Image and Vision Computing xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Image and Vision Computing

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

Joint prototype and metric learning for image set classification: Application to video face identification☆

Mengjun Leng, Panagiotis Moutafis, Ioannis A. Kakadiaris*

Computational Biomedical Lab Department of Computer Science, University of Houston, 4800 Calhoun Road, Houston, TX77004, United States

ARTICLE INFO

Article history:
Received 9 October 2015
Received in revised form 19 April 2016
Accepted 10 June 2016
Available online xxxx

Keywords:
Image set classification
Metric learning
Prototype learning
Video face recognition

ABSTRACT

In this paper, we address the problem of image set classification, where each set contains a different number of images acquired from the same subject. In most of the existing literature, each image set is modeled using all its available samples. As a result, the corresponding time and storage costs are high. To address this problem, we propose a joint prototype and metric learning approach. The prototypes are learned to represent each gallery image set using fewer samples without affecting the recognition performance. A Mahalanobis metric is learned simultaneously to measure the similarity between sets more accurately. In particular, each gallery set is represented as a regularized affine hull spanned by the learned prototypes. The set-to-set distance is optimized via updating the prototypes and the Mahalanobis metric in an alternating manner. To highlight the importance of representing image sets using fewer samples, we analyzed the corresponding test time complexity with respect to the number of images used per set. Experimental results using YouTube Celebrity, YouTube Faces, and ETH-80 datasets illustrate the efficiency on the task of video face recognition, and object categorization.

1. Introduction

Image set classification has been an active research field for more than twenty years [1–9]. The task is to assign each probe image set to its corresponding gallery subject. Templates stored in the gallery are also sets of images. Both gallery and probe sets contain various numbers of images, describing the same subject. In the field of biometrics, many applications can be formulated as an image set classification problem, such as video-based face recognition [1], gesture recognition [10], and person re-identification across camera networks [2]. Compared with traditional single image classification, the set-based approach provides richer information with multiple samples. Hence, more reliable results are expected. However, it also introduces several new challenges: First, not all the information provided is useful for the task at hand. There is information redundancy or even noise, especially for large scale image sets. Second the within-set variations are large (e.g., different views, illumination conditions, sensors). As a result, building a proper model is crucial. Third, the computational and storage cost are increased significantly with the rapid growth of data to be processed. For example, some videos could be thousands of frames long. To solve the image set classification problem, a straightforward approach is to model the set-to-set distance. The smaller the distance is, the more similar two image sets will be. According to how this distance is modeled, existing literature can be grouped into three categories: (i) subspace model, (ii) statistical model, and (iii) affine hull model.

1.1. Subspace model

Methods in this category can be further grouped into two subcategories: single subspace model and multi-subspace model. In the single subspace model, each image set is modeled as a single linear subspace [3, 4, 5, 11, 12] and can be treated as a point on a Grassmann manifold [3, 4]. Different mutual subspace distances were defined based on the principal angles between subspaces. Linear discriminative analysis [11], non-linear manifold kernels [3, 4], sparse dictionary learning [12], and direct manifold-to-manifold mappings [5] are employed to optimize the distances. However, the single subspace model cannot reflect the importance of different local variations under different scenarios. In the multi-subspace model, each image set is modeled as a mixture of several subspaces [6, 7, 8, 13]. These subspaces can be constructed using clustering algorithms (e.g., k-means clustering [6], hierarchical agglomerative clustering [7], and Maximum Linear Patches [8, 13]). The set-to-set distance is defined as the distance between the closest pair of local subspaces. It can represent the complex local variations in a better

☆ This paper has been recommended for acceptance by Patrick Flynn.

* Corresponding author.
E-mail address: ioannisk@uh.edu (I. Kakadiaris).

http://dx.doi.org/10.1016/j.imavis.2016.06.005
0262-8856/

way. However, computing a multi-subspace model is very expensive and a large amount of data is needed.

