Pattern Recognition 132 (2022) 108916

Contents lists available at ScienceDirect

Pattern Recognition

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

Multi-feature sparse similar representation for person identification

Meng Yang[a,b,d,*], Lei Liao[a], Kangyin Ke[a], Guangwei Gao[c,*]

[a] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[b] Key Laboratory of Machine Intelligence and Advanced Computing (SYSU), Ministry of Education, China
[c] Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing, China
[d] Guangdong Key Laboratory of Information Security Technology, China

ARTICLE INFO

Article history:
Received 31 October 2020
Revised 14 July 2022
Accepted 20 July 2022
Available online 21 July 2022

Keywords:
Multi-feature
Person identification
Sparse representation

ABSTRACT

Person identification with a single feature (e.g., face recognition, speaker verification, person re-identification, etc.) has been studied extensively for many years, while few works focus on multi-feature person identification. Though promising performance has been achieved by only using the information of facial images, voice, or pedestrian appearance, it is still challenging to recognize a person with only a single feature in some situations (e.g., a person at a distance or occluded by other objects, and a partial person out of view). In this paper, we present a multi-feature sparse similar representation (MFSSR) method to effectively fuse face features, body features, and global image features for the task of person identification. In MFSSR, we designed a reconstructed deep spatial feature for representing the appearance of human body by using the spatial correlation coding of partial deep spatial features. Then we presented a multi-feature sparse similar representation model for jointly using different features, e.g., face, body, and the global image. Besides, considering that the coding coefficients associated with good samples but not outliers should be more similar among different features, we jointly represent different features by imposing a weighted ℓ_1-norm distance regularization, instead of the conventional ℓ_2-norm regularization, on the coefficients. Experimental results on several multi-feature person identification databases have clearly shown the superior performance of the proposed model.

1. Introduction

With the emergence of advanced network optimization algorithms and developments of the Graphics Processing Unit (GPU), deep learning has made revolutionary progress with the support of high-performance computing. With the guarantee of various sources of big data, such as images and videos on the Internet, deep learning has a steady stream of motivation, with excellent network structures continuously proposed. In the research fields of deep learning and computer vision, the development of person identification technologies[1–4] is still the focus of industry and academia. Image-based person identification methods have used various features, among which facial features [5–7] and body features [8–11] play important roles because they are easy to obtain and very discriminative for identifying a person.

Although recent years have witnessed tremendous advances in person identification with a single feature, few works considered effectively combining multiple features. For single-feature person identification, there are still some issues that are too difficult to be solved. For instance, for face recognition, we can hardly handle the problems of face images with severe blur or nearly a complete occlusion, recognition at a distance, and even invisibility of faces due to different viewpoints and poses. Re-ID recognizes a person mainly based on the body appearance, but it may not be very reliable when the same person has different poses at different times. Meanwhile, the human body can be occluded by other objects or outside the field of view, which, named partial Re-ID, can not achieve satisfactory identification performance.

In a real-world situation, person identification is always performed in an unconstrained environment, making it very challenging to recognize a person. For example, to predict the identity of an actor in television, there is no guarantee that his or her frontal face or body appearance is complete or not occluded. Instead, most likely, only a profile or a blurred face image is provided (see Fig. 1). Fortunately, there are always multiple cues for us to recognize one person. Also, we take Fig. 1 as an example, in which the person is dressed in the same clothes in two different images. To recognize the same person across images, the body appearance of the person is another powerful feature besides face features. To promote

* Corresponding author.
* Corresponding Author
E-mail addresses: yangm6@mail.sysu.edu.cn (M. Yang), csggao@gmail.com (G. Gao).

https://doi.org/10.1016/j.patcog.2022.108916
0031-3203/

Fig. 1. Examples of iQIYI-VID dataset [12]. The left image is a sample in the training set, and the other is in the testing set. The iQIYI-VID dataset poses new challenges for identity recognition. Due to the changes in hairstyles and pose, as well as facial occlusion, it is necessary to combine face recognition, body recognition, and other feature recognition to achieve better results.

the accuracy of person identification, it is a more promising way to jointly take advantage of face, body features and so on, compared to the single-feature recognition with face image or body appearance. With more information provided by multiple sources, effectively combining different features will easily achieve higher recognition accuracy, and making person identification become more robust. For instance, Liu et al. [12] extracted different kinds of features, set the dimensions of them to the same, and fused them by a weighted sum. However, the correlation of different features is ignored and how to fully exploit the collaborative information of different features is still the key issue.

In recent years, multi-feature recognition has been effectively conducted by sparse representation, which has received wide attention in computer vision and pattern recognition, such as face recognition [5,13,14]. Following the line of sparse representation, multi-task joint sparse or collaborative representation methods have been investigated to deal with multi-feature visual problems, where the correlation among different features is exploited to improve recognition performance. In [15], a multi-task joint sparse representation is proposed to tackle visual classification, which imposes a $\ell_{1,2}$-norm regularization on coding coefficients of different features. Zhang et al. [16] adopted a joint dynamic sparsity prior for multiple observations of the same physical object. As for [17], the distance between coding coefficients of different features is minimized by a weighted distance regularization.

Although these methods have shown the superiority of fusing multiple features in various visual classification tasks, they have not considered the influence of specific samples (e.g., with outliers or without discrimination) when exploiting the similarity and distinctiveness between different features. Moreover, they have not touched on the challenging person identification task. To address multi-feature person identification, we propose a novel multi-feature sparse similar representation (MFSSR) model to fuse different types of features. Unlike conventional joint representation models with ℓ_2-norm regularization, the proposed MFSSR adopts a ℓ_1-norm weighted distance regularization, which is more robust in the coding stage. In our method, three features will be used, including face feature, body feature, and global image feature. Particularly, we proposed to represent the body feature by adaptively building the correlation of local regions and jointly reconstructing the deep spatial feature, which utilizes the robustness, joint information, and discrimination of local deep spatial features. Experimental results on several multi-feature person identification databases substantiate the effectiveness of reconstructed spatial features and ℓ_1-norm weighted distance regularization.

Our main contributions are summarized as follows.

- The discriminative body appearance feature is designed by reconstructing different spatial features with their spatial correlation and jointly combining all spatial feature representations.
- ℓ_1-norm weighted distance regularization is adopted in our multi-feature representation, which is more robust than the conventional ℓ_2-norm regularization.
- Competitive results have been achieved by our proposed model in addressing the multi-feature person identification problem.

The rest of this paper is organized as follows. Section 2 introduces the related works. Section 3 describes the proposed multi-feature sparse similar representation. Section 4 presents the experimental results on several databases. Section 5 concludes this paper.

2. Related works

Multiple feature extraction and multi-feature classification are two key steps in multi-feature person identification. Most of the feature extraction and classification methods are related to the deep neural network with a single feature, which is different from the multi-feature-based person identification. In this section, we will first review the methods of facial feature extraction and body appearance extraction, and then compare existing methods in multi-feature recognition tasks in general.

2.1. Facial feature extraction

Facial features for person identification (i.e., face recognition) have been studied for many years. Subspace-based methods for face recognition have attracted much attention during the past several decades. One of the classic methods is first to reduce the dimensionality of image features using the Principal Component Analysis (PCA) [18], and then use the Linear Discriminant Analysis (LDA) to find discriminant projections. Another representative way to conduct face recognition is collaborative (sparse) representation, which synthesizes an input by using all classes' dictionaries (e.g., training samples). Collaborative representation with various discriminative regularizations, such as sparse coefficient and Fisher discrimination, has promoted the accuracy of robust face recognition. John Wright et al. [5] proposed face recognition based on sparse representation, which reconstructs the input signal by using a linear combination of all training samples, with the combination coefficient having as few non-zero elements as possible. Yang et al. [19] proposed the Fisher discrimination dictionary learning (FDDL), which integrates the Fisher discriminant regularization into the dictionary representation to learn a structured discriminative dictionary.

