Journal of Visual Communication and Image Representation 53 (2018) 13–19

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

A set-to-set nearest neighbor approach for robust and efficient face recognition with image sets[☆,☆☆]

Ling Wang[a], Hong Cheng[b,*], Zicheng Liu[c]

[a] School of Electric Engineering, University of Electronic Science and Technology of China, 2006 Xiyuan Avenue, Chengdu 611731, China
[b] Center for Robotics, University of Electronic Science and Technology of China, 2006 Xiyuan Avenue, Chengdu 611731, China
[c] Microsoft Research Redmond, One Microsoft Way, Redmond, WA 98052, USA

ARTICLE INFO

Keywords:
Face recognition
Set-to-set
Robust analysis
Weighted correlation analysis

ABSTRACT

Set-to-set face recognition has drawn much attention thanks to its rich set information. We propose a robust and efficient Set-to-Set Nearest Neighbor Classification (S2S-NNC) approach for face recognition by using the maximum weighted correlation between sets in low-dimensional projection subspaces. A pair of face sets is represented as two sets of Mutual Typical Samples (MTS) based on their maximum weighted correlation, and the S2S distance is equivalent to that between two sets of MTS. For the variation of objects within a set, the faces are partitioned into patches and projected onto a correlation subspace to find the MTS between two sets. Furthermore, we develop a S2S-NNC approach for image set-based face recognition. Compared with existing approaches, the S2S-NNC unifies the image-to-image, image-to-set and set-to-set recognition problems into one model. Experimental results show the S2S-NNC approach significantly outperforms the state-of-art approaches on large video samples and small occluded samples.

1. Introduction

Video information is wildly used in our life thanks to the development of High Efficiency Video Coding (HEVC) [1–5]. But its also result the computation and accuracy problem of image set-based face recognition in the scenario of video recognition and surveillance [6–9]. For the multi-viewpoints of camera networks or long term observation, the images in the set are usually blurred, occluded or deformed. Traditional image set-based face recognition approaches are not usually robust enough to solve this problem. Furthermore, since the Euclidean distance between two vectors does not work very well to measure the distance between sets, the traditional image-to-image classifiers, such as Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) [10] and Naive Bayesian Nearest Neighbor (NBNN) [11], cannot be directly used in image set-based face recognition.

Image set recognition consists of two basic modules, set representation and set distance metric. In set representation, projection based approaches model each image set as a low-dimensional subspace [12–17], manifold [7,18–22] or affine hull [8]. Subspace/Manifold learning needs sufficient samples to obtain the subspace/manifold of an image set, which is very challenging for many face recognition problems. Thus, insufficient gallery samples could greatly degrade the performance of recognition. To alleviate this problem, an affine hull model with image samples and their mean was used to represent a face set [23]. However, this model does not really solve the small-size sample problem, especially in occluded cases. Moreover, traditional image set-based face recognition approaches suffer high computational complexity. It is hard to construct a robust yet efficient distance metric between image sets. Set-to-set distance metrics usually adopt complex optimization algorithms to evaluate the roles of different samples within a set, thus increasing the complexity of image set recognition approaches.

In this paper, we propose a robust and efficient Set-to-Set Nearest Neighbor Classification (S2S-NNC) approach for face recognition by using the maximum weighted correlation of sets in low-dimensional projection subspaces. Here, a pair of sets are described as Mutual Typical Samples (MTS) based on their maximum mutual correlation. As shown in Fig. 1, given two sets, $\mathbf{X}$ and $\mathbf{Y}$, we can learn their corresponding projection matrices $\mathbf{W}_x$ and $\mathbf{W}_y$ by using the maximum correlation between two sets. Furthermore, the MTS of each set is obtained by extracting principal samples from each image within a set. Then, we reduce the distance metric between two original sets with that between their sets of MTS. Given a testing sample set and the gallery sample sets of C classes, we can generate their MTS set pairs. Finally, we can assign

☆ This work is supported by the grant from "National Natural Science Foundation of China – China (NSFC)" (No. 61603077, No. 61273256, No. 61305033 and No. U1233103F01).
☆☆ This paper has been recommended for acceptance by Zicheng Liu.
* Corresponding author.
E-mail addresses: eewangling@uestc.edu.cn (L. Wang), hcheng@uestc.edu.cn (H. Cheng), zliu@microsoft.com (Z. Liu).

https://doi.org/10.1016/j.jvcir.2018.02.004
Received 14 March 2017; Received in revised form 24 January 2018; Accepted 3 February 2018
Available online 05 February 2018

