbf{S}_W}{\partial \lambda_j}\boldsymbol{\omega}^* + \mathbf{S}_W\frac{\partial \boldsymbol{\omega}^*}{\partial \lambda_j}\right), \tag{10}$$

and the derivation of $\partial(\boldsymbol{\omega}^{*T}\mathbf{S}_B\boldsymbol{\omega}^*)/\partial\lambda_j$ is similar to the right item of Eq. (10).

The partial derivative of $\boldsymbol{\omega}^*$ with respect to λ_j is calculated in each iteration:

$$\begin{aligned}\frac{\partial \boldsymbol{\omega}^*}{\partial \lambda_j} &= \frac{\partial}{\partial \lambda_j}\big(\mathbf{S}_W^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1)\big) \\ &= -\mathbf{S}_W^{-1}\left(\frac{\partial}{\partial \lambda_j}\mathbf{S}_W\right)\mathbf{S}_W^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1) + \mathbf{S}_W^{-1}\left(\frac{\partial}{\partial \lambda_j}(\boldsymbol{m}_2 - \boldsymbol{m}_1)\right).\end{aligned} \tag{11}$$

According to Eq. (3) and Eq. (4), we can compute $\partial\mathbf{S}_W/\partial\lambda_j$ as follows:

$$\begin{aligned}\frac{\partial \mathbf{S}_W}{\partial \lambda_j} &= \frac{\partial}{\partial \lambda_j}\big((\mathbf{X}_1 - \mathbf{M}_1)(\mathbf{X}_1 - \mathbf{M}_1)^T\big) \\ &\quad + \frac{\partial}{\partial \lambda_j}\big((\mathbf{X}_2 - \mathbf{M}_2)(\mathbf{X}_2 - \mathbf{M}_2)^T\big),\end{aligned} \tag{12}$$

and similarly, $\partial\mathbf{S}_B/\partial\lambda_j$ is computed as follows:

$$\frac{\partial \mathbf{S}_B}{\partial \lambda_j} = \frac{\partial}{\partial \lambda_j}\big((\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T\big). \tag{13}$$

Finally, the partial derivative of $\mathbf{X}$ with respect to λ_j is obtained by computing the partial derivative of each column of X with respect to λ_j, that is,

$$\frac{\partial \mathbf{X}}{\partial \lambda_j} = \frac{\partial f(\boldsymbol{\lambda}, \mathbf{P})}{\partial \lambda_j} = \left[\frac{\partial f(\boldsymbol{\lambda}, \boldsymbol{p}_1)}{\partial \lambda_j}, \ldots, \frac{\partial f(\boldsymbol{\lambda}, \boldsymbol{p}_n)}{\partial \lambda_j}\right]. \tag{14}$$

The partial derivative of $\boldsymbol{f}$ with respect to λ_j can be derived according to different forms of filters. For example, for the element-wise product, $\partial f/\partial\lambda_j = \boldsymbol{s}_j \otimes \boldsymbol{p}$, where $\boldsymbol{s}_j$ is a vector where only the value of the jth entry is 1 and the rest are zero. For the linear transformation, $\partial f/\partial\lambda_{i,j} = \boldsymbol{E}_j\boldsymbol{p}$, where $\boldsymbol{E}_j$ is a $d \times d$ matrix consisting of all zeros except for the (i, j)th entry, which is 1.

It is worth pointing out that the optimization problem (i.e., Eq. (6)) of the proposed method is formulated by taking advantage of the Fisher criteria used in the conventional LDA method. However, the proposed method and LDA are significantly different. Firstly, LDA obtains the projection matrix with a closed-form solution. In contrast, the proposed method obtains the filters (can take the forms of element-wise product, linear transform or convolution) in an iterative manner. Secondly, the objective of LDA is to obtain the optimal projection matrix (i.e., w in Eq. (2)), while that of the proposed method tries to obtain the filters (i.e., λ in Eq. (2)). In other words, the proposed method is not equivalent to LDA even if the linear transform is used.

3.3. Linear combination and regression learning

In this paper, each DIF is designed to discriminate a specific (non-neural) facial expression from the neural expression, which is a two-class classification problem. Suppose that there are N different expressions, N DIFs (denoted as $\{\boldsymbol{\lambda}_i\}_{i=1}^N$) will be trained. Given an image $\boldsymbol{p}$ with an unknown expression, we aim to figure out the expression based on the outputs of DIFs. Here, we denote the DIF (corresponding to the expression label of the image $\boldsymbol{p}$) as $\boldsymbol{\lambda}_1^+$ and the other $(N-1)$ DIFs as $\{\boldsymbol{\lambda}_i^-\}_{i=2}^N$.

To identify the expression of a test image, one simple way is to firstly train N expression-dependent classifiers (using the one-vs-all strategy), where each classifier is trained to discriminate a non-neural expression from the other expressions based on the outputs of one DIF. Then, the facial expression corresponding to the classifier giving the highest probability output is identified.

However, such a way is not reliable since the correlation between different expressions is not considered and the recognition results may be inaccurate when two expressions share similar appearance variations. In other words, for a test image, the outputs of the six classifiers (corresponding to different image filters) are not discriminative enough for selecting the correct image filter. Besides, unbalanced data distribution may lead to the classifier overfitting to the majority class (note that the one-vs-all strategy is used).

In this paper, we solve the above problem by using the strategy of linear combination and regression learning. The steps are briefly given as follows. Firstly, a set of DIFs ($\{\boldsymbol{\lambda}_i\}_{i=1}^N$) is applied to the input image $\boldsymbol{p}$ so as to obtain the N filtered images, defined as $\{\boldsymbol{s}^i\}_{i=1}^N$. Note that only one filtered image (corresponding to the test expression) is enhanced while the other $N-1$ filtered images are suppressed. In other words, the enhanced image, generated by $f(\boldsymbol{\lambda}_1^+, \boldsymbol{p})$, contains useful information for classification, while the suppressed images, generated by $f(\boldsymbol{\lambda}_i^-, \boldsymbol{p}), i = 2, \dots, N$, contain irrelevant information for classification. All the filtered images are linearly combined to generate the combined image. Secondly, based on the observation that the correlation between the filtered image (i.e., $f(\boldsymbol{\lambda}_1^+, \boldsymbol{p})$) and the input image $\boldsymbol{p}$ is higher than those between the filtered images (i.e., $f(\boldsymbol{\lambda}_i^-, \boldsymbol{p}), i = 2, \dots, N$) and the input image $\boldsymbol{p}$, we propose to use a regression learning technique to yield an effective representation for the combined image (which generates a new image representation for the input image $\boldsymbol{p}$).