1.2. Statistical model

Statistical characteristics are used to model the image sets. It can be further divided into two sub-categories: parametric and non-parametric. In the parametric statistical model, an image set is either modeled as a single Gaussian distribution [9] or a mixture of Gaussian [14]. The Kullback–Leibler divergence [9] or kernel based distance [14] is used to measure the distance between two sets. Methods in this category make strong assumptions concerning the distribution of the data which may not always be true. In non-parametric statistical model, each image set is described using its statistical properties: mean [15, 16], covariance matrix [16, 17], and other higher order statistics [15]. The distance is measured either in a Euclidean space [15, 16] or on a Riemannian manifold [16, 17]. The manifold-to-manifold dimensionality reduction [18] was developed to reduce the cost of computing a high dimensional Riemannian manifold. Multi-metric learning [15, 16] was employed to combine different properties together. The non-parametric statistic model relies only on a few statistical properties. As a result, it is robust, but may ignore significant local variation in the data.

1.3. Affine hull model

Each image set is modeled as an affine hull [19] or different kinds of reduced affine hull [19–21]. The geodesic distance between two hulls is then employed to measure the dissimilarity between sets. Mahalanobis metric is employed [22] for a more accurate dissimilarity measurement. More recently, the correlations between different gallery sets were taken into consideration [2, 23]. Although the hull-based approaches have a better tolerance on intra-class variation, the global data structure is weakly characterized. In addition, it is computationally expensive, especially when there is a large number of images in each set. In summary, even though there is a plethora of algorithms developed to address the image set classification problem, most of them only focus on exploring more discriminative similarity measurements. Very few efforts were focused on reducing high time/storage cost and information redundancy introduced by large scale image sets.

To address this gap, we extend the method of Köstinger et al. [24] to set-to-set matching and propose the Set-based Prototype and Metric Learning framework (SPML). Groups of discriminative prototypes and a Mahalanobis metric are jointly learned for image set classification. The prototype learning seeks to represent the gallery image sets with fewer templates, while maintaining or improving the recognition performance. The metric learning seeks to tailor a more accurate set-to-set similarity measurement based on the learned prototypes. We formulate the learning problem in a single loss function, and optimize the prototypes and Mahalanobis metric simultaneously. After processing by our SPML, a probe image set lies closer to those gallery prototype sets from the same subject, and further from those gallery prototype sets from different subjects, as illustrated in Fig. 1.

Parts of this work have appeared in our conference version [25]. In this paper, we offer three major extensions: (i) we present a time complexity analysis on existing distance models to highlight our motivation; (ii) we provide more detailed discussions and comparisons with methods from different categories; (iii) we include additional sensitivity analyses carried out to explore different aspects of the proposed algorithm.

The rest of the paper is organized as follows: In Section 2, we discuss related works. In Section 3 we introduce the mathematical model of the proposed framework. In Section 4 we discuss the implementation of our framework and the testing time complexity. In Section 5 we present the experimental settings and results; in Section 6 we summarize the limitations of our proposed framework; Section 7 concludes the paper.

2. Related work

In this section, we offer a brief introduction on algorithms that are closely related to our work. In particular, our work is built on the regularized nearest points (RNP) method [21], set-to-set distance metric learning (SSDML) [22], and the prototype learning for large margin nearest neighbor classifiers [24]. For the convenience of discussion, an overview of the notations used in this paper is summarized in Table 1.

In RNP, Yang et al. [21] proposed to model an image set $\boldsymbol{X}_i$ as a regularized affine hull (RAH), spanned by all its samples:

$$\mathcal{H}(\boldsymbol{X}_i) = \left\{ \boldsymbol{X}_i\boldsymbol{\alpha}_i \middle| \sum_{m=1}^{N_i} \alpha_{i,m} = 1, \| \boldsymbol{\alpha}_i\|_{l_p} < \sigma \right\}, \tag{1}$$

with a regularization on the l_p norm of the the combination coefficient $\|\boldsymbol{\alpha}\|_{l_p} < \sigma$, where $\boldsymbol{\alpha}_i = [\alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,m}]^T$. The distance between two image sets $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ is then defined as the geodesic distance between $\mathcal{H}(\boldsymbol{X}_i)$ and $\mathcal{H}(\boldsymbol{X}_j)$,

$$\begin{aligned} &\mathcal{D}^2(\boldsymbol{X}_i, \boldsymbol{X}_j) = \min_{\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j} \left[(\boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{X}_j\boldsymbol{\alpha}_j)^T (\boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{X}_j\boldsymbol{\alpha}_j) \right] \\ &\text{s.t. } \| \boldsymbol{\alpha}_i\|_{l_p} < \sigma_1, \ \|\boldsymbol{\alpha}_j\|_{l_p} < \sigma_2, \sum_{m=1}^{N_i} \alpha_{i,m} = 1, \sum_{m=1}^{N_j} \alpha_{j,m} = 1. \end{aligned} \tag{2}$$