As deep learning rapidly develops, deep features have been widely adopted to replace the hand-crafted features. A common deep network structure for image feature extraction is the convolutional neural network and its variants, such as ResNet [20], AlexNet [21], VGGNet [22], and so on. The loss function of a deep network, which measures the classification ability of the model, has been one of the main-stream researches to optimize deep networks. In 2014, DeepFace [23] used a SoftMax function in its classification layer. Then Google's FaceNet [24] used a triplet loss function to replace the SoftMax cross-entropy loss function and center-Loss [25] added a loss function that maintains the category center and makes features close to their category center, thereby achieving a similar effect to Triple Loss. These algorithms usually focus on metric learning [6,24–27] and margin-based loss [7,28–33], achieving promising results in the databases of face recognition (e.g., ArcFace [7] got an accuracy of 99.83% on LFW [34] dataset). Besides, Sadiq et al. proposed an Attentive Occlusion-adaptive Deep Network (AODN) [14], which introduces the attention module consisting of Channel-wise Attention (CA) and Spatial Attention (SA)

to improve its ability to deal with occlusions and enhance feature representation ability simultaneously. A novel geodesic-guided convolution (GeoConv) [35] was proposed by Chen et al., which can be applied to fine-grained face analysis tasks.

2.2. Body appearance extraction

Body appearance is another important feature for person identification, especially in person re-identification (Re-ID). As one of the most active research topics in the field of computer vision, the task of Re-ID is to match the same person across different non-overlapping camera views. The past several years have witnessed remarkable strides of Re-ID [11] [36] [37]. Some of the latest works even exceed human performance [11]. To solve the inconsistent distributions under different views, Zhao et al. [10] proposed a method of joint transfer constraint to learn the similarity function by combining multiple common subspaces, each in charge of a sub-region. Re-ID can also be combined with semi-supervised learning [38] and weakly supervised multi-type attribute learning [39] to complete recognition tasks.

Nevertheless, Re-ID is still very challenging, in particular when only a partial human body is available. Some effective methods have been proposed to address the partial Re-ID problem [8] [9] [37]. For instance, to deal with identifying a partial body of a person, Zheng et al. [9] firstly decomposed images into small local patches and then computed a patch-level ambiguity score between a probe and each gallery by performing a local-to-local matching, which is named Ambiguity-sensible Matching Classifier (AMC). Besides, they further proposed a sliding window matching (SWM) method to perform a global-to-local matching. He et al. [8] proposed a deep spatial feature reconstruction model (DSR), in which a Fully Convolutional Network (FCN) is trained to generate spatial feature maps of a certain size and the probe image is matched to a gallery image with the minimum reconstruction error of all spatial features. After that, He et al. focused on recognizing partial input with the assistance of proposed Part-Part Correspondence Learning (PPCL) [37], which is a self-supervised learning framework that learns correspondence between image patches without any additional part-level supervision. Chen et al. [4] proposed to learn global and local attention aware features for person ReID, which introduces two additional branches to realize the proposed attention aware feature learning in the training stage and removes them in the inference time to keep the same model size and inference speed.

In DSR, a fully convolutional network (FCN) is trained on Market1501 database [40] and fine-tuned on Partial-REID database [9]. Suppose a pair of person images $\mathbf{I}$ and $\mathbf{J}$ are given. Then, we can extract spatial feature maps $\mathbf{x} = FCN(\mathbf{I}, \theta)$ and $\mathbf{y} = FCN(\mathbf{J}, \theta)$ by FCN, where θ is the parameters of FCN. Suppose the size of feature map $\mathbf{x}$ is $w \times h \times d$, where w, h and d represent the width, the height, and the number of channels, respectively. Then the feature map $\mathbf{x}$ is divided into $N = w \times h$ spatial sub-features, which are represented as the set of $\mathbf{x}$, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{d \times N}$. In the same way, if $\mathbf{y}$ have M spatial sub-features because of the different weight and height, the set of sub-feature maps $\mathbf{y}$ is denoted as follow: $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M\} \in \mathbb{R}^{d \times M}$. To measure the similarity of two sub-feature sets, DSR [8] firstly searches the similar sub-feature for the set $\mathbf{X}$ by using the linear combination of the sub-features in $\mathbf{Y}$, i.e., $\min \|\mathbf{X} - \mathbf{YW}\| + \beta \|\mathbf{W}\|_1$, where $\mathbf{W} \in \mathbb{R}^{M \times N}$ denotes the sparse coefficients of all blocks of $\mathbf{X}$ and β is a small scalar (e.g., 0.001) that controls the sparsity of the coding coefficients. After the sparse coefficients are obtained, the similarity metric between block set $\mathbf{X}$ and $\mathbf{Y}$ is computed using the reconstruction error

$$e = \frac{1}{N} \|\mathbf{X} - \mathbf{YW}\|_F^2 \tag{1}$$

Although the sparse representation has enough flexibility to tolerate the spatial variation and different sub-feature numbers, it ignores the location information and the collaborative information between different sub-features, which can further advance the final identification performance.

2.3. Multi-feature recognition

We review some multi-feature representation methods related to our work on multi-feature person recognition. Liu et al. proposed a baseline approach for multi-modal person identification on the iQIYI-VID dataset in [12]. They described a video by dividing it into four parts, including face, head, body, and audio features, which are extracted by some state-of-the-art models such as ArcFace [7] and ResNet [20]. Then a multi-feature attention module in the baseline is used to fuse different kinds of features. However, they only concatenate different features of the human body, while ignoring their spatial information.

A relaxed collaborative representation model was proposed for multi-feature recognition in [17] based on the collaborative representation [41], where a weighted distance regularization is imposed on the coding coefficients of different types of features. In the coding stage, the between-feature-coefficient distance of a sample's representation should be minimized. With ℓ_2-norm distance regularization, all the coding coefficient elements would be close to the weighted mean values. However, this model ignores the distinctiveness of different coding vector elements (e.g., the correlation between a sample and a dictionary atom may not be similar in different features). The coefficient associated with the dictionary atom including outliers in some feature should be different from those in other features. A better regularization robust to the outliers should be designed.

In addition to the RCR [17], multi-feature alignment is also used in the Semantic-Guided Shared Feature Alignment (SGSFA)[42] proposed by Ren et al. in 2020. SGSFA contains a Semantic Guided Alignment(SGA) branch and a Spatial Feature Alignment(SFA), which produces several pixel-wise attention maps, highlighting the visible body part while suppressing the occluded region or background. Usually, facial features account for a large proportion in the multi-feature recognition task. Although the SGSFA divides the images into three broad zones and uses the spatial information to reduce the effect of occlusion of information about people, it is still not sufficient to better avoid the problem of occlusion. Therefore, in the multi-feature recognition task, how to avoid the impact of obscuring the person's identity information on the basis of how to better fuse multiple features is also a problem we need to solve.

3. Multi-feature sparse similar representation

To solve the aforementioned issues related to feature extraction and multi-feature classification, we proposed to adopt multiple features to better exploit the discrimination of person images and designed a new flexible regularization to jointly learn the representation of multiple features and preserve the representation distinctiveness of dictionary atoms.

In this section, we present a novel multi-feature sparse similar representation (MFSSR) model for multi-feature person identification. In the proposed model, we extracted three different kinds of features, including the reconstructed deep spatial feature, global image feature, and face feature, in which the global image feature and face feature are generated from pre-trained deep models, and the reconstructed deep spatial feature is specially designed to tackle partial body appearance problem in person identification. Especially, considering the lack of individual feature spatial location information in the baseline approach proposed in [12] we build the adaptive spatial correlation of local regions and exploit

their joint information in the coding phase. For the multi-feature joint representation model, we proposed a novel ℓ_1-norm weighted distance regularization to exploit the representation distinctiveness associated with dictionary atoms, which can make a better balance between the commonality and distinctiveness of different features.

3.1. Reconstructed deep spatial feature

In this section, we will present the reconstructed deep spatial feature as a powerful body appearance representation for our person identification task. Different from the DSR [8] model which is designed for partial re-identification, our proposed reconstructed deep spatial feature generates a discriminative feature for the whole body appearance. Based on the DSR model, we will introduce the spatial correlation to conduct a spatial correlation coding, and then combine all partial features with tolerance to spatial misalignment and a partial observation of a person.

Given a pair of person images, we conduct the feature map extraction via a fully convolutional network. After deep feature extraction as DSR [8], two person images can be represented as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{d\times N}$ and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M\} \in \mathbb{R}^{d\times M}$, where d is the feature dimensionality, N and M are the number of spatial blocks (i.e., sub-features) in these two images, respectively.