Linear Combination We first linearly combine $\{\boldsymbol{s}^i\}_{i=1}^N$ to generate a combined filtered images $\boldsymbol{s}$, and

$$\begin{aligned}\boldsymbol{s} &= \sum_{i=1}^{N} \boldsymbol{s}^i \\ &= f(\boldsymbol{\lambda}_1^+, \mathbf{p}) + \sum_{i=2}^{N} f(\boldsymbol{\lambda}_i^-, \mathbf{p}),\end{aligned} \tag{15}$$

Note that the weights corresponding to different filtered images are all set to 1. Therefore, given n training images $\mathbf{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \dots, \boldsymbol{p}_n] \in \mathbb{R}^{d\times n}$ containing different expression facial images, $N \times n$ filtered images are generated, which are linearly combined as follows:

$$\mathbf{S} = \sum_{i=1}^{N} f(\boldsymbol{\lambda}_i, \mathbf{P}), \tag{16}$$

where $\mathbf{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \dots, \boldsymbol{s}_n] \in \mathbb{R}^{d\times n}$ contains n linearly combined filtered images. Different from the DIF learning, the computation of $\mathbf{S}$ is general and not class-specific. Each $\boldsymbol{s}_i$ $(i = 1, \dots, n)$ consists of an enhanced filtered images (corresponding to the expression label of $\boldsymbol{p}_i$) and $(N-1)$ suppressed filtered images. The information in the suppressed filtered images can be considered as noises that should be removed without affecting the valuable expression information in the enhanced filtered images.

Regression learning To extract the useful information and remove the irrelevant information in $\mathbf{S}$, we propose to take advantage of the linear ridge regression (LRR) method, where we obtain a new representation for the input image. Mathematically, LRR solves the following optimization problem,

$$\min_{\mathbf{G}} \|\mathbf{P} - \mathbf{G}^T\mathbf{S}\|^2 + \beta\|\mathbf{G}^T\mathbf{I}\|^2, \tag{17}$$

where $\mathbf{I}$ is a diagonal matrix (which is usually the identity matrix); $\|\mathbf{G}^T\mathbf{I}\|^2$ is the regularization term, and β is the regularization parameter; $\mathbf{G}$ is a transformation matrix, which projects the combined filtered images in $\mathbf{S}$ onto a new space (i.e., generating new image representations); $\|\cdot\|$ denotes the Frobenius norm.

The closed-form solution of Eq. (17) can be written as:

$$\mathbf{G}^* = (\mathbf{S}\mathbf{S}^T + \beta\mathbf{I})^{-1}\mathbf{S}\mathbf{P}^T. \tag{18}$$

The optimal transformation matrix $\mathbf{G}^*$ is expression-aware, since it encodes expression information, which not only improves the capability of distinguishing different expressions, but also reduces the influence of facial identity differences.

Based on $\mathbf{G}^*$, the projected images are obtained as:

$$\mathbf{Y} = \mathbf{G}^{*T}\mathbf{S}, \tag{19}$$

where $\mathbf{Y}$ is defined as the IFSL images. Each projected image $\boldsymbol{y}_n$ in $\mathbf{Y}$ contains useful information for its corresponding expression label in $\boldsymbol{p}_n$, which can be used for classification. $\boldsymbol{y}_n$ and $\boldsymbol{p}_n$ have the same dimension.

Note that the least squares (LS) method is also a popular regression learning technique. However, compared with LRR used in the proposed method, LS encounters the following problems. Firstly, LS is effective only if the independent variables are not well-correlated. However, the characteristics of facial expressions are usually well-correlated [57], which can greatly affect the performance of LS. Secondly, the variance estimation of LS may be large when the number of samples used is small. Thus, the results obtained by LS become unreliable when a limited number of training samples are used. Thirdly, suppose that $\mathbf{S} \in \mathbb{R}^{d\times n}$ consists of n d-dimensional feature samples obtained by the above procedures. $\mathbf{S}\mathbf{S}^T$ becomes a singular matrix if $n < d$, and thus the results obtained by LS can be overfitted. In contrast, LRR [23] effectively solves these problems by adding a regularization term to balance the deviation [55]. Therefore, the useful information in $\mathbf{S}$ can be successfully preserved while irrelevant information is removed by using LRR.

The objective of regression learning step is to learn an expression-aware transformation matrix (i.e., $\mathbf{G}^*$). Based on $\mathbf{G}^*$, we are able to obtain a new subspace, where the information in the enhanced images is preserved while that in the suppressed images is removed. Such a way not only improves the capability of distinguishing different expressions, but also reduces the influence of facial identity differences. As a result, for an arbitrary image, we can obtain an image representation encoding effective expression information by projecting the combined filtered image onto the new subspace.

In summary, the advantages of combining linear combination and regression learning are two-fold. 1) We can effectively improve the discriminative capability for facial expression recognition by alleviating the influence of the distracting factors (such as facial identity). 2) For the test stage, we do not need to decide which image filter to be used. Instead, we combine the filtered outputs and project them onto a subspace to obtain the facial image representation by using $\mathbf{G}^*$.

3.4. The complete algorithm

In the previous subsections, we have developed all ingredients for the LR facial expression recognition method. Now we put them together to describe a complete algorithm for facial expression recognition.

The overall training stage of the proposed IFSL method is summarized in Algorithm 1, which returns a set of DIFs $\{\boldsymbol{\lambda}_i\}_{i=1}^N$, an expression-aware transformation matrix $\mathbf{G}^*$, and a classification model Φ. The test stage is straightforward. Specifically, given a test image, IFSL firstly uses a set of DIFs $\{\boldsymbol{\lambda}_i\}_{i=1}^N$ to generate N filtered images. After linearly combining these N filtered images, the transformation matrix $\mathbf{G}^*$ is used to obtain the corresponding IFSL image (i.e., a new image representation for the test image). The final classification result is performed by applying the trained model Φ.

Algorithm 1: The training stage of the proposed IFSL method.

```
Input: A set of training images $\mathbf{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n] \in \mathbb{R}^{d\times n}$,
with the neutral expression set $\mathbf{P}_n$ and $N$ non-neutral expression set $\{\mathbf{P}_i\}_{i=1}^N$;the maximum number of iterations $\tau$;
Output: $\{\boldsymbol{\lambda}_i\}_{i=1}^N$, $\mathbf{G}^*$, and $\Phi$.

for $i = 1, \ldots, N$ do
    Randomly initialize $\boldsymbol{\lambda}_i^{(0)}$;
    Select $\mathbf{P}_n$ and $\mathbf{P}_i$ as the inputs of Eq. (7);
    $t$ = 0;
    while $(t < \tau)$ Do
        Compute $\boldsymbol{\omega}^*$ according to Eq. (5);
        Compute $\frac{\partial O(\boldsymbol{\lambda}_i^{(t)})}{\partial \boldsymbol{\lambda}_i^{(t)}}$ following Eq. (8) to Eq. (14);
        Update $\boldsymbol{\lambda}_i^{(t)}$ using the conjugate gradient descent technique;
        $t = t + 1$;
    end while
    Obtain an optimal image filter $\boldsymbol{\lambda}_i$ corresponding to the $i$th expression;
end for
Combine the images filtered using the learned DIFs $\{\boldsymbol{\lambda}_i\}_{i=1}^N$ to obtain $\mathbf{S}$ by Eq. (16);
Compute an expression-aware transformation matrix $\mathbf{G}^*$ by Eq. (18);
Obtain the projected images $\mathbf{Y}$ according to Eq. (19);
Obtain a facial expression classifier $\Phi$ using $\mathbf{Y}$, and the corresponding labels using the training data.
```