By relaxing $\sum_{m=1}^{N_i} \alpha_{i,m} = 1$ and $\sum_{m=1}^{N_j} \alpha_{j,m} = 1$ to $\sum_{m=1}^{N_i} \approx 1$ and $\sum_{m=1}^{N_j} \alpha_{j,m} \approx 1$ and using the Lagrangian formulation, Eq. (2) with $l_p = 2$ can be integrated as

$$\mathcal{D}^2(\boldsymbol{X}_i, \boldsymbol{X}_j) = \min_{\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j} \left(\|\boldsymbol{u} - \hat{\boldsymbol{X}}_i\boldsymbol{\alpha}_i - \hat{\boldsymbol{X}}_j\boldsymbol{\alpha}_j\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}_i\|_2^2 + \|\boldsymbol{\alpha}_j\|_2^2 \right), \tag{3}$$

where $\boldsymbol{u} = [\mathbf{0}; \mathbf{1}; \mathbf{1}]$, $\hat{\boldsymbol{X}}_i = [\boldsymbol{X}_i; \mathbf{1}^T; \mathbf{0}^T]$, $\hat{\boldsymbol{X}}_j = [-\boldsymbol{X}_j; \mathbf{0}^T; \mathbf{1}^T]$, and the column vectors $\mathbf{0}$ and $\mathbf{1}$ have the appropriate sizes associated with their corresponding context. Although the regularization can effectively restrict the expansion of the hull area, the natural geodesic distance might not reflect the dissimilarity for the task at hand properly. To tailor a more accurate set-to-set distance, Zhu et al. [22] extended the Mahalanobis distance metric learning [26] to the geodesic distance between hulls:

$$\begin{aligned} &\mathcal{D}^2_{\boldsymbol{M}}(\boldsymbol{X}_i, \boldsymbol{X}_j) = (\boldsymbol{X}_i\hat{\boldsymbol{\alpha}}_i - \boldsymbol{X}_j\hat{\boldsymbol{\alpha}}_j)^T \boldsymbol{M} (\boldsymbol{X}_i\hat{\boldsymbol{\alpha}}_i - \boldsymbol{X}_j\hat{\boldsymbol{\alpha}}_j) \\ &(\hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\alpha}}_j) = \arg\min_{\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j} \left[(\boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{X}_j\boldsymbol{\alpha}_j)^T \boldsymbol{M} (\boldsymbol{X}_i\boldsymbol{\alpha}_i - \boldsymbol{X}_j\boldsymbol{\alpha}_j) \right] \\ &\text{s.t. } \|\boldsymbol{\alpha}_i\|_{l_p} < \sigma_1, \ \|\boldsymbol{\alpha}_j\|_{l_p} < \sigma_2, \sum_{m=1}^{N_i} \alpha_{i,m} = 1, \sum_{m=1}^{N_j} \alpha_{j,m} = 1, \end{aligned} \tag{4}$$

where $\boldsymbol{M}$ is a positive semi-definite matrix to be learned. It can be learned using any distance metric learning model. In both RNP and SSDML, the restricted affine hull is spanned by all the samples in the image set. This is not only computationally expensive, but also sensitive to outliers. In single-shot image classification, Köstinger et al. [24] proposed to reduce and optimize the templates used for each subject, and a distance metric is learned jointly. In this paper, we extend this idea to set-to-set matching. In particular, we build a framework, which can jointly learn a prototype representation and a Mahalanobis distance metric for the geodesic distance between hulls.

Please cite this article as: M. Leng, et al., Joint prototype and metric learning for image set classification: Application to video face identification, Image and Vision Computing (2016), http://dx.doi.org/10.1016/j.imavis.2016.06.005