Each sub-feature in each spatial block is independently coded in DSR [8]. In this manner, misalignment and partial features can be handled, however, the joint discrimination of partial features in correlated spatial blocks is weakened. Denote by $\mathbf{C}^X$ and $\mathbf{C}^Y$ the spatial correlation matrix of all partial features from $\mathbf{X}$ and $\mathbf{Y}$, respectively. $\mathbf{C}^X$ and $\mathbf{C}^Y$ are square matrices containing the relative position information. According to the picture partition algorithm in the model, we set the corresponding correlation value of two overlapping partitions or two adjacent partitions in the spatial correlation matrix to 1; otherwise, we set it 0. That is, if the i-th block and the j-th block in $\mathbf{X}$ are adjacent or overlapped, $\mathbf{C}^X_{i,j} = 1$; otherwise, $\mathbf{C}^X_{i,j} = 0$. Then $\mathbf{C}^X$ and $\mathbf{C}^Y$ can be obtained by considering all pairs of blocks in the $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Therefore, considering the spatial correlation coding with $\mathbf{C}^X$ and $\mathbf{C}^Y$, the optimized formula of the reconstructed deep spatial feature is:

$$\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{Y}\mathbf{W}\|_F^2 + \beta\|\mathbf{W}\|_F^2 - \gamma \sum_{i=1}^{N}\sum_{j=1}^{M} \mathbf{W}(:,i)^T \mathbf{C}^Y \mathbf{W}(:,j)^T \mathbf{C}^X_{i,j} \quad (2)$$

where γ is a parameter that controls the proportion of spatial position information. The third term is to increase the similarity between different partial features, where $\mathbf{C}^X_{i,j}$ denotes their correlation in the spatial domain. In this term, the similarity between coefficient vectors (e.g., $\mathbf{W}(:,i)$ and $\mathbf{W}(:,j)$) are regularized by the spatial correlation matrix $\mathbf{C}^Y$ of the coding dictionary $\mathbf{Y}$.

To effectively use the location information and collaboration between different blocks, we then design the reconstructed deep spatial feature as a new representation for the whole body. As shown in Fig 2, suppose there are S_j training samples in class j. After extracting their deep spatial feature and organizing them as block sets, we can get $\left\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{S_j}\right\}$. Let $\mathbf{Y}_s$ be the block sets of a training sample. When a testing sample comes, likewise, its block set is represented as $\mathbf{X}$. Then, after computing each coding matrix $\mathbf{W}_s$ associated to $\mathbf{Y}_s$ by solving Eq. (2), the block sets reconstructed for $\mathbf{X}$ with $\mathbf{Y}$ are:

$$\left\{\mathbf{Y}_1\mathbf{W}_1, \mathbf{Y}_2\mathbf{W}_2, \ldots, \mathbf{Y}_{S_j}\mathbf{W}_{S_j}\right\} \quad (3)$$

where $\mathbf{X} \approx \mathbf{Y}_s\mathbf{W}_s$ and $\mathbf{W}_s$ are the coding matrix associated to $\mathbf{Y}_s$.

Although the deep spatial feature represented by $\mathbf{X} \approx \mathbf{Y}_s\mathbf{W}_s$ can generate a discriminative representation for $\mathbf{X}$, the inter-block representation of $\mathbf{Y}_s\mathbf{W}_s$ without considering other intra-class samples will ignore the collaboration of different samples and introduce some disturbance. As shown in Fig. 3, the distance of $\mathbf{X}$ to some instance of the wrong class may be smaller than that to any sample of the correct class, although the subspace of the correct class represents $\mathbf{X}$ better than that of the wrong class.

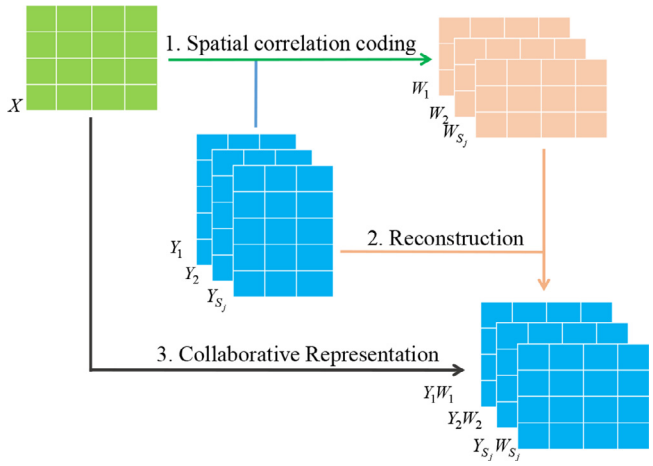


Fig. 2. Reconstructed Deep spatial Feature.

In order to utilize all intra-class information, the testing sample $\mathbf{X}$ can be represented by all reconstructed spatial features from class j, which can be written as a collaborative representation [41]

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{y}_b - \boldsymbol{D}_b\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2 \quad (4)$$

where $\boldsymbol{y}_b = vec(\mathbf{X})$ and $\boldsymbol{D}_b = [vec(\mathbf{Y}_1\mathbf{W}_1), vec(\mathbf{Y}_2\mathbf{W}_2), \ldots, vec(\mathbf{Y}_{S_j}\mathbf{W}_{S_j})]$. Here vec is a function to concatenate all columns of a matrix, organizing a matrix as a vector.

3.2. Multi-feature sparse similar representation model

Our method directly extracts the global features by using ResNet-50 [20] pre-trained on ImageNet [43] dataset and the face features by using the state-of-art face recognition model ArcFace [7], respectively. These two features are directly extracted from the image on fully convolutional networks, which are pre-trained on Market1501 [40]. Different from DSR[8] which directly extracts the deep spatial features from the picture, we use the proposed spatial correlation coding to obtain the reconstructed deep spatial features, which are added to the model as the third type of feature.

Although joint representation models, e.g., RCR [17], have been proposed to exploit multiple features, the conventional regularization of ℓ_2-norm is not well enough to exploit the correlation of different features and the representation distinctiveness associated with different atoms, e.g., outliers in atom-level.

The proposed multi-feature sparse similar representation (MFSSR) model is defined as

$$\min_{\boldsymbol{\alpha}_k} \sum_{k=1}^{K} (\|\boldsymbol{y}_k - \boldsymbol{D}_k\boldsymbol{\alpha}_k\|_2^2 + \lambda\|\boldsymbol{\alpha}_k\|_2^2 + \tau\omega_k\|\boldsymbol{\alpha}_k - \overline{\boldsymbol{\alpha}}\|_1) \quad (5)$$

where K is the number of features (e.g., body feature, global feature, and face feature), $\mathbf{D}_k$ represents the k^{th} feature dictionary from some class, $\boldsymbol{\alpha}_k$ is associated with $\mathbf{D}_k$, λ and τ are scalar constants, ω_k indicates the weight of the feature $\mathbf{y}_k$, and $\overline{\boldsymbol{\alpha}}$ is a center of multi-feature representation. To neglect the coefficients associated with dictionary atoms of outliers when exploiting the similarity between different features, we proposed to minimize a ℓ_1-norm

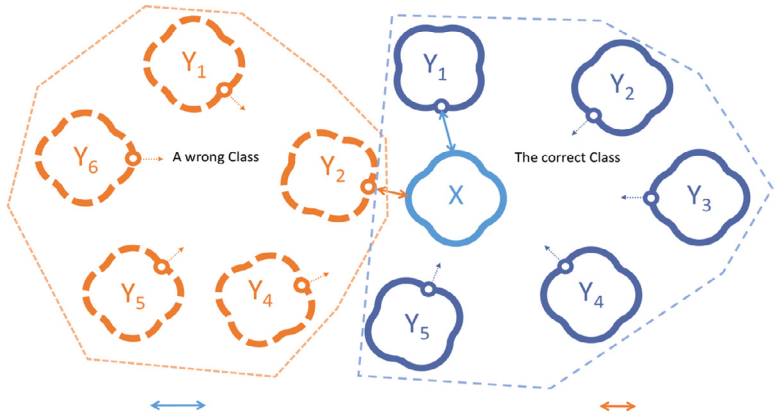


Fig. 3. The deep spatial feature representation.