3.5. Discussions

Firstly, the main advantage of the proposed IFSL method is that irrelevant or useless expression information can be significantly removed, while the useful information for LR facial expression recognition can be effectively preserved. The proposed IFSL method contains two key elements that contribute to the overall performance and effectiveness. (1) Image filtering. A discriminative image filter (DIF) is learned to distinguish a non-neural expression from the neural expression (by optimizing the cost function of LDA). The image filtered by the learned DIF contains the critical information for discriminating the non-neural expression. (2) Subspace learning. An expression-aware transformation matrix is learned to encode the expression information and remove the identity information (by using the linear ridge regression technique). Fig. 2 shows the visualization of different facial expression images and the corresponding IFSL images. We can see that the similarities of the IFSL images obtained by the proposed method are higher than those of the raw images for each of the six expressions. The irrelevant information (e.g., facial identity difference) is suppressed and the valuable expression information around facial keypoints is preserved in the obtained IFSL images. Therefore, the preserved information in the IFSL images bears high discriminability for facial expression recognition.

Secondly, the reasons why the proposed IFSL method can be applied to LR facial expression recognition are two-fold: (1) The LR facial images usually contain noises due to the variations caused by illumination, pose and degradation in resolution [58]. The learned DIFs can effectively remove noises while preserving useful information in LR facial images. (2) The proposed IFSL method is a holistic recognition method, which performs subspace learning based on the whole facial appearances of LR images. Compared with the local recognition methods, the holistic recognition methods are less sensitive to the image resolution [1–3].

Finally, it is worth noting that there are some potential problems, when the proposed method is applied to HR facial expression recognition. Firstly, it is difficult for the image filter to learn the good parameters for recognizing HR facial images, when the number of training samples is limited. This is because that the number of parameters of image filter is relatively large for HR facial images. Secondly, the proposed image filter becomes more sensitive to the misalignment problem for HR facial images. Thirdly, the proposed method suffers from high computational complexity if the sizes of images are large (see Section 4.3 for more details).

4. Experimental results

In this section, extensive experiments are conducted to evaluate the performance of the proposed IFSL method. In Section 4.1, we demonstrate the discriminability of the DIF on a synthetic dataset. In Section 4.2, we firstly introduce several popular facial expression datasets, and evaluate the performance of different filter functions. Moreover, we discuss the influence of different image sizes on the performance of IFSL. Then, we evaluate the performance of IFSL on the facial expression datasets, and compare IFSL with several

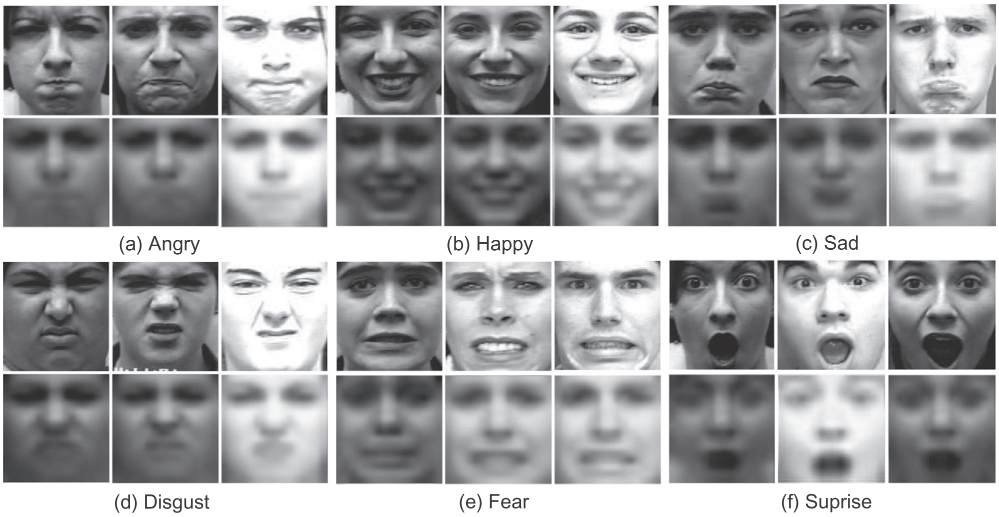
(a) Angry
(b) Happy
(c) Sad
(d) Disgust
(e) Fear
(f) Suprise

Fig. 2. Visualization of six types of facial expression images (1^{st} and 3^{rd} rows) and the corresponding IFSL images (2^{nd} and 4^{th} rows). Here, the element-wise product filter function is used.

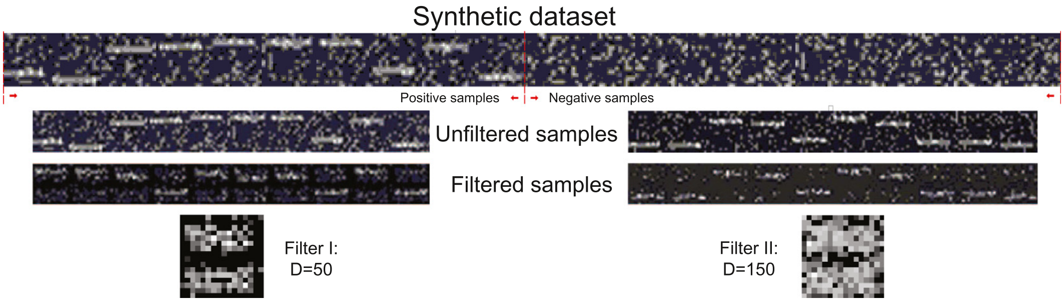

Fig. 3. Experimental results by applying the learned DIF to a synthetic dataset. The first row presents a randomly generated synthetic dataset, showing 10 positive samples in the left half part and 10 negative samples in the right half part. The second to the fourth rows respectively show the unfiltered images, the corresponding filtered images and the trained filters. Filter I is trained when D is set to 50 while filter II is trained when D is set to 150.

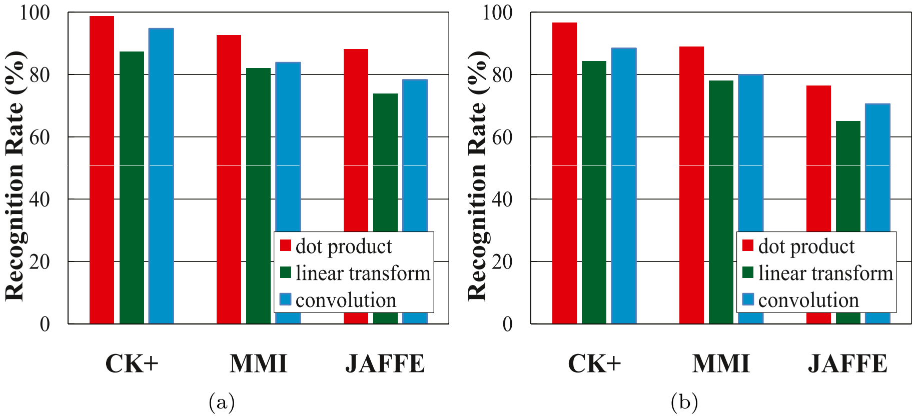

Fig. 4. Recognition results obtained by the proposed IFSL method with different filter functions on the three facial expression datasets. (a) The recognition rates obtained by IFSL with SVM (b) The recognition rates obtained by IFSL with k-NN.

state-of-the-art methods. In Section 4.3, we analyze the limitations of the proposed method.

