Neurocomputing 135 (2014) 306–312

Contents lists available at ScienceDirect

Neurocomputing

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

Joint sparse representation for video-based face recognition

Zhen Cui [a,b,c], Hong Chang [a,*], Shiguang Shan [a], Bingpeng Ma [c], Xilin Chen [a]

[a] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[b] School of Computer Science and Technology, Huaqiao University, Xiamen 361021, China
[c] University of China Academy Science, Beijing 100190, China

ARTICLE INFO

Article history:
Received 3 July 2013
Received in revised form
9 December 2013
Accepted 19 December 2013
Communicated by Qingshan Liu
Available online 8 January 2014

Keywords:
Face recognition
Sparse representation
Structured sparse representation
Accelerated proximal gradient

ABSTRACT

Video-based Face Recognition (VFR) can be converted into the problem of measuring the similarity of two image sets, where the examples from a video clip construct one image set. In this paper, we consider face images from each clip as an ensemble and formulate VFR into the Joint Sparse Representation (JSR) problem. In JSR, to adaptively learn the sparse representation of a probe clip, we simultaneously consider the class-level and atom-level sparsity, where the former structurizes the enrolled clips using the structured sparse regularizer (*i.e.*, $L_{2,1}$-norm) and the latter seeks for a few related examples using the sparse regularizer (*i.e.*, L_1−norm). Besides, we also consider to pre-train a compacted dictionary to accelerate the algorithm, and impose the non-negativity constraint on the recovered coefficients to encourage positive correlations of the representation. The classification is ruled in favor of the class that has the lowest accumulated reconstruction error. We conduct extensive experiments on three real-world databases: Honda, MoBo and YouTube Celebrities (YTC). The results demonstrate that our method is more competitive than those state-of-the-art VFR methods.

1. Introduction

In traditional face recognition task, face images are identified from only a few samples per subject under controlled environments. Many current algorithms have achieved pretty good performance. However, with the popularization of video cameras such as surveillance cameras and cell phone cameras, we can easily capture large-scale face video clips in the wild, in which face images usually accompany with dramatic appearance changes in lighting, pose, expression, blur, etc. Therefore, the efficient classification of face video clips remains challenging and meaningful in practical applications.

The popular methods are the explosive development of Image Set based Classification (ISC) techniques [1–10]. Generally speaking, these approaches consist of two key steps: representing image set and defining between-set similarity. As image set representation is concerned, popular methods include Gaussian models [1,2], subspaces [3,7,8], nonlinear manifolds [4,11,9], etc. Gaussian model based methods can reasonably extend to unseen data with well-estimated parameters. However, if the data distribution does not follow the Gaussian assumption, the estimated model will not properly fit with the real distribution of image set. Instead, the non-parametric methods revive in more recent years. They usually represent one image set as a linear subspace [3,7,8] or a nonlinear manifold [4,11,9]. Compared with Gaussian based methods, these non-parametric methods have demonstrated many favorable properties (*e.g.*, no assumption on data distribution) with more excellent performance in VFR.

The second concern is how to define between-set distance. Generally, different distance metrics are used for different set representation methods. For Gaussian model, Kullback–Leibler Divergence [2] may be used to define between-set similarity. For subspace model, principal angles between two subspaces are often used as the distance metric. The classic works include Mutual Subspace Method (MSM) [12], Orthogonal Subspace Method (OSM) and their variants [3,13]. To develop more robust distance of two subspaces, specifically, some recent studies attempt to constrain the space of synthetic face images. For example, Cevikalp et al. [7] constrained the subspace spanned from face images of a clip into a convex hull, and then calculate the nearest distance of two convex hulls as the between-set similarity. Hu et al. [8] further extended it and proposed Sparse Approximated Nearest Point (SANP) to make the nearest points between two convex hulls lie on some facets by using the sparse regularizer. In addition, by assuming that face images of a clip lie on a nonlinear manifold, Wang et al. [4] extended Subspace–Subspace Distance (SSD) to Manifold–Manifold Distance (MMD), where a nonlinear manifold is partitioned into several local linear subspaces and then MMD is defined as pair-wise SSDs. However, MMD implicitly suffers a computational bias due to the uncertainty of subspace partitions

* Corresponding author.
E-mail address: hong.chang@vipl.ict.ac.cn (H. Chang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.12.004

[9]. To this end, Cui et.al [9] attempted to align all image sets to a pre-specified reference set and then measured the corresponding subspaces, which inevitably leads to the dependence on the choice of the reference set for the classification accuracy.

More recently, the sparse representation based methods [14–16] are developed to address the task of face recognition. Especially, the Sparse Representation-based Classification (SRC) [14] method sparsely represents a probe face image with a dictionary constructed from all gallery examples and then classifies it into the subject with the smallest reconstruction error. If the examples from the same subject construct its own subspace, *i.e.*, the intersection of subspaces spanned from any two subjects is null, the non-zero coefficients of the reconstructed example can ideally focus on the gallery examples from the same subject. Following this assumption, SRC has shown favorable properties in face recognition, especially when face images are partially occluded.