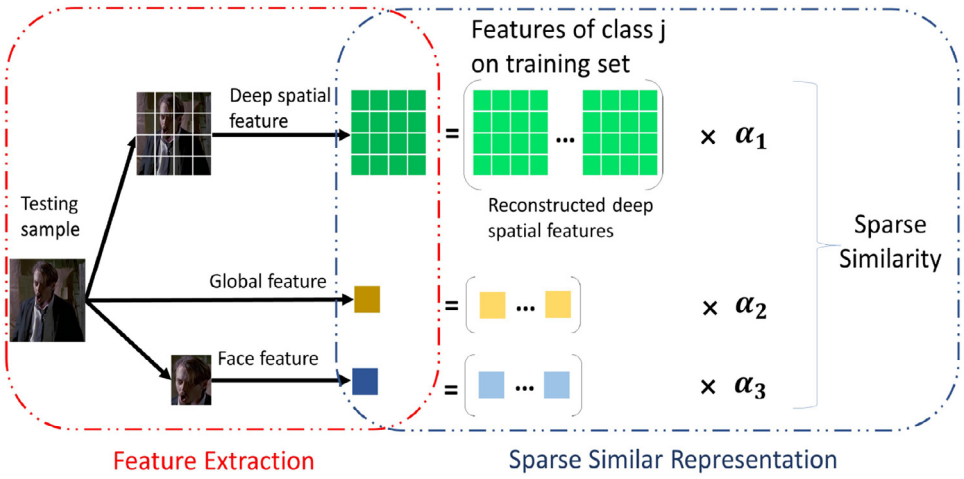


Fig. 4. The flowchart of the proposed MFSSR, which includes feature extraction of three features and sparse similar representation.

weighted distance regularization term

$$\min_{\boldsymbol{\alpha}_k} \omega_k \|\boldsymbol{\alpha}_k - \overline{\boldsymbol{\alpha}}\|_1 \tag{6}$$

where $\overline{\boldsymbol{\alpha}}$ is an unknown variable. Under the influence of ℓ_1-norm on weighted distance regularization, the similarity of coding coefficients for multiple features can be regularized in the feature level with ω_k and in dictionary atom level by using the ℓ_1-norm distance. ω_k controls the collaboration between different features. For instance, the bigger ω_k is, the higher participation of the k^{th} feature there is in joint representation for classification. By using the ℓ_1-norm distance, lots of elements of coding vector $\boldsymbol{\alpha}_k$ will be close to those of $\overline{\boldsymbol{\alpha}}$, except for little elements with too different values in atom level. On one hand, Eq. (6) will enhance the similarity of different features at the aspect of coding. On the other hand, the coefficients associated with the dictionary atom of outliers are selectively ignored to some extent.

Fig. 4 shows the flowchart of MFSSR. The left part describes the extraction of deep spatial features, global features, and face features, and the right part shows the matching similarity between a testing sample and class j based on the sparse similar representation. From Eq. (5) and Fig. 4, it can be observed that there are several advantages of MFSSR to tackle person identification. First, the three discriminative features are jointly used in our method. For example, the proposed reconstructed deep spatial feature is very effective for partial body appearance problems. Second, the three features provide complementary information to identify a person, which is more accurate than that with a single feature. The complementary information can recover the identity from several aspects such as the global clothing characteristics, face characteristics, and partial appearance characteristics, which can effectively solve the challenges in the iQIYI-VID dataset. What's more, with the ℓ_1-norm weighted distance regularization, we can enhance the similarity between different features and decrease the influence of outliers in the coding stage, which will improve the performance of MFSSR in the final classification.

3.3. Optimization algorithm

With predefined values of ω_k, the objective function Eq. (5) can be solved by alternatively updating $\bar{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}_k$.

Initialization: Firstly, we need to extract three types of features with pre-trained models. In particular, for the deep spatial feature, the reconstruction coefficient matrix $\mathbf{W}$ in Eq. (2) needs to be calculated, and we use an iterative method to optimize it. By removing the spatial location information of Eq. (2), a closed-form solution of $\mathbf{W}$ can be obtained, which is used as the initial value of $\mathbf{W}$ in the iteration. The iterative process of $\mathbf{W}$ is shown as following:

$$\begin{aligned} \mathbf{W^0} &= (\mathbf{Y}^T\mathbf{Y} + \lambda\mathbf{I})^{-1}(\mathbf{Y}^T\mathbf{X}) \\ \mathbf{W^n} &= (\mathbf{Y}^T\mathbf{Y} + \lambda\mathbf{I})^{-1}(\mathbf{Y}^T\mathbf{X} - \gamma\mathbf{C}^X\mathbf{W^{n-1}}(\mathbf{C}^Y)^T)), n = 1, 2, \cdots \end{aligned} \tag{7}$$

Then, we initialize $\boldsymbol{\alpha}_k$ by removing the ℓ_1-norm weighted distance regularization term and directly compute the closed-form solution of coding coefficients for each feature. When we remove the

ℓ_1-norm weighted distance regularization, Eq. (5) becomes

$$\min_{\boldsymbol{\alpha}_k} \|\boldsymbol{y}_k - \boldsymbol{D}_k\boldsymbol{\alpha}_k\|_2^2 + \lambda\|\boldsymbol{\alpha}_k\|_2^2 \tag{8}$$

which is a standard collaboration representation formulation. By requiring the derivative of collaboration representation formulation to be zero, an equation for finding a closed solution can be obtained for each feature:

$$\boldsymbol{\alpha}_k = (\boldsymbol{D}_k^T\boldsymbol{D}_k + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}_k^T\boldsymbol{y}_k \tag{9}$$

where $\boldsymbol{I}$ is an identity matrix.

Alternative Step 1: After the coding coefficients of different features are obtained, we can solve the unknown center (i.e., $\overline{\boldsymbol{\alpha}}$) of all $\boldsymbol{\alpha}_k$. Now Eq. (5) becomes

$$\min_{\overline{\boldsymbol{\alpha}}} \sum_{k=1}^{K} \omega_k\|\boldsymbol{\alpha}_k - \overline{\boldsymbol{\alpha}}\|_1 \tag{10}$$

Then we introduce a equation of $\boldsymbol{b}_k = \boldsymbol{\omega}_w\boldsymbol{\alpha}_k - \boldsymbol{\omega}_w\overline{\boldsymbol{\alpha}}$. By relaxing the equation of $\boldsymbol{b}_k$, the ℓ_1-norm weighted distance regularization of Eq. (10) changes to the following equation

$$\min_{\boldsymbol{b}_k} \sum_{k=1}^{K} \tau\|\boldsymbol{b}_k\|_1 + \|\omega_w\boldsymbol{\alpha}_k - \omega_w\overline{\boldsymbol{\alpha}} - \boldsymbol{b}_k\|_2^2 \tag{11}$$

where τ is a relaxation factor.

To solve Eq. (11), we first initialize $\overline{\boldsymbol{\alpha}}$ with the median of all $\boldsymbol{\alpha}$s, then keep $\overline{\boldsymbol{\alpha}}$ unchanged, and finally solve the closed solution of the ℓ_1-norm weighted distance regularization by discussing the case of $\boldsymbol{b}_k$ as below:

$$b_{k,j} = \begin{cases} \omega_w\alpha_{k,j} - \omega_w\overline{\alpha}_j - \frac{\tau}{2}, & (\omega_w\alpha_{k,j} - \omega_w\overline{\alpha}_j) \geq \frac{\tau}{2} \\ 0, & otherwise \\ \omega_w\alpha_{k,j} - \omega_w\overline{\alpha}_j + \frac{\tau}{2}, & (\omega_w\alpha_{k,j} - \omega_w\overline{\alpha}_j) \leq -\frac{\tau}{2} \end{cases} \tag{12}$$

where $b_{k,j}, \alpha_{k,j}$, and $\overline{\alpha}_j$ are the j^{th} entries of $\boldsymbol{b}_k, \boldsymbol{\alpha}_k$, and $\overline{\boldsymbol{\alpha}}$, respectively. After getting $\boldsymbol{b}_k$, we can calculate the value of $\overline{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_k - \boldsymbol{b}_k/\omega_k$.