To show the influence of a classifier on the proposed method, we use two classifiers (i.e., SVM and k-NN) for comparison: (1) The support vector machine (SVM) classifier has been proposed as one of the most popular classifiers to deal with the task of facial expression recognition [57]. SVM uses a kernel function to project samples to a high-dimensional space. Popular kernels include linear, polynomial, and radial basis functions (RBF). To avoid overfitting, we use the linear kernel in the following experiments. (2) The k-nearest neighbor (k-NN) classifier is regarded as the simplest instance-based classifier [23]. A sample is classified by a majority vote of its k nearest neighbors. We set the value of k to 3 in the following experiments.

4.1. *Experiments on a synthetic dataset*

In this experiment, we validate the discriminability of the proposed DIF (the element-wise product filter is employed) by using a synthetic dataset, where we visually demonstrate that the learned DIF can effectively extract discriminative information. We generate a synthetic dataset consisting of one positive class and one negative class. Note that here we do not use the linear combination step and the regression learning step (described in Section 3.3), since this is a two-class classification task.

More specifically, the synthetic dataset is comprised of D (the value of D is set to 50 and 150, respectively) synthetic samples, including $D/2$ positive samples and $D/2$ negative samples. The patch size of each synthetic sample is 16×16 (thus $d = 256$). For each positive sample, we generate a horizontal white line at a random position crossing from the left side to the right side, while such a line does not exist for the negative samples, as illustrated in Fig. 4. Both negative and positive samples are contaminated by randomly generated white noises. The DIF is then trained using all the samples. Therefore, the synthetic dataset is used to evaluate whether the learned DIF can suppress the useless information (i.e., white noises) in the positive samples while preserving the useful information (i.e., horizontal white lines).

Fig. 3 also shows the experimental results obtained by using the proposed DIF on the synthetic dataset. We can observe that the noises in the positive samples are successfully suppressed by the learned DIF and the white lines are well preserved. Moreover, the filtering performance obtained by DIF (when $D = 150$) is better than that obtained by DIF (when $D = 50$), since the noises in the filtered positive samples when $D = 150$ are much less than those when $D = 50$. Moreover, the positive horizontal lines are effectively preserved when $D = 150$. However, these lines are slightly suppressed when $D = 50$ because the limited number of samples is used. In general, the learned DIF can effectively extract useful information while at the same time filtering out irrelevant information for classification.

4.2. *Experiments on facial expression recognition*

In this section, we extensively demonstrate the performance of the proposed IFSL for facial expression recognition.

4.2.1. *Facial expression datasets*

We evaluate the performance of the proposed method on both in-the-lab facial expression datasets (including CK+, JAFFE and MMI) and in-the-wild facial expression datasets (including SFEW [62] and RAF-DB [42]). A brief introduction of these datasets is given as follows.

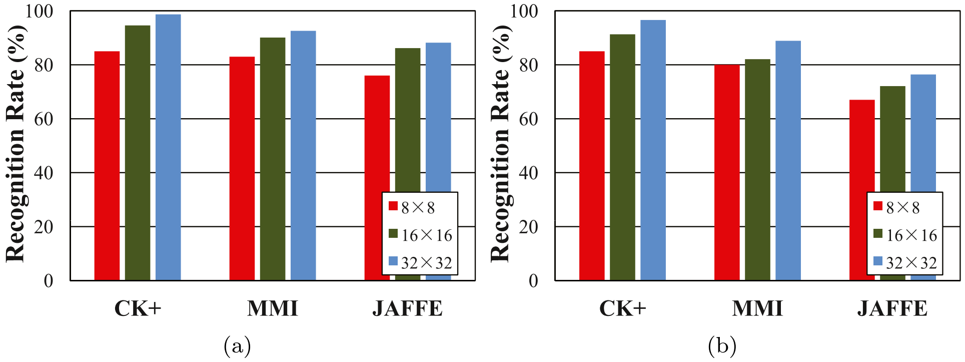

Fig. 5. Recognition results obtained by the proposed IFSL method with different image sizes on the three facial expression datasets. (a) The recognition rates obtained by IFSL with SVM (b) The recognition rates obtained by IFSL with *k*-NN.

The extended Cohn-Kanade (CK+) facial expression dataset[1], which is an extended version of the previous Cohn-Kanade (CK) dataset, consists of 593 short videos from 123 subjects with different ages under uniform illumination [59]. All videos vary in duration (i.e., from 10 to 60 frames) and start from the first neutral frame to the last frame with a peak expression. The MMI dataset[2] includes more than 43 subjects, who express facial emotions non-uniformly and spontaneously. 213 video sequences in MMI have been labeled with six basic expressions, where some subjects wear hats, hoods, or glasses. The JAFFE dataset[3] is an expression dataset consisting of 219 images from 10 Japanese subjects who are female [60]. There are three or four images for each subject with each expression. The Static Facial Expressions in the Wild (SFEW) dataset[4] is collected by selecting static frames from Acted Facial Expressions in the Wild (AFEW) [63]. The SFEW dataset contains 95 subjects with unconstrained facial expressions (such as different poses, ages). RAF-DB[5] is a large-scale dataset captured from the Internet. This dataset contains about 30,000 facial images of thousands of subjects annotated with basic or compound expressions by 40 trained human labelers. In our experiment, only images with basic facial expressions are used. In total, there are 1007 different facial expression images selected from CK+, 606 images selected from MMI, 219 images selected from JAFFE, 663 images selected from SFEW, and 15,339 images selected from RAF-DB.

In all experiments, six basic non-neutral expressions and one neutral expression are selected from each of the three datasets. The six basic non-neutral expressions include angry, disgust, fear, happy, surprise, and sad expressions, which are respectively abbreviated as An, Di, Fe, Ha, Su, and Sa in this paper. Thus $N = 6$ for the following experiments.

Following [37], for each image in the dataset, we firstly manually locate the eye position and crop a 110×150 patch covering the facial region. Then, the manually cropped facial images are resized to the size of 32×32 (i.e., $d = 1024$) and converted to the gray-scale images. We conduct the 10-fold cross-validation on all subjects to evaluate the performance of the proposed method, as done in [30]. The training set is used to train the six DIFs corresponding to the six expressions and learn the transformation matrix $\mathbf{G}^*$ in Eq. (18). Their corresponding IFSL images are used to train a classifier. The neutral expression training images are shared during the training process of the six DIFs. For the parameter settings, we empirically set $C = 0.1$ in Eq. (6). The value of the regularization parameter β is set to 2.0.

We report the recognition rates obtained by the proposed IFSL on each of the six expressions and the average recognition rates using either SVM or *k*-NN.