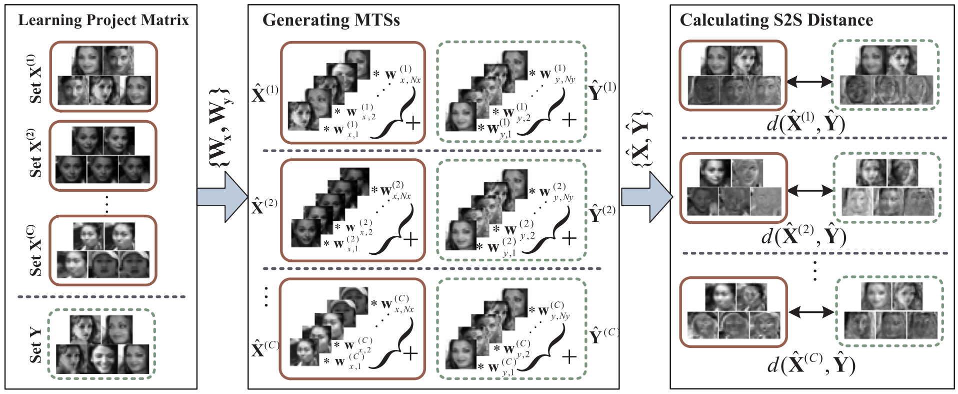

Fig. 1. An illustration of MTS based S2S-NNC. The $\mathbf{X}^{(c)}$ and $\mathbf{Y}$ are two sets of given classes. By computing the maximum correlation coefficient $\rho(\mathbf{X}^{(c)},\mathbf{Y})$, we get the projection matrix $(\mathbf{W}_x^{(c)},\mathbf{W}_y^{(c)})$ and MTS $(\widehat{\mathbf{X}}^{(c)},\widehat{\mathbf{Y}})$. Then S2S distance $d_{\mathrm{MTS}}(\mathbf{X}^{(c)},\mathbf{Y})$ is calculated in this MTS space, With the NN classifier, the recognition result is the minimum distance in $d_{\mathrm{MTS}}(\mathbf{X}^{(c)},\mathbf{Y}),c \in C$.

the testing set to the class with the minimum distance between MTS set pairs. By doing so, the proposed approach can work well even in heavy occlusions without any preprocessing steps or explicit dimensionality reduction. More interestingly, we do not use complex optimization algorithms in set representation and set-to-set distance metric. Even in small size samples, the proposed approach still works very well. We also unify the Image-to-Image (I2I), Image-to-Set (I2S) and Set-to-Set (S2S) distance metric in one framework.

The rest of this paper is organized as follows. We first give an overview of the related work of S2S recognition approach in Section 2. In Section 3, we highlight the image set representation and its distance metric with MTS and patched-MTS. By using these MTS representations, we propose a S2S nearest neighbor classification approach. Comprehensive experiments are presented in Section 4. Finally, we draw conclusions and discuss future work in Section 5.

2. Related work

The existing techniques on image set-based face recognition can, for the most part, be divided into two classes: parametric approaches and non-parameter based approaches.

The parametric approaches usually tend to represent the image set by a parametric model or distribution function. Lee et al. use Principal Component Analysis (PCA) to represent gallery images in low-dimensional appearance subspaces [6] and then cluster the subspaces by K-means algorithm. Arandjelovic et al. propose a semi-parametric model for learning probability densities confined to highly non-linear but intrinsically low-dimensional manifolds [18]. This approach is based on a stochastic approximation of Kullback-Leibler divergence between the estimated densities. Zheng et al. used kernel canonical correlation analysis to solve the facial expression recognition problem [24]. They used Gabor wavelet transformation to convert landmark points of faces into a labeled graph vector to representing the facial features. The limitation of these parametric-based approaches is that if the image set does not have strong statistical correlation for the parameters, the estimated model cannot represent the image set very well.

In contrast, the non-parameter based approaches, usually relax the assumptions on distribution of the data set and are more flexible. One of the important nonparametric approaches is the subspace/manifold based approach. Yamaguchi et al. propose Mutual Subspace Method (MSM) to define the similarity between two image sequences [12]. Kim et al. represent the images by subspaces and recognition is carried out by subspace-to-subspace discrimination matching [25]. Wang et al. propose Manifold Discriminant Analysis (MDA) and Manifold-Manifold Distance (MMD) approaches, by modeling the covariance matrix of the image set as a manifold [7,26,27]. Yang et al. propose a Multi-Manifold Discriminant Analysis (MMDA) method for image feature extraction and pattern recognition [28]. Kim et al. extend the concept of principal angles between linear subspaces to manifolds with arbitrary nonlinearity [29]. Cevikalp et al. represent images as points in a linear or affine feature space [8]. To overcome the problem of local linear or nonlinear model, image sets are mapped into Grassmannian or Riemannian manifold space. And then discriminant analysis or information entropy methods are used for recognition [19–22]. Hu et al. introduce a Sparse Approximate Nearest Points (SANP) approach [23], where the nearest points between two sets are sparsely approximated from the respective set. A joint representation of the image set is defined and included both the sample images and their affine hull models. Yang et al. use a joint regularized nearest points to represent image sets of different classes [15]. A joint feature projection matrix learning and dictionary structuring method is proposed in [17]. Chen et al. propose a multivariate sparse representation for video-to-video face recognition [30]. Cui et al. divide images into patches and sparsely encode them. Then whitened PCA and pairwise constrained multiple metric learning techniques are used to reduce the feature dimension and to integrate the descriptors [31]. Wolf et al. present patch-based Local Binary Pattern (LBP) multiple descriptors to capture statistics of local features in a set [32,33]. Moreover, Lu et al. [16,34] and Vemulapalli et al. [35] model the image set by using set-statistics based approaches. Compared to the parametric-based approaches, the subspace/manifold representation does not need any assumptions about the data distribution. However, most of them need training steps and feature extraction, which require some priori information of the dataset and need man-made feature design.

Compared with image set representation, the problem of set distance metric seems less of a challenge. The maximum posteriori formulation [6], geometric distance function [8,23,26,27,31,36] and discriminative learning [7,28,32,33,37,38] approaches are used in classification. Zhu et al. propose a I2S classification by extending Mahalanobis distance [39]. Huang et al. propose a Euclidean-to-Riemannian metric to solve the I2S classification problem [40]. Zhu et al. propose a convex or regularized hull to collaboratively represent all of the gallery image sets and query sets [14]. The distance metric between query set and gallery sets can be calculated by the represent coefficients, this method is named as RH-ISCRC. Those existing techniques usually are based on Euclidean distance or its deformation, Cosine distance, and Jaccard distance, et al. A main drawback of these measures is that they cannot unify the I2I, I2S and S2S measure in one framework. Furthermore, most of the S2S measures need an iterative process to obtain the optimal solutions, which leads to increasing of computation in the case of large number of samples.

In this paper, some notations of variables are listed in Table 1.