Alternative Step 2: We add the ℓ_1-norm weighted distance regularization back to Eq. (5). To optimize Eq. (5), we can rewrite $\boldsymbol{\alpha}_k - \overline{\boldsymbol{\alpha}}$ as $\boldsymbol{\alpha}'_k$. And the objective function Eq. (5) becomes

$$\min_{\boldsymbol{\alpha}'_k} \sum_{k=1}^{K} (\|\boldsymbol{y}_k - \boldsymbol{D}_k(\boldsymbol{\alpha}'_k + \overline{\boldsymbol{\alpha}})\|_2^2 + \lambda\|\boldsymbol{\alpha}'_k + \overline{\boldsymbol{\alpha}}\|_2^2 + \tau\omega_k\|\boldsymbol{\alpha}'_k\|_1) \tag{13}$$

There are many methods to tackle this ℓ_1-norm minimization problem. In this paper, we use the Iterative Projection Method [44] (IPM) to optimize Eq. (13). At last, when $\boldsymbol{\alpha}'_k$ is solved, the coding vector $\boldsymbol{\alpha}_k$ equals to $\boldsymbol{\alpha}'_k + \overline{\boldsymbol{\alpha}}$.

Alternatively run Step 1 and Step 2: As Step 1 and Step 2 are iteratively executed, the coding vector center $\tilde{\boldsymbol{\alpha}}$ and the coding vector $\boldsymbol{\alpha}_k$ will become stable. When the iteration number is larger than the maximal number or the vectors are stable enough, the alternative optimization will stop.

Convergence analysis: Eq. (5) is divided into two parts, Eq. (8) and ℓ_1-norm part. In Eq. (8), let $\boldsymbol{f}(\boldsymbol{\alpha}_k) = \|\boldsymbol{y}_k - \boldsymbol{D}_k\boldsymbol{\alpha}_k\|_2^2 + \lambda\|\boldsymbol{\alpha}_k\|_2^2$. Since $\nabla\boldsymbol{f}(\boldsymbol{\alpha}_k)$ satisfies the L-Lipschitz condition, $\boldsymbol{f}(\boldsymbol{\alpha}_k)$ is approximated by a second-order Taylor expansion around $\boldsymbol{\alpha}''_k$. That is $\hat{\boldsymbol{f}}(\boldsymbol{\alpha}_k) \simeq \boldsymbol{f}(\boldsymbol{\alpha}''_k) + \langle\nabla\boldsymbol{f}(\boldsymbol{\alpha}_k), \boldsymbol{\alpha}_k - \boldsymbol{\alpha}''_k\rangle + \frac{L}{2}\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}''_k\|^2 = \frac{L}{2}\|\boldsymbol{\alpha}_k - (\boldsymbol{\alpha}''_k - \frac{1}{L}\nabla\boldsymbol{f}(\boldsymbol{\alpha}''_k))\|_2^2 + const$, where $const$ is a constant unrelated to $\boldsymbol{\alpha}_k$ and $\langle\cdot,\cdot\rangle$ indicates inner product. If $\boldsymbol{f}(\boldsymbol{\alpha}_k)$ is minimized by the gradient descent method, then each iteration is actually equivalent to minimizing the quadratic function $\hat{\boldsymbol{f}}(\boldsymbol{\alpha}_k)$. When this idea extended to Eq. (5), each iteration should be:

$$\boldsymbol{\alpha}_{k+1} = \min_{\boldsymbol{\alpha}_k} \frac{L}{2}\|\boldsymbol{\alpha}_k - \boldsymbol{z}\|_2^2 + \tau\omega_k\|\boldsymbol{\alpha}_k - \overline{\boldsymbol{\alpha}}_k\|_1 \tag{14}$$

where $\boldsymbol{z} = \boldsymbol{\alpha}''_k - \frac{1}{L}\nabla\boldsymbol{f}(\boldsymbol{\alpha}''_k)$. Let $\boldsymbol{\alpha}_k^i$ denote the ith component of $\boldsymbol{\alpha}_k$. Expand the Eq. (14), we can find that there is no item as $\boldsymbol{\alpha}_k^i\boldsymbol{\alpha}_k^j (i \neq j)$ in the formula. That is, the components of $\boldsymbol{\alpha}_k$ do not affect each other. Therefore, the optimization algorithm is convergent.

3.4. Classification

The classification rule is based on the lowest total reconstruction error for each class. When a query sample $\boldsymbol{y}$ comes, we can compute its coding coefficients for all classes. Then, the overall reconstruction error of class j is computed as follow.

$$e_j = \sum_{k=1}^{K} \omega_k\|\boldsymbol{y}_k - \boldsymbol{D}_k^j\boldsymbol{\alpha}_k^j\|_2^2 \tag{15}$$

where $\boldsymbol{D}_k^j$ is the dictionary of class j, and $\boldsymbol{\alpha}_k^j$ is the coding vector associated with $\boldsymbol{D}_k^j$. At last, the label of $\boldsymbol{y}$ is assigned by the class with minimum error:

$$identity(\boldsymbol{y}) = arg\min_j \{e_j\} \tag{16}$$

4. Experiments

In this section, we conduct two groups of experiments. In the first group, to evaluate the effectiveness of the reconstructed deep spatial features, we apply the features to person identification based on partial body appearance. DSR [8] will be compared with our method that inputs the reconstructed deep spatial features of multiple training samples. Furthermore, we will compare the reconstructed features with that extracted by the state-of-art model AlignedReID [11] in ReID task and Part-based Convolutional Baseline (PCB) [1], to verify the reconstructed features are more appropriate for identifying a person with partial body appearance.

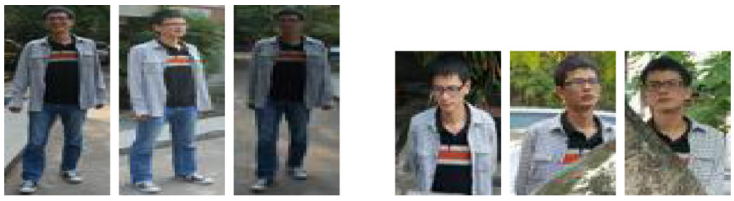
In the second group, we perform multi-feature person identification experiments. In these experiments, CRC [41] is used as the baseline, because it has no joint representation regularization to fuse different features in the coding phase. RCR [17] is used to compare with the proposed MFSSR, to verify our ℓ_1-norm distance regularization is more effective than ℓ_2-norm used in RCR. In addition, we also compared MSFFR with the baseline method on the iQIYI-VID dataset proposed in [12], and verified the effectiveness of this model.

4.1. Experiment settings

Features. In our experiments, to identify a person, we use three types of features, including face feature, reconstructed deep spatial feature, and global feature. Face feature is the most reliable feature for person identification, which is extracted by the state-of-art face recognition model ArcFace [7]. The deep spatial feature is firstly extracted by the same deep network as DSR [8], then encoded by the proposed spatial correlation coding, and finally, represented as the reconstructed deep spatial feature, which is considered as the body feature. As for global feature, considering the images of the same person are similar to some extent, we directly extract image features by ResNet-50 [20] that is pre-trained on ImageNet [43] dataset.

Datasets. Three person datasets are used in our experiments. They are Partial-REID dataset [9], iQIYI-VID [12], and CSM [45].

Partial-REID is a person dataset designed for the task of partial re-identification. It contains 600 images belonging to 60 persons. Each person has 5 full body images and 5 partially occluded images. The images are collected at a university campus with different viewpoints, backgrounds, and different types of severe occlusions. Some samples of the Partial-REID dataset are shown in Fig. 5. For our person identification task, we select the full body images for training and the partial ones for testing.



(a) (b)

Fig. 5. Samples of the gallery set (a) and the query set (b) in Partial-REID dataset.

(a) (b)

Fig. 6. Samples of the training set (a) and the validation set (b) in iQIYI-VID dataset.

Fig. 7. Samples of one cast in CSM dataset.

IQIYI-VID dataset [12] is one of the largest celebrities video datasets designed for multi-feature person identification, which includes more than 500 thousand video clips for about 5000 celebrities. Some frames of different video clips are shown in Fig. 6. The face features of iQIYI-VID dataset [12], extracted by ArcFace model [7], are provided with the dataset. We use a subset of iQIYI-VID dataset in our experiments. The subset consists of 125 celebrities with 10 videos in the training set, and these 1250 videos are used as our training data. In addition, the videos of the same celebrities in the validation set are served as testing data (760 videos in total). Please note that we perform face detection on each frame of each video clip, take the largest face frame and its corresponding confidence, calculate the quality score of the face quality assessment, and use the one with the highest quality score as the clearest face image.