4.2.2. Influence of different filter functions

As we mention previously (see Section 3.1), a variety of filter functions can be used in the optimization process of LDA. In this section, we evaluate the performance of IFSL with different filter functions, including the element-wise product (also called dot product), linear transform and convolution functions. Here, we use the three in-the-lab facial expression datasets in this experiment for performance evaluation.

The performance obtained by the proposed IFSL with different filter functions is shown in Fig. 4, where the results obtained using the SVM classifier are given in Fig. 4(a) and those obtained using the *k*-NN classifier are shown in Fig. 4(b). We can observe that the proposed IFSL with the dot product filter function achieves better performance than that with the other two filter functions on all the three datasets. The proposed IFSL with the linear transform achieves the worst results among the three filter functions. This is mainly because that the dot product filter can directly have an influence on the pixel-level values in the facial image, while the other two filter functions operate on the whole facial image. Therefore, the direct change of pixel-level values seems to be more effective for facial expression recognition since the number of training samples is limited. In other words, the discriminative local facial regions can be enhanced and the irrelevant local facial regions are suppressed by using the dot product filter. Furthermore, the recognition rates obtained by IFSL using SVM are higher than those obtained by IFSL using *k*-NN.

Therefore, in the following experiments, we will choose the dot product filter as the filter function of the proposed IFSL method.

4.2.3. Influence of different image sizes

In this section, we evaluate the performance of IFSL with different image sizes, including 8×8, 16×16 and 32×32. We also use the three in-the-lab facial expression datasets for performance evaluation.

The performance obtained by the proposed IFSL with different image sizes is shown in Fig. 5, where the results obtained using the SVM classifier are given in Fig. 5(a) and those obtained using the *k*-NN classifier are shown in Fig. 5(b). We can observe that the proposed IFSL with the image size of 32×32 achieves the best recognition rates, while IFSL with the image size of 8×8 obtains the worst results on the three datasets. The higher the image resolution is, the better the recognition performance is. This is mainly due to the fact that the proposed ISFL method can be beneficial from exploiting more information in the higher-resolution images.

[1] http://www.pitt.edu/~emotion/ck-spread.htm
[2] http://www.mmifacedb.com
[3] http://www.kasrl.org/jaffe.html
[4] https://cs.anu.edu.au/few/emotiw2015.html
[5] http://www.whdeng.cn/RAF/model1.html

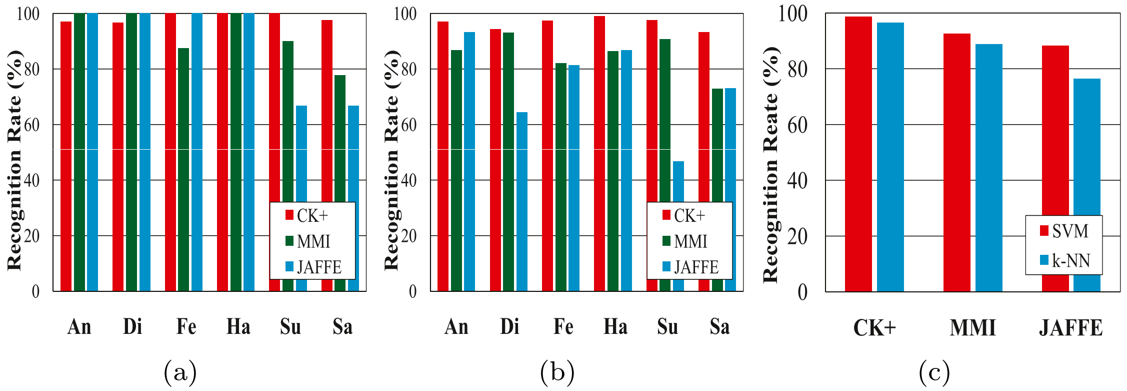

Fig. 6. Recognition results obtained by the proposed method with two different classifiers on the three facial expression datasets. (a) The recognition rate for six different expressions obtained by IFSL with SVM (b) The recognition rate for six different expressions obtained by IFSL with *k*-NN. (c): The average recognition rates obtained by the proposed IFSL using either SVM or *k*-NN on the three datasets.

Furthermore, the recognition rates obtained by IFSL using SVM are higher than those obtained by IFSL using *k*-NN.

4.2.4. Performance of the proposed method

In the following experiments, we show the performance of the proposed IFSL method to handle the task of multi-class facial expression recognition.

Table 1 shows the confusion matrix obtained by the proposed IFSL using the SVM classifier on the CK+ dataset. IFSL achieves good performance on all the expressions. However, we can also observe that some samples corresponding to the sad and disgust expressions are misclassified as the angry expression. This is because that both the disgust and sad expressions have the wrinkled eyebrows, and they share some similarities to the angry expression. Moreover, the subjects with the sad or disgust expressions do not have obvious motions around the mouth area, which also leads to the incorrect classification of these two expressions for some samples.

We also show the classification results for each expression obtained by the proposed method using the SVM and *k*-NN classifiers on all the datasets in a more compact way in Fig. 4. From the classification results for the best expressions on the three datasets with SVM (see Fig. 4(a)), IFSL achieves the top recognition rates on the happy expression while it obtains the worst results on the sad expression. These results can also be observed when using *k*-NN (see Fig. 4(b)). This is because that the happy expression has very obvious appearance variations around the mouth area compared with the other expressions. In contrast, the sad expression does not have significant appearance variations around the salient facial feature points. Actually, similar observations have been discussed in some other works [9,37]. We also observe that the sad and fear expressions sometimes show similar appearance for the subjects in the three facial expression datasets. Thus, IFSL can not accurately discriminate these two expressions for some subjects. As we can see in Fig. 4(c), IFSL with either SVM or *k*-NN achieves the best performance on the CK+ dataset. This is because CK+ is the simplest and the largest dataset among these three datasets. However, IFSL achieves the worst performance on the JAFFE dataset, since the number of the training samples is limited.

Table 1
The confusion matrix obtained by IFSL with SVM on the CK+ dataset. The best results are boldfaced.

(%)	An	Di	Fe	Ha	Su	Sa
An	**97.01**	2.99	0	0	0	0
Di	2.27	**96.59**	0	1.14	0	0
Fe	0	0	**100**	0	0	0
Ha	0	0	0	**100**	0	0
Su	0	0	0	0	**100**	0
Sa	2.5	0	0	0	0	**97.5**

4.2.5. Comparison with the state-of-the-art methods

We compare the proposed IFSL with several state-of-the-art methods. These methods include PCA [38], multi-class LDA [38], m-LBP [36], Boosted-LBP [37], PLBP [6], CSPL [35], HMFF [61], SalientPatch [9], CS-APL [64], MSCNN [65], pACNN [21], gACNN [21], DLP-CNN [41], and DAM-CNN [40]. The choice of these competing methods is based on the following reasons: 1) PCA and multi-class LDA are the two widely-used subspace learning methods for facial expression recognition. We use these two methods as the baseline. 2)The LBP-based methods (i.e., m-LBP, Boosted-LBP and PLBP) are regarded as the powerful feature extraction methods which achieve the state-of-the-art performance. 3) SalientPatch, CSPL and CS-APL are proposed to address part-based image representation and they effectively extract expression information in local facial regions, which is similar to the proposed method. 4) HMFF uses a subspace analysis method based on a hierarchical feature extraction framework, which also aims to enhance the discriminability of different expressions. 5) MSCNN, pACNN, gACNN, DLP-CNN and DAM-CNN are the representative CNN-based facial expression recognition methods. MSCNN can effectively extract spatial features under the supervision of recognition and verification signals. pACNN, gACNN adopt the attention mechanism in CNN. DLP-CNN employs the deep locality-preserving feature learning for FER. DAM-CNN designs a deep multi-path convolutional neural network by taking advantage of salient region attention. For all the competing methods, we use the default parameters from the respective papers.

