Neurocomputing 174 (2016) 866–874

Contents lists available at ScienceDirect

Neurocomputing

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

Iterative deep learning for image set based face and object recognition

Syed Afaq Ali Shah [a,*], Mohammed Bennamoun [a], Farid Boussaid [b]

[a] School of Computer Science and Software Engineering, The University of Western Australia, Perth 6009, Australia
[b] School of Electrical, Electronics and Computer Engineering, The University of Western Australia, Perth 6009, Australia

ARTICLE INFO

Article history:
Received 16 March 2015
Received in revised form
4 August 2015
Accepted 1 October 2015
Communicated by Jiayu Zhou
Available online 22 October 2015

Keywords:
Face/object recognition
Image set classification

ABSTRACT

We present a novel technique for image set based face/object recognition, where each gallery and query example contains a face/object image set captured from different viewpoints, background, facial expressions, resolution and illumination levels. While several image set classification approaches have been proposed in recent years, most of them represent each image set as a single linear subspace, mixture of linear subspaces or Lie group of Riemannian manifold. These techniques make prior assumptions in regards to the specific category of the geometric surface on which images of the set are believed to lie. This could result in a loss of discriminative information for classification. This paper alleviates these limitations by proposing an Iterative Deep Learning Model (IDLM) that automatically and hierarchically learns discriminative representations from *raw* face and object images. In the proposed approach, low level translationally invariant features are learnt by the Pooled Convolutional Layer (PCL). The latter is followed by Artificial Neural Networks (ANNs) applied iteratively in a hierarchical fashion to learn a discriminative non-linear feature representation of the input image sets. The proposed technique was extensively evaluated for the task of image set based face and object recognition on YouTube Celebrities, Honda/UCSD, CMU Mobo and ETH-80 (object) dataset, respectively. Experimental results and comparisons with state-of-the-art methods show that our technique achieves the best performance on all these datasets.

1. Introduction

Face and object recognition have various real life applications, such as human-computer interaction, surveillance, pervasive computing and access control (non-intrusive security), to name a few [1–9]. Variations in illumination and viewing direction make face/object recognition a very challenging task [10–13]. Recently, face/object recognition using image sets has attracted significant attention from the research community [14–16]. This is mainly because of the following two reasons. Firstly, compared with single mug-shot, image sets offer significantly more information about the variations in appearance of the target face or object [17]. These variations exist in the context of face/object recognition, where multiple shots of the face/object under varying facial expressions, changing illumination conditions, background, occlusions and viewpoint variations captured over long periods of time are available [18,19]. Secondly, the availability of modern low cost portable sensors make image sets a more natural choice of input for face and object recognition tasks, e.g. surveillance scenario and/or indoor robotic navigation and localization.

Image set based methods are expected to give better performance than the ones based on individual images [17]. This is because image set allows the decision process to be based on comparisons of the most similar pairs of the query and gallery images. Moreover, they incorporate information about the variability of the face/object appearance. In many applications (e.g. surveillance systems involving face/object tracking), image sets are the most natural form of input to the system [17].

While image set based face/object recognition provides a better classification performance, compared to a single image, it also poses many challenges [20]. The main challenge lies in effectively and automatically modeling the image set into a meaningful representation without losing discriminative information. Existing methods commonly model appearance variability information within images of a set on a non-linear manifold geometry such as the Grassmannian [14], Lie Group [21] of Riemannian manifold, an image set modeled by a subspace [22,23] or a combination of subspaces [24,25]. These approaches make prior assumptions in regards to the specific category of the geometric surface on which image sets are believed to lie [26], resulting in a loss of discriminative information. This paper addresses this problem by presenting a novel deep learning framework for image set based face/object recognition. The paper makes the following three main contributions:

* Corresponding author.
E-mail address: syed.shah@research.uwa.edu.au (S. Afaq Ali Shah).

http://dx.doi.org/10.1016/j.neucom.2015.10.004
0925-2312/

1. A novel Iterative Deep Learning Model (IDLM) is proposed to hierarchically learn class-specific image set representations. IDLM achieves translational invariance at lower hierarchy levels and learns discriminative compositional features at higher hierarchy levels.
2. The proposed deep network automatically learns discriminative features from sets of *raw* images. It does not rely on any hand-crafted image features to achieve high performance.
3. The proposed IDLM does not make any prior assumptions about the underlying geometry of faces and objects. Instead, the proposed framework automatically learns the structure of the complex non-linear surface on which sets of images reside.

The rest of this paper is organized as follows. Related work is presented in the next section. The proposed technique is described in Section 3. Section 4 describes the image set based classification algorithm. Experimental results and analysis are presented in Section 5. Finally, the paper is concluded in Section 7.