CSM (Cast Search in Movies) [45] is a large-scale person search dataset, which contains 127 thousand manually annotated tracklets of 1218 cast identities from 192 movies (see Fig. 7). A subset, consisting of 53 persons with no less than 17 tracklets in the first ten movies, is used in our experiments. Considering the images in a tracklet are very similar, we simply choose the first image in each tracklet. We randomly selected 15 pictures of each label as training samples, with half of the remaining images as verification samples and the other half as test samples. Finally, 795 pictures are obtained for the training set, 1383 pictures for the verification set, and 1383 pictures for the test set.

Algorithm 1 The optimization of $\boldsymbol{\alpha}'_k$.

1: **Input:** $\lambda, \tau, \bar{\boldsymbol{\alpha}}, \omega_k > 0$

2: **Initialization,** $\widetilde{\boldsymbol{\alpha}}'^{1}_k = \boldsymbol{\alpha}_k - \bar{\boldsymbol{\alpha}}, h = 1.$

3: **while** not converged or the maximal iteration number are not reached **do**

$h = h + 1$

$\widetilde{\boldsymbol{\alpha}}'^{h}_k = \mathbf{S}_{\tau\omega_k/\sigma}\left(\widetilde{\boldsymbol{\alpha}}'^{h-1}_k - \frac{1}{2\sigma}\nabla\mathbf{F}(\widetilde{\boldsymbol{\alpha}}'^{h-1}_k)\right)$

where $\nabla\mathbf{F}(\widetilde{\boldsymbol{\alpha}}'^{h-1}_k)$ is the derivate of $\|\mathbf{y}_k - \mathbf{D}_k(\boldsymbol{\alpha}'_k + \bar{\boldsymbol{\alpha}})\|_2^2 + \lambda\|\boldsymbol{\alpha}'_k + \bar{\boldsymbol{\alpha}}\|_2^2$, and $\mathbf{S}_{\tau\omega_k/\sigma}$ is a soft threshold operator defined in [44].

4: **return:** $\boldsymbol{\alpha}'_k = \widetilde{\boldsymbol{\alpha}}'^{h}_k$.

Table 1
Recognition rates with different training samples on the dataset of CSM .

N_{tr}	3	6	9	12	15
DSR	11.57%	16.49%	18.58%	21.04%	24.30%
reconstructed	**20.73%**	**27.21%**	**32.94%**	**37.05%**	**41.79%**

Table 2
Recognition rates based on deep spatial feature .

Method	Partial-REID	iQIYI-VID	CSM
PCB[1]	46.7%	11.5%	9.9%
AlignedReID[11]	53%	18.15%	20.67%
DSR[8]	54.67%	20.53%	24.30%
AlignedReID[11]+CRC[41]	57.33%	28.29%	34.71%
DSR[8]+CRC[41]	58.00%	31.84%	39.55%
reconstructed+CRC[41]	**59.39%**	**37.05%**	**46.23%**

4.2. Effectiveness of reconstructed deep spatial feature

To evaluate the effectiveness of the reconstructed deep spatial features, we firstly conduct an experiment on the subset of CSM [45] with a different number of training samples per subject. We randomly select 3, 6, 9, 12, and 15 images per subject for training, and the testing ones remain unchanged. The collaborative representation classifier(CRC) [41] with reconstructed features as input in Eq. (4) is compared with DSR(multi-shot). The experimental result is shown in Table 1. With the increase of training samples per subject, our method based on the reconstructed feature gets 9.16%, 10.72%, 14.36%, 16.01%, and 17.49% improvements, respectively. It shows that the reconstructed feature is effective for our method to get better performance than DSR [8] in person identification.

Further, we compare our reconstructed features with those extracted by AlignedReID [11] model and PCB [1], which are both trained on Market1501 [40] in our experiments. And these features are also classified by CRC. The experimental results on Partial-REID [9], iQIYI-VID [12], and CSM [45] datasets are listed in Table 2. First of all, it can be seen that the performance of DSR is bet-

ter than PCB, with 7.97%, 9.03%, and 14.40% improvements on the three datasets, respectively. The reason is that DSR is specially designed for tackling partial person re-identification. Compared with AlignedReID, the performance of DSR is also better, with 1.67%, 2.42%, and 3.63% improvement on the three datasets, respectively. Besides, CRC [41] with the proposed reconstructed deep spatial features is more powerful than AlignedReID [11] and DSR[8]. Especially, CRC with the proposed reconstructed features as input outperforms DSR on Partial-REID, iQIYI-VID, and CSM datasets with 4.72%, 16.52%, and 21.93% improvements, respectively. We can also see that the CRC with reconstructed deep spatial features is better than the features extracted by AlignedReID, increasing by 2.06%, 8.76%, and 11.52% on Partial-REID, iQIYI-VID, and CSM respectively. With the same classifier of CRC, our proposed reconstructed deep spatial features outperform the DSR features by 1.39%, 5.21%, and 6.68% on Partial-ReID, iQIYI-VID, and CSM, respectively. It proves that the reconstructed features are more effective than previous features of partial body appearance in dealing with person identification, and the proposed spatial correlation coding can effectively use different partial features to improve the representation discrimination.

Among the experiments on three datasets, the CSM dataset gets the most significant improvement since there are 15 training samples per subject in this experiment while the number of those on Partial-REID and iQIYI-VID is relatively small. Obviously, to some extent, the dictionary with more reconstructed deep spatial features as its atoms brings more improvements in person identification because of more complementary information from multiple training samples. It substantiates that the reconstructed features that represent the whole body appearance are effective in our method, which combines multiple training samples from the same class in the dictionary for person identification. As for Partial-REID dataset, the improvement is not obvious in the experiment, compared with those on the other datasets. Only a few training samples per subject is one reason. What's more, the gallery samples of the same person are very similar to each other and have little extra information to combine their reconstructed deep spatial features.

4.3. Multi-feature person identification

In this subsection, we apply the proposed MFSSR to multi-feature person identification on the datasets of iQIYI-VID [12] and CSM [45]. In the comparison, we randomly divided each dataset into a training set and a test set and the experiments of all methods are repeated 20 times. Collaborative representation classifier (CRC) [41], which has no regularization term between coding coefficients of different features, is used as the baseline. Relaxed collaborative representation model [17] is compared with the proposed method MFSSR, to verify the effectiveness of our ℓ_1-norm weighted distance regularization in exploiting the similarity and distinctiveness between different features.

4.3.1. Parameters settings discussion

MFSSR has four parameters that need to be adjusted, which are γ, λ, τ, and ω_k as we can see in Eq. (2) and Eq. (5). γ is a parameter that controls the proportion of spatial position information. λ serves to balance the contribution between reconstruction error and coefficient value, ω_k is used to adjust the effect of multiple features on the representation center, and τ is used to balance the contribution between the joint representation and the ℓ_1-norm weighted distance regularization. Changes in γ do not affect the results too much. We evaluate the sensitivity of our model to the changes in γ by empirically setting the values of γ to 0.001, 0.005, 0.01, and 0.015. The final person recognition rates are 90.23%, 90.53%, 90.37%, and 90.34%, respectively. In general, the model is less affected by the change of gamma, and the accuracy rate obtained is relatively stable. Therefore, we chose the best result and set λ to 0.005. Based on the recognition rates of different features, we empirically find the weight values of the three features we used. The recognition rates of different features on a particular data set are detailedly described in the following section. The weights of reconstructed deep spatial feature, global feature, and face feature are empirically set to 0.3, 0.2, and 1, respectively.

Table 3
Recognition rates of different λ.

λ	0.0005	0.0010	0.0015	0.0020	0.0025
CSM	66.62%	66.42%	**66.79%**	66.52%	66.73%
iQIYI-VID	**90.74%**	90.43%	90.56%	90.41%	90.26%

Table 4
Recognition rates of different τ.

τ	0.0005	0.0010	0.0015	0.0020	0.0025
CSM	66.36%	**66.79%**	66.70%	66.14%	66.12%
iQIYI-VID	90.69%	**90.74%**	90.44%	90.32%	90.21%

Table 5
Recognition rates of different features on iQIYI-VID .

