Signal Processing 144 (2018) 296–305

Contents lists available at ScienceDirect

Signal Processing

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

Feature fusion with covariance matrix regularization in face recognition

Ze Lu*, Xudong Jiang, Alex Kot

Nanyang Technological University, 50 Nanyang Drive, 637553, Singapore

ARTICLE INFO

Article history:
Received 24 August 2017
Accepted 22 October 2017

Keywords:
Feature fusion
CNN
Overfitting
Regularization
Face recognition

ABSTRACT

The fusion of multiple features is important for achieving state-of-the-art face recognition results. This has been proven in both traditional and deep learning approaches. Existing feature fusion methods either reduce the dimensionality of each feature first and then concatenate all low-dimensional feature vectors, named as DR-Cat, or the vice versa, named as Cat-DR. However, DR-Cat ignores the correlation information between different features which is useful for classification. In Cat-DR, on the other hand, the correlation information estimated from the training data may not be reliable especially when the number of training samples is limited. We propose a covariance matrix regularization (CMR) technique to solve problems of DR-Cat and Cat-DR. It works by assigning weights to cross-feature covariances in the covariance matrix of training data. Thus the feature correlation estimated from training data is regularized before being used to train the feature fusion model. The proposed CMR is applied to 4 feature fusion schemes: fusion of pixel values from 3 color channels, fusion of LBP features from 3 color channels, fusion of pixel values and LBP features from a single color channel, and fusion of CNN features extracted by 2 deep models. Extensive experiments of face recognition and verification are conducted on databases including MultiPIE, Georgia Tech, AR and LFW. Results show that the proposed CMR technique significantly and consistently outperforms the best single feature, DR-Cat and Cat-DR.

1. Introduction

Face recognition has been a very active research area due to its increasing security demands, commercial applications and law enforcement applications [1–6]. It is often the case in face recognition that no single feature is rich enough to capture all of the available information [7]. The robust face recognition requires multiple feature sets to be taken into account [8], which can be features of different color channels [9–12], different types of features [8,13,14] and features extracted by different deep models [15–17]. Recently, Convolutional Neural Networks (CNN) provides an effective tool for feature learning in face recognition and very promising results have been obtained as in [18,19]. The pre-trained VGG-Face model [18] was learned from a large face dataset containing 2.6M web images of 2622 celebrities and public figures. It is widely used as a feature extractor for classifying face images as in [20–22]. Different from the architecture of VGG-Face, ResNet in [19] consists of residual modules which conduct additive merging of signals. The authors in [19] argue that residual connections are inherently important for training very deep architectures. It is natural to study the combination of VGG-Face with ResNet, which would allow two models to reap the benefits of each other. Thus we train a ResNet-like CNN model using images from the recently released CASIA-WebFace dataset [23] and combine it with the pre-trained VGG-Face model by feature fusion.

Feature fusion often results in very high dimensionality. For example, multi-scale descriptors in [24] are densely extracted from dense landmarks and concatenated together to form a 100K-dimensional feature vector. The high dimensionality of feature vectors imposes great burdens on the robust face recognition task. Therefore, dimensionality reduction is a critical module of feature fusion. Existing feature fusion methods can be generally classified into two categories: DR-Cat and Cat-DR. DR-Cat applies dimensionality reduction to each feature before the concatenation of multiple features and Cat-DR does vice versa. Choi and et al. [11] use DR-Cat to reduce the dimension of each color local texture feature separately before concatenating all low-dimensional features in the column order. Tan and et.al [13] use PCA to reduce the dimensionality of Gabor wavelets and LBP prior to fusing them by averaging their similarity scores (same as DR-Cat). DR-Cat is also used in [12,25–27]. By reducing the dimensionality of each feature separately before concatenating them together, DR-Cat ignores the correlation

* Corresponding author.
E-mail addresses: zlu008@ntu.edu.sg (Z. Lu), exdjiang@ntu.edu.sg (X. Jiang), eackot@ntu.edu.sg (A. Kot).

https://doi.org/10.1016/j.sigpro.2017.10.024
0165-1684/

information between different features. But the correlation information plays an important role in the process of feature fusion. In order to utilize the correlation information, Yang and et.al [28] employ Cat-DR to concatenate three color components into one pattern vector first and then perform PCA or EFM on the concatenated pattern vector. Cat-DR is also used in [24] to fuse multi-scale descriptors centered at dense facial landmarks. The dimension of the concatenated feature is reduced by PCA and LDA. Multiple deep ConvNets are used in [15] to learn face features from images of various scales, where Cat-DR is employed by applying PCA to the concatenation of multiple features. In the case of perfect training data, Cat-DR utilizing the correlation information usually achieves better performance than DR-Cat. However, in practice, the limited training data may result in unreliable estimates of cross-feature correlations. This often leads to overfitting and performance degradation in Cat-DR.