Comparison results on the CK+ dataset Table 2 compares the proposed IFSL method with the state-of-the-art methods on the CK+ dataset. We can see that the proposed IFSL method achieves the best recognition rate, and significantly outperforms the traditional feature learning methods (such as GSPL, HMFF, SalientPatch) on the CK+ dataset. The PCA and multi-class LDA achieve poor performance on the CK+ dataset. This is mainly because the learned subspace obtained by either PCA or multi-class LDA is not discriminative to distinguish different facial expressions. The performance obtained by IFSL with SVM is better than that obtained by IFSL with *k*-NN in terms of the average recognition performance. Moreover, compared with some competing methods, such as m-LBP, Boost-LBP, GSPL and MSCNN, IFSL with the simple *k*-NN classifier achieves promising performance. As we mention previously, *k*-NN makes a classification decision by using the majority vote of

Table 2
Comparison results obtained by all the competing methods on the CK+ dataset. The best results are boldfaced.

Methods	Accuracy (%)
PCA (*k*-NN)	43.8
PCA (SVM)	47.3
multi-class LDA (*k*-NN)	84.7
multi-class LDA (SVM)	87.1
m-LBP [36]	88.4
Boost-LBP [37]	91.1
PLBP [6]	95.2
GSPL [35]	89.9
HMFF [61]	96.1
SalientPatch [9]	94.0
CS-APL [64]	98.6
MSCNN [65]	95.5
pACNN [21]	97.0
gACNN [21]	96.4
DAM-CNN [40]	95.9
IFSL (SVM)	**98.7**
IFSL (*k*-NN)	96.6

Table 3
Comparison results obtained by all the competing methods on the MMI dataset. The best results are boldfaced.

Methods	Accuracy (%)
PCA (*k*-NN)	65.6
PCA (SVM)	67.9
multi-class LDA(*k*-NN)	68.3 multi-class LDA
(SVM)	71.0
Boost-LBP [37]	86.9
PLBP [6]	91.0
GSPL [35]	73.5
MSCNN [65]	77.1
pACNN [21]	70.4
gACNN [21]	69.0
DLP-CNN [41]	78.5
IFSL (SVM)	**92.6**
IFSL (*k*-NN)	88.9

Table 4
Comparison results obtained by all the competing methods on the JAFFE dataset. The best results are boldfaced.

Methods	Accuracy (%)
PCA (*k*-NN)	52.4
PCA (SVM)	55.6
multi-class LDA (*k*-NN)	62.7
multi-class LDA (SVM)	64.4
Boost-LBP [37]	82.0
MSCNN [65]	85.1
DAM-CAM [40]	**99.3**
IFSL (SVM)	88.2
IFSL (*k*-NN)	76.4

Table 5
Comparison results obtained by all the competing methods on the SFEW dataset. The best results are boldfaced.

Methods	Accuracy (%)
PCA (*k*-NN)	23.4
PCA (SVM)	28.1
multi-class LDA (*k*-NN)	34.9
multi-class LDA (SVM)	39.3
MSCNN [65]	47.9
gACNN [21]	**51.7**
pACNN [21]	49.8
DLP-CNN [41]	51.1
IFSL (SVM)	46.5
IFSL (*k*-NN)	43.2

its k nearest neighbors, which indicates that intra-class variations are small and inter-class variations are large in the transformed subspace obtained by IFSL. In other words, the distributions of the samples corresponding to different expressions are well-separated in the subspace obtained by the proposed IFSL method. Compared with the CNN-based methods (such as MSCNN, pACNN, gACNN and DAM-CNN), the proposed IFSL still achieves better performance, when only limited training data are available. Therefore, the proposed method can effectively extract the discriminative and compact features from the LR facial images.

Comparison results on the MMI dataset Table 3 shows the comparison results obtained by the proposed IFSL method and some state-of-the-art methods on the MMI dataset. MMI is a well-known challenging facial expression dataset due to non-uniformly posed expressions and various head dressing. The proposed method with SVM obtains higher accuracy than PLBP [6], and it achieves much better performance than GSPL [35]. PLBP uses the images with the size of 110 × 150 and GSPL uses the images with the size of 95 × 95. The image resolutions used in these two methods are much larger than the image resolution used in IFSL (i.e., 32 × 32). From Table 3, we can see that PCA obtains much worse recognition rate than multi-class LDA. This is because multi-class LDA effectively reduces the within-class scatter while enlarging the between-class scatter. However, multi-class LDA is not able to discriminate the classes close to each other since large class distances are often overemphasized during training. In contrast, the proposed IFSL method benefits from the class specific filter learning and linear ridge regression techniques, which can distinguish a specific facial expression from the neutral expression and extract discriminative expression information from the facial image, respectively. IFSL also achieves better performance than MSCNN, which shows the effectiveness of the proposed method for LR facial expression recognition. The main reason is that MSCNN suffers from the problem of insufficient training data. pACNN, GACNN and DAM-CNN achieves worse results than the proposed method in the MMI datasets. This is mainly because that these methods use the CNN model trained on other datasets for feature extraction without fine-tuning.

Comparison results on the JAFFE dataset Table 4 shows the comparison results obtained by the proposed IFSL method and some state-of-the-art methods on the JAFFE dataset. The performance of only few existing methods is evaluated on JAFFE, since it is a small dataset. IFSL achieves relatively lower accuracy on JAFFE than that on MMI and CK+. Similarly, Boost-LBP also achieves the worst performance on JAFFE, compared with its performance on the other two datasets. This observation is especially obvious for multi-class LDA and PCA. This is mainly because JAFFE has a very small number of samples for training, which causes that the obtained facial expression features are less effective. Note that IFSL achieves worse performance than DAM-CAM. The main reason is that DAM-CAM fine-tunes the VGG model training on a large-scale dataset. Moreover, the input image size of DAM-CAM (i.e., 224 × 224) is much larger than that of IFSL (i.e., 32 × 32). In contrast, IFSL learns the parameters of DIF by only using the small training set.

Comparison results on the SFEW dataset Table 5 compares the proposed IFSL method with several state-of-the-art methods on the SFEW dataset. Among all the competing methods, gACNN and DLP-CNN respectively achieve the best and second performance (51.7% and 51.1%, respectively), which are better than the proposed IFSL with SVM by a moderate margin (5.2% and 4.6%, respectively). This

Table 6
Comparison results obtained by all the competing methods on the RAF-DB dataset. The best results are boldfaced.