Feature	Reconstructed deep spatial feature	Global feature	Face feature
accuracy	36.35%	35.57%	85.83%

To find the optimal combination of parameters in the model, we apply MFSSR to multi-feature person identification on the validation set of CSM [45] by tuning the parameter values. For iQIYI-VID, the videos of the same celebrities as the training data in the validation set are served as testing data (760 videos in total). The optimal values of all parameters are determined on the validation set of the CSM dataset and used in the test set of the CSM and iQIYI-VID. In particular, for the parameter λ, since the dictionary becomes larger on the iQIYI-VID dataset, the dimensionality of coding coefficient becomes larger and the corresponding L2-norm value increases, so we empirically reduce the value of λ from 0.0015 to 0.0005. We first tune the value of λ in the set of $\{0.0005, 0.0010, 0.0015, 0.0020, 0.0025\}$ by fixing τ as 0.001, which is usually a default value. The identification rates in different settings of person recognition on iQIYI-VID [12] and CSM [45] are listed in the Table 3. It can be observed when λ is taken as 0.0015 on CSM [45], the correct rate is the highest. Through experiments we found that MFSSR on the iQIYI-VID dataset is not sensitive to changes in the λ, and it is close to the optimal value when λ is 0.0005. Then we fix λ to be 0.0015 and 0.0005, and change the value of τ in the set of $\{0.0005, 0.0010, 0.0015, 0.0020, 0.0025\}$. The correct rates of multi-feature recognition on IQIYI [12] and CSM [45] are shown in the Table 4. When τ is taken 0.0010, the correct rate is the highest in the CSM dataset. Referring to the above results, the tau value is also set to 0.0010 on the iQIYI-VID dataset. In the following experiments, the parameters λ and τ are empirically set as 0.0015 and 0.0010 on CSM [45], while λ and τ are empirically set as 0.0005 and 0.0010 on iQIYI-VID [12].

4.3.2. IQIYI-VID person identification

In the section, we evaluate our proposed MFSSR on the dataset of iQIYI-VID. The three features are the reconstructed deep spatial feature, face feature, and global feature, which have been introduced in the previous subsection.

We first evaluate the performance of different features (e.g., reconstructed deep spatial feature, global image feature, and face feature) with the baseline classifier of CRC on this dataset. The recognition rates of each type of feature through CRC [41] are listed in Table 5. It can be seen that face recognition rate 89.36% is much

Table 6
The mean and the standard deviation of multi-feature person identification accuracy on iQIYI-VID by randomly dividing the training set and testing set .

Methods	S&G	S&F	G&F	S&G&F
baseline[12]	6.63%(0.78)	87.94%(0.60)	88.95%(0.44)	88.20%(0.62)
CRC[41]	36.42%(0.69)	85.84%(0.74)	86.71%(0.75)	87.69%(0.14)
RCR[17]	38.74%(0.40)	89.95%(0.10)	89.96%(0.28)	89.79%(0.25)
MFSSR	**41.29%**(0.56)	**90.39%**(0.01)	**89.99%**(0.35)	**90.42%**(0.11)

higher than those of the other two features. One reason is that in these videos there are much fewer obvious occlusions in face images than those in body appearance. In addition, the face image feature has less intra-class variation than body appearance and is more discriminative with current deep face models (e.g., ArcFace [7]) for person identification than the whole image feature.

Then, we combine different types of features to conduct multi-feature person identification experiments. In our experiments, in addition to comparing with the related methods CRC [41] and RCR [17], we also compared with the baseline approach [12] on the iQIYI-VID dataset. This baseline uses average pooling and NetVLAD [46] to process the individual features, concatenating the individual features and using multi-modal attention to process the fused features before using a classifier to obtain the final result. Through experiments under twenty different training set and test set settings, we calculated the mean accuracy and the standard deviation of the accuracy of each method as shown in Table 6, where *S, G* and *F* represent reconstructed deep spatial feature, global feature and face feature, respectively. All methods achieve better results with multiple features, which shows that multiple features always provide more information to improve person identification accuracy. Furthermore, CRC does not differ much from face features alone when they are fused with other features, in contrast to MSFFR, which allows the individual features to perform better when they are fused. It can also be seen that CRC is worse than RCR and MFSSR since it does not exploit the similarity between different features in the coding stage. With the weighted distance regularization in the coding stage, both RCR [17] and MFSSR outperform CRC [41] on each combination of different features, with at least 2.55%, 0.44%, 0.03%, and 0.63% improvement on S&G, S&F, G&F, and S&G&F, respectively. MFSSR achieves the best results in all cases. Especially, MFSSR has 0.69% improvement over RCR on the combination of the three features. MFSSR also has better performance compared to the multi-feature baseline approach [12] proposed for the iQIYI-VID dataset, with a 2.22% improvement in the fusion of three features. In [12], the main contribution of the baseline approach is to combine NetVLAD [46] and multi-feature attention, and focuses on the fusion of other features on the basis of facial features to assist in completing the identity recognition. The model without facial features has weak discrimination, so the results obtained under the experimental S&G setting are much lower than the results obtained under the setting containing facial features. Such results are also present in the CSM dataset.

Since there are several sets of experiments with small gaps in Table 6, we conduct the significance hypothesis test in the experiment by randomly dividing the training set and the test set from the dataset. We set the confidence interval to 5% and conduct a one-sided significance hypothesis test. Through the results on Table 7, our model has significant improvements over other methods under all different settings except for RCR under the G&F setting. One possible reason is that facial feature is much more powerful than global feature and global feature can only bring few discrimination. This validates that our ℓ_1-norm weighted distance regularization is more effective to combine multiple features.

Table 7
The result of significance hypothesis test with MFSSR on iQIYI-VID by randomly dividing the training set and testing set, which 1 means that there is a significant difference from MFSSR, otherwise it is expressed as 0.

Methods	S&G	S&F	G&F	S&G&F
baseline[12]	1	1	1	1
CRC[41]	1	1	1	1
RCR[17]	1	1	0	1

Table 8
Recognition rates of different features on CSM .

Feature	Reconstructed deep spatial feature	Global feature	Face feature
accuracy	37.75%	38.15%	59.87%

Table 9
The mean and the standard deviation of multi-feature person identification accuracy on CSM by randomly dividing the training set and testing set .

Methods	S&G	S&F	G&F	S&G&F
baseline[12]	11.36%(0.46)	54.59%(0.47)	54.28%(0.35)	54.27%(0.57)
CRC[41]	41.93%(0.07)	60.83%(0.20)	58.01%(0.14)	61.26%(0.15)
RCR[17]	42.37%(0.37)	61.06%(0.31)	60.23%(0.17)	63.30%(0.12)
MFSSR	**42.53%**(0.55)	**63.85%**(0.29)	**65.21%**(0.29)	**66.50%**(0.28)

Table 10
The result of significance hypothesis test with MFSSR on CSM by randomly dividing the training set and testing set, which 1 means that there is a significant difference from MFSSR, otherwise it is expressed as 0.

Methods	S&G	S&F	G&F	S&G&F
baseline[12]	1	1	1	1
CRC[41]	1	1	1	1
RCR[17]	0	1	1	1

4.3.3. CSM Person identification

As for CSM [45], we extract its reconstructed deep spatial features and global features in the same way as iQIYI-VID [12]. The face features of CSM are also released with the dataset and are extracted by a ResNet-101 [20] trained on MS-Celeb-1M [47].

Likewise, we firstly test different features independently on person identification. From Table 8, we can see that face feature is still more discriminative than the others, followed by reconstructed deep spatial feature and global feature. Compared with iQIYI-VID, the recognition rate based on the face feature on CSM is relatively low. One reason is that the image resolution is much lower than that of iQIYI-VID. Besides, there is severe lighting variation in the CSM dataset.

With multiple features as input, the average accuracy and the standard deviation of accuracy of 20 randomly dividing dataset are shown in Table 9. All methods have improvements on each combination of different features. Without the joint representation regularization to fuse different types of features, CRC is not competitive with MFSSR and RCR. MFSSR achieves the best results over CRC and RCR in all cases. Especially in the combination of three types of features, MFSSR has 12.23%, 5.24% and 3.20% improvements over iQIYI-VID baseline approach, CRC, and RCR, respectively.

Like the iQIYI-VID data set, we set the confidence interval to 5% and conduct a one-sided significance hypothesis test. Through the results on Table 10, except that the RCR method has no significant gap with our model under the S&G setting, our model has significant improvements over other methods under other different settings. This verifies that the proposed ℓ_1-norm weighted dis-

tance regularization used in MFSSR is more effective than that of ℓ_2-norm for coding and classification.