2. Related work

In this section, a brief overview of existing image set based face recognition methods is provided. The key issues in image set based classification include how to model a set and consequently how to compute the distance/similarity between query and gallery sets. In the literature, two approaches have been proposed for image set modeling: (a) parametric-model and (b) non-parametric-model methods. The details of these methods are as follows.

2.1. Parametric model methods

The parametric model methods [27] represent an image set in terms of parametric distribution [26]. These techniques use Kullback–Leibler (KL) divergence to measure the similarity between the distributions. The limitations of these techniques lie in that they require a strong statistical relationship between the gallery and query image sets to achieve good performance [21]. To address the shortcomings of these methods, non-parametric model methods (described below) for image set classification have been proposed [22,24,28,17].

2.2. Non-parametric model methods

Non-parametric model based methods approximate an image set either on a geometric surface or by its representative exemplars. These methods have shown promising results and are being actively developed [14,15,21,29,16,30,26].

2.2.1. Geometric surface based methods

This category of non-parametric methods represent a complete image set by a point on a geometric surface [24,28,14,21]. The image set can be modeled either on a complex non-linear manifold or by a subspace and/or mixture of subspaces. For image set representations on manifolds, appropriate distance metrics have been adopted such as the geodesic distance [31], the projection kernel metric [32] on the Grassmann manifold, and the log-map distance metric [33] on the Lie group of Riemannian manifold. In order to discriminate image sets on the manifold surface, different learning strategies have been developed. Mostly, a discriminant analysis method is formulated for different set representations. Examples include Covariance Discriminative Learning (CDL) [21], Manifold Discriminant Analysis (MDA) [28], Discriminative Canonical Correlations (DCC) [22] and Graph Embedding Discriminant Analysis (GEDA) [14].

To determine the distance between image sets represented by a linear subspace, principal angles are commonly used [21]. The p principal angles $0 \le \theta_1 \le \cdots \le \theta_p \le \pi/2$ between two subspaces are defined as the smallest angles between any vector in one subspace and any other vector in the second subspace. The sum of the cosines of the principal angles is then used to compute the similarity between subspaces. The performance of geometric surface based methods significantly degrades when the image set has a small sample size but big variations [26].

2.2.2. Representative exemplars

For image sets represented in terms of representative exemplars, the set–set distance can be defined as the Euclidean distance between the set representatives [21]. These can simply be adaptively learnt set samples [17,15] or the set mean [24]. Cevikalp and Triggs [17] used the affine hull or convex hull models of the set images to learn the set samples. In their technique, the set–set distance is termed as Affine Hull Image Set Distance (AHISD) or Convex Hull Image Set Distance (CHISD). Hu et al. [15] proposed Sparse Approximated Nearest Points (SANPs). The SANPs of two sets are first computed from the mean image and the affine hull model of the corresponding sets. The SANPs are then sparsely approximated from the set's sample images while simultaneously searching for the closest points between sets. The similarity between image sets is then defined on the basis of distance between their SANPs. The set representative based methods require the computation of a one-to-one set distance. These methods are therefore capable of handling intra set variations very effectively. However, their performance is highly prone to outliers [26,21]. They are also computationally very expensive as a one-to-one match of the query set with all sets in the galley is required.

Recently, Adaptive Deep Network (ADNT) [26] has been proposed for image set classification. This deep learning framework consists of encoder and decoder layers, which are used for reconstruction of input images. ADNT has shown to achieve a superior performance compared to existing techniques. However, unlike existing deep networks which use raw data for feature learning, ADNT uses hand-crafted LBP features for training and testing. The use of LBP features helps to achieve a good performance, but it adds an additional computational complexity and limits the main purpose of deep network i.e. to automatically learn features from raw data.

From the above review, it can be concluded that most existing methods represent image sets either on a geometric surface [17,15,29], as a point on Grassmanian manifold [24], a Lie group of Riemannian manifold [21] or model the image set by a subspace [22,23]. These methods therefore make prior assumptions about the underlying geometry on which images of a set are believed to lie [26]. Moreover, the recently proposed deep learning based technique in [26] uses hand-crafted LBP features for high performance. In contrast, our proposed method defines an Iterative Deep Learning Model (IDLM) that can automatically learn the underlying geometric structure from the raw images in the set. Our IDLM consists of a single layer CNN [34], called Pooled Convolutional Layer (PCL) and multiple Artificial Neural Network (ANN) layers stacked together through non-linear activation functions. PCL was adopted due to its translational invariance properties. The IDLM is therefore capable of discovering the complex geometric surface on which images of a set are present.

Note that there are significant differences between the proposed IDLM and ADNT [26], which is based on an autoencoder and consists of encoder, decoder and hidden layers. *First*, the number and size of the layers is randomly selected in ADNT. *Second*, the performance of ADNT is highly dependent on hand-crafted LBP features. *Third*, the autoencoder of ADNT is based on the concept of sparse coding [35]. The shortcoming of sparse coding is that for