Methods	Accuracy (%)
PCA (*k*-NN)	40.4
PCA (SVM)	42.1
multi-class LDA (*k*-NN)	48.6
multi-class LDA (SVM)	50.3
MSCNN [65]	77.2
gACNN [21]	**85.1**
pACNN [21]	83.3
DLP-CNN [41]	84.1
IFSL (SVM)	76.9
IFSL (*k*-NN)	72.6

is mainly because gACNN and DLP-CNN make use of the additional large-scale training sets (i.e., the AffectNet which contains 280,000 training samples and the RAF-DB which contains 15,339 training samples). gACNN adopts the local-global attention mechanism to capture subtle expression variations. DLP-CNN exploits the deep locality-preserving CNN to extract effective facial features. In contrast, the proposed IFSL method takes advantage of filter learning to obtain discriminative filters, which can enable the model to pay attention to distinctive facial regions.

Comparison results on the RAF-DB dataset Table 6 compares the proposed IFSL method with several state-of-the-art methods on the challenging RAF-DB dataset. Among all the competing methods, the CNN-based methods (such as gACNN, pACNN and DLP-CNN) achieves significantly better results ($>30\%$ improvement in recognition rates) than the traditional handcrafted feature-based methods (such as PCA and LDA), which shows the excellent performance achieved by deep learning. This is mainly because the large-scale training data are beneficial for boosting the performance of CNN. Although the proposed IFSL achieves worse results than these CNN-based methods, the input sizes of pACNN, gACNN and DLP-CNN are respectively 256×256, 256×256, 224×224, which are much higher than the input image size (32×32) of the proposed method. These CNN-based methods take advantage of high-resolution images for feature learning. In other words, these methods can extract effective features for sufficient information in these HR facial images. Note that RAF-DB contains a large number of training data. The proposed IFSL still achieves the comparable performance compared with a CNN-based method (MSCNN), which demonstrates the superiority learning capability of the proposed filter learning method.

In summary, the above experimental results show that the proposed IFSL can achieve excellent recognition performance for LR facial expression recognition, which indicates IFSL is good at extracting useful expression information in the LR facial images.

4.3. Limitations and future work

Although the proposed IFSL method achieves promising results, it also has some limitations. Firstly, the proposed method only works on frontal or near-frontal facial image samples. Handling facial expressions with large pose variations is a more challenging task while the proposed method does not address this challenge at the current stage. In future, we can take advantage of the image synthesis methods (such as GANs [29]) to generate the frontal facial images and combine the generated images with the image filter to achieve pose-invariant expression recognition. Secondly, when the size of sample images increases, the speed of computing the derivation of DIF and LRR decreases at the training stage. Thus, how to improve the computational speed of the proposed method is still an open question. For example, we can adopt some optimization methods (such as [66]) to efficiently compute the gradients. Note that as the training stage is usually performed offline, the computational complexity of the proposed method will not greatly constrain its applications to real-world tasks.

In addition, the proposed method trains an image filter for each class. Therefore, the proposed method is not suitable for the classification problem that involves a large number of classes (such as face recognition with millions of persons). However, the image filter can be useful to filter out the irrelevant information for the classification problem.

5. Conclusion

In this paper, we propose a novel image filter based subspace learning (IFSL) method for effective image representations. We show that a discriminative image filter (DIF) can be effectively learnt by incorporating the image filter into the cost function of LDA. The learned DIF can not only filter out useless information, but also preserve useful information for discriminating facial expressions. Furthermore, we develop a regression learning approach to explore the most discriminative information in the combined filtered images (generated by DIFs) by constructing an expression-aware transformation matrix, which successfully encodes expression information while reducing the influence of facial identity differences. Experimental results on several popular facial expression datasets are presented to demonstrate the effectiveness of the proposed IFSL on LR facial expression recognition. Compared with several state-of-the-art methods, the proposed method achieves superior results.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2017YFB1302400, by the National Natural Science Foundation of China under Grants 61571379, U1605252, 61872307, by the Natural Science Foundation of Fujian Province of China under Grants 2017J01127 and 2018J01576, and by the Fundamental Research Funds for the Central Universities under Grant 20720170045.

References

[1] M. Pantic, L. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1424–1445.
[2] C.A. Corneanu, M. Oliu, J.F. Cohn, S. Escalera, Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications, IEEE Trans. Pattern Anal. Mach. Intell. 38 (12) (2016) 1548–1568.
[3] Y. Huang, Y. Yan, S. Chen, H. Wang, Expression-targeted feature learning for effective facial expression recognition, J. Vis. Commun. Image Represent. 55 (2018) 677–687.
[4] Y. Fu, Q. Ruan, Z. Luo, Y. Jin, G. An, J. Wan, FERLRtc: 2d+3d facial expression recognition via low-rank tensor completion, Signal Process. 161 (2019) 74–88.
[5] S. Li, W. Deng, Deep facial expression recognition: a survey, 2018, arXiv preprint, arXiv:1804.08348.
[6] R.A. Khan, A. Meyer, H. Konik, S. Bouakaz, Framework for reliable, real-time facial expression recognition for low resolution images, Pattern Recognit. Lett. 34 (10) (2013) 1159–1168.
[7] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2014) 295–307.
[8] J. Jiang, J. Ma, C. Chen, X. Jiang, Z. Wang, Noise robust face image super-resolution through smooth sparse representation, IEEE Trans. Cybern. 47 (11) (2017) 3991–4002.
[9] S.L. Happy, A. Routray, Automatic facial expression recognition using features of salient facial patches, IEEE Trans. Affect. Comput. 6 (1) (2015) 1–12.

[10] J. Shao, I. Gori, S. Wan, J.K. Aggarwal, 3D dynamic facial expression recognition using low-resolution videos, Pattern Recognit. Lett. 65 (C) (2015) 157–162.

[11] S.C. Park, K.P. Min, M.G. Kang, Super-resolution image reconstruction: a technical overview, IEEE Signal Proc. Mag. 20 (3) (2003) 21–36.

[12] Z. Wang, D. Liu, J. Yang, W. Han, T. Huang, Deep networks for image super-resolution with sparse prior, in: Proceedings of the IEEE Conference on Computer Vision, 2015, pp. 370–378.

[13] Y. Tian, T. Kanade, J.F. Cohn, Facial expression recognition, Springer, 2011.

[14] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, L. Prevost, Facial action recognition combining heterogeneous features via multikernel learning, IEEE Trans. Syst. Man Cybern. B 42 (4) (2012) 993–1005.

[15] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1683–1699.

[16] G. Zhao, M. Pietikinen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 27 (6) (2007) 915–928.

[17] J. Whitehill, C.W. Omlin, Haar features for FACS AU recognition, in: Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition, 2006, pp. 217–222.

[18] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Recognizing facial expression: machine learning and application to spontaneous behavior, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 568–573.