5. Conclusion

In this paper, a multi-feature joint representation method for person identification is proposed from two perspectives of robust feature extraction and multiple-feature classifier. For more robust features to spatial occlusions, we have designed a more powerful body appearance feature, the reconstructed deep spatial feature with spatial correlation coding, which can effectively use the collaborative information of different partial features in a sample. The designed spatial correlation coding has well represented a local region by the combination of other regions with considering the spatial neighborhood information to regularize the coding coefficients. Moreover, we proposed a multi-feature sparse similar model by designing a ℓ_1-norm weighted distance regularization, which not only jointly exploits the discrimination in multiple features (e.g., the reconstructed deep spatial feature, global feature, and face feature), but also allows the representation distinctiveness in dictionary-atom level to handle different sample features (e.g., outliers). Experimental results on several datasets have evaluated the effectiveness of the reconstructed deep spatial feature and demonstrated that our ℓ_1-norm weighted distance regularization is more advantageous to the fusion of different types of features. However, our current work has only focused on the fusion of individual features within a modality of the image, and the feature extraction and multi-feature classifier are not jointly learned. How to fuse features from other modalities such as video and audio and jointly learn an end-to-end multi-feature person identification to further improve the correct identification rate is one of the issues that we need to address in our future work.

Declaration of Competing Interest

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.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (Grant no.62176271, 61972212, and 61772568), the Guangdong Basic and Applied Basic Research Foundation (Grant no.2019A1515012029), the Natural Science Foundation of Jiangsu Province under Grant BK20190089 and Science and Technology Program of Guangzhou (Grant no. 202201011681)..

References

[1] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 480–496.
[2] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, Y. Xu, Deep-person: learning discriminative deep features for person re-identification, Pattern Recognit 98 (2020) 107036.
[3] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C.H. Hoi, Deep learning for person re-identification: a survey and outlook, IEEE Trans Pattern Anal Mach Intell 44 (6) (2021) 2872–2893.
[4] Y. Chen, H. Wang, X. Sun, B. Fan, C. Tang, H. Zeng, Deep attention aware feature learning for person re-identification, Pattern Recognit 126 (2022) 108567.
[5] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans Pattern Anal Mach Intell 31 (2) (2009) 210–227.
[6] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891–1898.
[7] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.
[8] L. He, J. Liang, H. Li, Z. Sun, Deep spatial feature reconstruction for partial person re-identification: alignment-free approach, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7073–7082.
[9] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, S. Gong, Partial person re-identification, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4678–4686.
[10] C. Zhao, X. Wang, W. Zuo, F. Shen, L. Shao, D. Miao, Similarity learning with joint transfer constraints for person re-identification, Pattern Recognit 97 (2020) 107014.
[11] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, J. Sun, Alignedreid: surpassing human-level performance in person re-identification, arXiv preprint arXiv:1711.08184 (2017).
[12] Y. Liu, P. Shi, B. Peng, H. Yan, Y. Zhou, B. Han, Y. Zheng, C. Lin, J. Jiang, Y. Fan, et al., Iqiyi-vid: a large dataset for multi-modal person identification, arXiv preprint arXiv:1811.07548 (2018).
[13] M. Yang, L. Zhang, J. Yang, D. Zhang, Robust sparse coding for face recognition, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, 2011, pp. 625–632.
[14] M. Sadiq, D. Shi, Attentive occlusion-adaptive deep network for facial landmark detection, Pattern Recognit 125 (2022) 108510.
[15] X.-T. Yuan, X. Liu, S. Yan, Visual classification with multitask joint sparse representation, IEEE Trans. Image Process. 21 (10) (2012) 4349–4360.
[16] H. Zhang, N.M. Nasrabadi, Y. Zhang, T.S. Huang, Multi-observation visual recognition via joint dynamic sparse representation, in: 2011 IEEE International Conference on Computer Vision (ICCV 2011), 2011, pp. 595–602.
[17] M. Yang, L. Zhang, D. Zhang, S. Wang, Relaxed collaborative representation for pattern classification, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2224–2231.
[18] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans Pattern Anal Mach Intell 19 (7) (1997) 711–720.
[19] M. Yang, L. Zhang, X. Feng, D. Zhang, Fisher discrimination dictionary learning for sparse representation, in: 2011 IEEE International Conference on Computer Vision (ICCV 2011), 2011, pp. 543–550.
[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[21] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[22] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[23] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1701–1708.
[24] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.
[25] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European Conference on Computer Vision, 2016, pp. 499–515.
[26] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2, 2014, pp. 1988–1996.
[27] X. Chang, Z. Ma, X. Wei, X. Hong, Y. Gong, Transductive semi-supervised metric learning for person re-identification, Pattern Recognit 108 (2020) 107569.
[28] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6738–6746.
[29] W. Liu, Y.-M. Zhang, X. Li, Z. Yu, B. Dai, T. Zhao, L. Song, Deep hyperspherical learning, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 3953–3963.
[30] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-margin softmax loss for convolutional neural networks, in: International Conference on Machine Learning, 2016, pp. 507–516.
[31] F. Wang, J. Cheng, W. Liu, H. Liu, Additive margin softmax for face verification, IEEE Signal Process Lett 25 (7) (2018) 926–930.
[32] X. Zhu, X.-Y. Jing, F. Zhang, X. Zhang, X. You, X. Cui, Distance learning by mining hard and easy negative samples for person re-identification, Pattern Recognit 95 (2019) 211–222.
[33] Q. Meng, S. Zhao, Z. Huang, F. Zhou, Magface: A universal representation for face recognition and quality assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14225–14234.
[34] L. Wolf, T. Hassner, Y. Taigman, Similarity scores based on background samples, in: Proceedings of the 9th Asian Conference on Computer Vision-Volume Part II, 2009, pp. 88–97.
[35] Y. Chen, G. Song, Z. Shao, J. Cai, T.-J. Cham, J. Zheng, Geoconv: geodesic guided convolution for facial action unit recognition, Pattern Recognit 122 (2022) 108355.
[36] Y. Ge, F. Zhu, D. Chen, R. Zhao, H. Li, Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, pp. 11309–11321.

[37] T. He, X. Shen, J. Huang, Z. Chen, X.-S. Hua, Partial person re-identification with part-part correspondence learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9105–9115.
[38] X. Xin, J. Wang, R. Xie, S. Zhou, W. Huang, N. Zheng, Semi-supervised person re-identification using multi-view clustering, Pattern Recognit 88 (2019) 285–297.
[39] C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, Multi-type attributes driven multi–camera person re-identification, Pattern Recognit 75 (2018) 77–89.
[40] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.
[41] L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: Which helps face recognition? in: Proceedings of the 2011 International Conference on Computer Vision, 2011, pp. 471–478.
[42] X. Ren, D. Zhang, X. Bao, Semantic-guided shared feature alignment for occluded person re-IDentification, in: Asian Conference on Machine Learning, 2020, pp. 17–32.
[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[44] L. Rosasco, A. Verri, M. Santoro, S. Mosci, S. Villa, Iterative projection methods for structured sparsity regularization, MIT Technical Reports (2009). MIT–CSAIL–TR–2009–050, CBCL–282
[45] Q. Huang, W. Liu, D. Lin, Person search in videos with one portrait through visual and temporal links, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 425–441.
[46] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, J. Sivic, NetVLAD: CNN architecture for weakly supervised place recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5297–5307.
[47] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, in: European Conference on Computer Vision, 2016, pp. 87–102.

Meng Yang is currently an associate professor at School of Data and Computer Science, Sun Yat-sen University. He received his Ph.D. degree from The Hong Kong Polytechnic University in 2012. Before joining Shenzhen University, he has been working as Postdoctoral fellow in the Computer Vision Lab of ETH Zurich. His research interest includes sparse coding, dictionary learning, object recognition and machine learning. He has published more than 70 academic papers, including more than 10 CVPR/ICCV/ECCV/ICML/AAAI papers and several IJCV, IEEE TNNLS and TIP journal papers. Now his Google citation is over 10000.

Lei Liao is currently a graduate student at School of Data and Computer Science, Sun Yat-sen University.

Kangyin Ke was currently a graduate student at School of Data and Computer Science, Sun Yat-sen University.

Guangwei Gao received the Ph.D. degree in pattern recognition and intelligence systems from Nanjing University of Science and Technology, Nanjing, China, in 2014. Now, he is an associate professor in the Institute of Advanced Technology, Nanjing University of Posts and Telecommunications. His research interests include face recognition, face hallucination and biometrics.