To solve problems in feature fusion methods of DR-Cat and Cat-DR, we propose a covariance matrix regularization (CMR) technique. Instead of modifying eigenvalues of covariance matrices as in conventional regularization techniques [29–33], CMR works by regularizing the off-diagonal cross-feature covariances in the covariance matrix of training data. Thus the trace of covariance matrices remains unchanged and the feature correlation estimated from the training data is suppressed before being used to train the feature fusion model. In this way, the obtained model does not adapt too much to the estimated correlation and hence the overfitting is reduced. In the experimental part conducted on four public face databases including MultiPIE, GT, AR and LFW, we first show that our proposed ResNetShort model achieves state-of-the-art face verification performance on LFW. After that, we vary the value of weights in CMR to show how it solves the problem of overfitting and improves the face recognition performance. Then, we study the relationship between the optimal value of weights in CMR and the number of training images per subject. Finally, we compare the performance of CMR against the best single feature, DR-Cat and Cat-DR by fusing features of multiple color channels, multiple types of features, and features extracted by multiple deep models.

2. Feature fusion in face recognition

2.1. Feature fusion schemes

Face recognition is an area that is well-suited for the fusion of multiple descriptors due to its inherent complexity and need for fine distinctions [8]. Multiple descriptors can be features extracted from different color channels. Y, I, Q components possess the property of decorrelation, which helps reduce redundancy and is an important property in pattern classifier design. Thus features extracted from Y, I, Q color channels are fused in [9]. Similarly, R, Q, Cr features are fused in [10,11] and Z, R, G features are fused in [12]. Furthermore, multiple descriptors can be different types of features. Authors in [8,13] combine Gabor wavelets and LBP to achieve considerably better performance than either alone. The two features are complimentary in the sense that LBP captures small appearance details while Gabor wavelets encode facial shape over a broader range of scales. Fourier features, Gabor wavelets are combined in [14] to achieve better performance for face recognition. Global Fourier features describe the general characteristics of the holistic face and they are often used for coarse representation. Differently, local Gabor features reflect and encode more detailed variations within some local facial regions. Moreover, multiple features may be extracted using different deep models. Authors in [15] train 60 ConvNets, each of which extracts two 160-dimensional DeepID vectors from 60 face patches with ten regions, three scales, and RGB or gray channels. Combing 60 different deep models increases the face verification accuracy by 5.27% over the best single model. The deep learning structure proposed in [16] is composed of a set of elaborately designed CNN models, which extract complementary facial features from multimodal facial data.

To investigate the effectiveness of the proposed feature fusion method for face recognition, this paper explores 4 different feature fusion schemes: (1), fusion of pixel values in 3 color channels *R, G, B*; (2), fusion of LBP features in 3 color channels *R, G, B*; (3), fusion of pixel values and LBP features of a single color channel *R*; (4), fusion of CNN features extracted by 2 deep models. Many recent face recognition works conduct experiments on pixel values to evaluate the face recognition performance of their methods [34–37]. LBP has been proven to be highly discriminative for face recognition [24,38]. Thus these two features are used for the task of fusing features of different color channels *R, G, B* and the task of fusing different types of features in channel *R*. As *R* channel has been shown to perform better than other intensity images including Gray for face retrieval [11,34], we take the *R* channel as an example channel for the fusion of different types of features. For the fusion of multiple deep learning features, we utilize the pre-trained VGG-Face model and propose a new deep model, ResNet-Short, presented in the following section.

2.2. Deep learning feature fusion: VGG-Face and ResNetShort

Convolutional Neural Networks have significantly improved the state of the arts in face recognition [39]. VGG-Face is a deep neural network proposed by Simonyan et al. in [18]. This network is characterized by using 3×3 convolutional layers stacked on top of each other in increasing depth. The architecture of VGG-Face comprises 21 layers, which consist of 13 convolutional layers, 5 maxpooling layers and 3 fully connected layers. The first two fully connected layers are 4096 dimensional and the dimension of the last fully connected layer depends upon the loss functions used for optimisation. The pre-trained VGG-Face model was learned from a large face dataset (see Fig. 1 for sample images) containing 2.6M images of 2622 celebrities and public figures. Faces are detected using the method described in [40] and a 2D similarity transformation is applied to map the face to a canonical position. VGG-Face is first trained as a multi-class classification problem by minimizing the softmax loss and then fine-tuned by the recently proposed triplet loss [41]. The pre-trained VGG-face model has been widely used by researchers to extract CNN features from face images as in [20–22].

Unlike traditional sequential network architectures such as VGG, ResNet consists of “network-in-network” modules. First introduced by He et al. in [19], ResNet has become a seminal work, demonstrating that the degradation problem of deep networks can be solved through the use of residual modules. ResNet layers are formulated as learning residual functions with reference to the layer inputs. By referring to the CNN model used in [42] and residential modules, we propose a model as shown in Fig. 2 and name it ResNetShort. The size of filters in convolution layers is 3×3 with stride 1, followed by PReLU [43] non-linear units. The max-pooling grid is 2×2 and the stride is 2. The number of feature maps in convolutional layers or the dimension of fully connected layers is indicated by the number on top of each layer. ‘$\times h$’ represents a residual module that repeats for h times. Joint supervision of softmax loss and center loss [42] is adopted. The value of λ, which is used for balancing the softmax and center loss functions, is set as 0.005.

The recently released CASIA-WebFace [23] database is used to train the ResNetShort model. CASIA-WebFace contains 494,414 images of 10,575 subjects. According to [44], adding the individuals with only a few instances do not help to improve the recognition performance. Indeed, these individuals will harm the systems performance. Thus the 10,575 subjects are ranked in the descent or-