[19] F. Zhang, Q. Mao, X. Shen, Y. Zhan, M. Dong, Spatially coherent feature learning for pose-invariant facial expression recognition, ACM Trans. Multimedia Comput. Commun. Appl. 14 (1s) (2018) 27:1-27:19.

[20] S. Xie, H. Hu, Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks, IEEE Trans. Multimed. 21 (1) (2019) 211–220.

[21] Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using CNN with attention mechanism, IEEE Trans. Image Process. 28 (5) (2019) 2439–2450.

[22] D.L. Swets, J.J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836.

[23] C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[24] Z. Zhang, H. Wang, Y. Yan, Discriminative filter based regression learning for facial expression recognition, in: Proceedings of the IEEE International Conference on Image Processing, 2013, pp. 1192–1196.

[25] L. Lin, Y. Zhang, W. Zhang, Z. Chen, Y. Yan, T. Yu, A real-time smile elegance detection system: a feature-level fusion and SVM based approach, in: Proceedings of the IS&T International Symposium on Electronic Imaging, 2017, pp. 80–85.

[26] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, L. Zhang, Convolutional sparse coding for image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision, 2016, pp. 1823–1831.

[27] X. Ma, J. Zhang, Q. Chun, Hallucinating face by position-patch, Pattern Recognit. 43 (6) (2010) 2224–2236.

[28] X. Wang, X. Tang, Hallucinating face by eigen transformation, IEEE Trans. Syst. Man Cybern. C 35 (3) (2005) 425–434.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, in: Proceedings of the Advanced Neural Information Processing Systems, 2014, pp. 2672–2680.

[30] X. Yu, F. Porikli, Ultra-resolving face images by discriminative generative networks, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 318–333.

[31] C.X. Ren, D.Q. Dai, H. Yan, Coupled kernel embedding for low-resolution face image recognition, IEEE Trans. Image Process. 21 (8) (2012) 3770–3783.

[32] J. Jiang, R. Hu, Z. Wang, Z. Cai, CDMMA: coupled discriminant multi-manifold analysis for matching low-resolution face images, Signal Process. 124 (2016) 162–172.

[33] X. Xing, K. Wang, Couple manifold discriminant analysis with bipartite graph embedding for low-resolution face recognition, Signal Process. 125 (2016) 329–335.

[34] Y. Chu, T. Ahmad, G. Bebis, L. Zhao, Low-resolution face recognition with single sample per person, Signal Process. 141 (2017) 144–157.

[35] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, D. Metaxas, Learning active facial patches for expression analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2562–2569.

[36] C. Shan, S. Gong, P. McOwan, Robust facial expression recognition using local binary patterns, in: Proceedings of the IEEE International Conference on Image Processing, 2005, pp. 914–917.

[37] C. Shan, S. Gong, P. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.

[38] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[39] A.J. Calder, A.M. Burton, P. Miller, A.W. Young, S. kamatsu, A principal component analysis of facial expressions, Vis. Res. 41 (9) (2001) 1179–1208.

[40] S. Xie, H. Hu, Y. Wu, Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition, Pattern Recogni. 72 (2019) 177–191.

[41] S. Li, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, IEEE Trans. Image Process. 28 (1) (2019) 356–370.

[42] S. Li, W. Deng, A deeper look at facial expression dataset bias, 2019, arXiv preprint, arXiv:1904.11150.

[43] A. Majumder, L. Behera, V.K. Subramanian, Automatic facial expression recognition system using deep network-based data fusion, IEEE Trans. Cybern. 48 (1) (2018) 103–114.

[44] Y. Zong, X. Huang, W. Zheng, Z. Cui, G. Zhao, Learning from hierarchical spatiotemporal descriptors for micro-expression recognition, IEEE Trans. Multimed. 20 (11) (2018) 3160–3172.

[45] O. Gupta, D. Raviv, R. Raskar, Illumination invariants in deep video expression recognition, Pattern Recognit. 76 (2018) 25–35.

[46] M. Alam, L.S. Vidyaratne, K.M. Iftekharuddin, Sparse simultaneous recurrent deep learning for robust facial expression recognition, IEEE Trans. Neural Netw. Learn. Syst. 29 (10) (2018) 4905–4916.

[47] J. Chen, Z. Chen, Z. Chi, H. Fu, Facial expression recognition in video with multiple feature fusion, IEEE Trans. Affect. Comput. 9 (1) (2018) 38–50.

[48] J. Whitehill, J. Movellan, Discriminately decreasing discriminability with learned image filters, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2488–2495.

[49] Y. Yan, H. Wang, D. Suter, Multi-subregion based correlation filter bank for robust face recognition, Pattern Recognit. 47 (11) (2014) 3487–3501.

[50] J.F. Henriques, C. Rui, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2014) 583–596.

[51] A. Rodriguez, V.N. Boddeti, B.V.K.V. Kumar, A. Mahalanobis, Maximum margin correlation filter: a new approach for localization and classification, IEEE Trans. Image Process. 22 (2) (2013) 631–643.

[52] P. Burkert, F. Trier, M.Z. Afzal, A. Dengel, M. Liwicki, Dexpression: deep convolutional neural network for expression recognition, 2016, arXiv preprint, arXiv:1509.05371.

[53] W. Zheng, Y. Zong, X. Zhou, M. Xin, Cross-domain color facial expression recognition using transductive transfer subspace learning, IEEE Trans. Affect. Comput. 9 (1) (2016) 21–37.

[54] S. An, W. Liu, S. Venkatesh, Face recognition using kernel ridge regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–7.

[55] Z. Li, D. Lin, X. Tang, Nonparametric discriminant analysis for face recognition, IEEE Trans. on Pattern Anal. Mach. Intell. 31 (4) (2009) 755–761.

[56] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2006.

[57] S. Chew, S. Lucey, P. Lucey, S. Sridharan, J. Conn, Improved facial expression recognition via uni-hyperplane classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2554–2561.

[58] Z. Wang, Z. Miao, Q. Wu, Y. Wan, Z. Tang, Low-resolution face recognition: a review, Vis. Comput. 39 (4) (2014) 359–386.

[59] P. Lucey, J. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2010, pp. 94–101.

[60] M. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (12) (1999) 1357–1362.

[61] Z. Zhang, C. Fang, X. Ding, A hierarchical algorithm with multi-feature fusion for facial expression recognition, in: Proceedings of the IEEE Conference on Pattern Recognition, 2012, pp. 2363–2366.

[62] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2011, pp. 2106–2112.

[63] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Collecting large, richly annotated facial-expression databases from movies, IEEE Multimed. 19 (3) (2012) 34–41.

[64] K. Zhao, W. Chu, F. De la Torre, J.F. Cohn, H. Zhang, Joint patch and multi-label learning for facial action unit and holistic expression recognition, IEEE Trans. Image Process. 25 (8) (2016) 3931–3946.

[65] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Trans. Image Process. 26 (9) (2017) 4193–4203.

[66] M. M. Schmidt, N.L. Roux, F.R. Bach, Convergence rates of inexact proximal–gradient methods for convex optimization, in: Proceedings of the Advanced Neural Information Processing Systems, 2011, pp. 1458–1466.