However, SRC treats every examples in gallery set equally and does not consider the structure of gallery data especially the class label. Intuitively, all examples from the same subject should be treated as an ensemble instead of multiple isolated images, which implies that the dictionary (*i.e.*, gallery set in SRC) may be characterized with the structure of group. For this, Elastic Net [17] and Group Lasso [18–20] are proposed to improve SRC. Specifically, Elhamifar et al. [20] casted the classification task as a structured sparse recovery problem, where the images from the same subject in gallery set form a group, and the sparsity is imposed on these groups, *i.e.*, the class-level or group-level sparsity. However, these methods only address the representation of a singe probe example and do not consider within-class appearance variations of an image set.

In addition, SRC is only designed to encode a single probe image. In a video clip, however, there are multiple frame images of the same subject, *i.e.*, multiple views of a subject. Note that here each clip only contains images of the same subject. Obviously, in a clip there exist strong correlations across different frames because a clip may be regarded as an approximately continuous stream. Therefore, when representing a clip, instead of frame-wise regression, the joint representation of all frames should be more meaningful for resisting the noises and increasing the representation stability. Generally, this problem of jointly estimating models from multiple related images is referred to "multi-view learning" [21,22] in the machine learning literatures. Yuan et al. [21] proposed Multi-Task Joint Sparse Representation (MTJSR), which aims to recover a test sample with multiple features from as few training subjects as possible and simultaneously enforces sparse coefficients on common atoms. However, MTJSR assumes that each sample has the same type of features, which naturally leads to the counterparts between multiple features. In the task of VFR, given any two clips, it is very intractable to obtain the counterparts between two clips (*i.e.*, images with the same poses, expressions, *etc.*). Thus it is impossible to encode one image with the same type of examples (or atoms) across different clips.

In this paper, inspired by recent progresses on sparse learning [14,18,21,23], we formulate VFR into a Joint Sparse Representation (JSR) problem (as shown in Fig. 1). In JSR, two sparse constraints are considered. The first sparse constraint is put on class-level by using $L_{2,1}$ mixed-norm, which assumes that a probe image set (or a clip) can be represented by a few gallery image sets (or gallery clips). The second sparse constraint enforces the sparsity on within-class images by using L_1-norm, with the aim to choose a few related views. Intuitively, different subjects lead to class-level sparsity, while appearance variations cause atom-level sparsity among images of all persons. In addition, in order to make the model more robust, two improvements are further provided: one is to learn a compact dictionary to reduce time cost, and the other is to impose nonnegative constraints on the representation. To solve this model of JSR, the Accelerated Proximal Gradient (APG) [24] optimization strategy is employed with fast convergence rate guaranteed. We conduct extensive experiments on three video databases: Honda [25], MoBo [26] and YouTube Celebrities(YTC) [27]. The results demonstrate that the proposed method is more competitive than the state-of-the-art methods for video-based face recognition.

The remainder of this paper is organized as follows. In Section 2.2, we present the proposed joint sparse representation model. The optimization details along with the final classification rule are stated in Sections 2.3 and 2.4. The applications of our method to face recognition are reported in Section 3. Finally, we reach a conclusion in Section 4.

2. Joint sparse representation

In this section, we first introduce the basic idea of joint sparse representation, then give the mathematical formulation in detail, and finally provide the optimization and the classification rule.

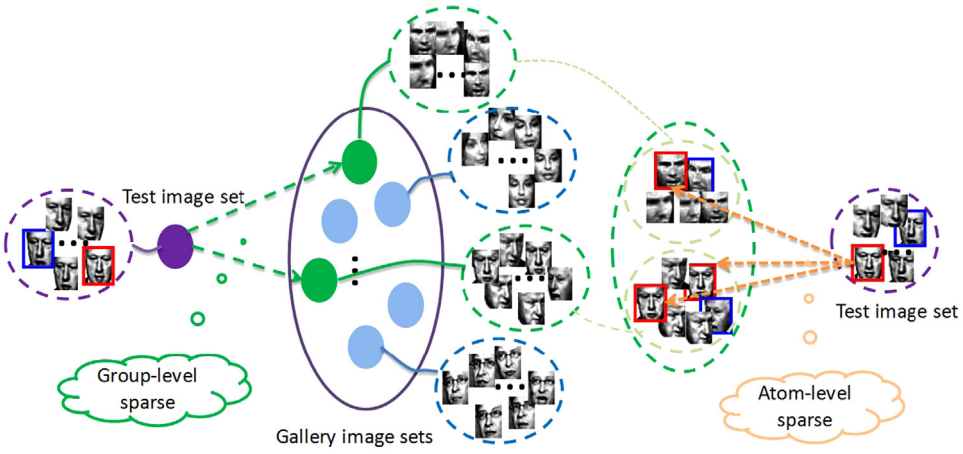

Fig. 1. Illustration of our idea. Given a test image set (or a video clip), the group-level (or class-level) sparse recovery is used to search the most relevant subjects from gallery image sets (or gallery clips), while the atom-level sparse regression is imposed on each image of the test set to find the similar appearance images. Further, such two sparse constrains are jointly imposed on an image set (or a clip) rather than an isolated image, which suppresses noises and leads to more robust representations. Behind that an intuitive explanation is that different subjects lead to the class-level sparsity, while appearance variations cause the atom-level sparsity among images of each person. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)