Neurocomputing xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

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

Image set classification based on synthetic examples and reverse training

Lin Zhang[a,b,*], Qingjun Liang[a], Ying Shen[a], Meng Yang[c], Feng Liu[c]

[a] School of Software Engineering, Tongji University, Shanghai, China
[b] Shenzhen Institute of Future Media Technology, Shenzhen, China
[c] School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

ARTICLE INFO

Keywords:
Face recognition
Image set classification
Reverse training

ABSTRACT

Classification based on image sets has recently attracted increasing interests in computer vision and pattern recognition community. It finds numerous applications in real-life scenarios, such as classification from surveillance videos, multi-view camera networks, and personal albums. Image set based face classification highly depends on the consistency and coverage of the poses and view point variations of a subject in gallery and probe sets. This paper explores a synthetic method to create the unseen face features in the database, thus achieving better performance of image set based face recognition. By considering the high symmetry of human faces, multiple synthetic instances are virtually generated to make up the missing parts, so as to enrich the variety of the database. With respect to the classification framework, we resort to reverse training due to its high efficiency and accuracy. The performance of the proposed approach, Synthetic Examples based Reverse Training (SERT), has been fully evaluated on Honda/UCSD, CMU Mobo and YouTube Celebrities, three benchmark datasets comprising facial image sequences. Extensive comparisons with the other state-of-the-art methods have corroborated the superiority of our approach.

1. Introduction

Image classification has attracted much attention from researchers recently since it has many significant potential applications [1–5]. As a special kind of image classification problems, face recognition has been studied for decades [6–9]. Traditional face recognition can be regarded as a single image classification problem. With the significant development in imaging technology, multiple images of a person are becoming readily available in a number of real-world scenarios, such as video surveillance, multi-view camera networks, and personal albums collected during a period of time. Face recognition based on multiple images can be formulated as an image set classification problem, where each set contains images belonging to the same person but covering a wide range of variations. These variations could be caused by illumination variations, viewpoint variations, different backgrounds, expressions, occlusions, disguise, etc. More robust and promising face recognition can be expected by using image sets since they contribute more information than one single image.

In the past decade, the image set based recognition has gained significant attention from the research community. Generally speaking, there are two major steps involved in image set classification, to find a suitable representation of the images in the set and to define an appropriate distance metric for computing the similarity between these representations. According to the types of representations, existing image set classification methods can be classified into two categories, parametric model based methods and non-parametric model based methods [10,11].

Parametric-model based approaches tend to utilize a specific statistical distribution model to represent an image set and measure the similarity between two distribution models using KL-divergence [12,13]. The main drawback of such methods is that they may fail to produce a desirable performance if there is no strong statistical relationship between the training and the test image sets.

Unlike parametric-model based methods seeking for global characteristics of the sets, non-parametric model based ones put more emphasis on matching local samples. They do not model image sets as statistical distributions. Instead, they attempt to find the overlapping views between two sets and measure the similarity upon those parts of data. Non-parametric model based approaches have shown promising results and have received much attention recently. Several representative ones belonging to this category will be briefly reviewed here.

For non-parametric model based methods, there are usually two ways to represent an image set, either by its representative exemplars or by a point on a geometric surface. Then, different distance metrics to determine the between-set distance will be defined with respect to different types of representations. For image sets represented by

* Corresponding author at: School of Software Engineering, Tongji University, Shanghai, China.
E-mail address: cslinzhang@tongji.edu.cn (L. Zhang).

http://dx.doi.org/10.1016/j.neucom.2016.04.067
Received 20 December 2015; Received in revised form 16 March 2016; Accepted 10 April 2016
Available online xxxx
0925-2312/

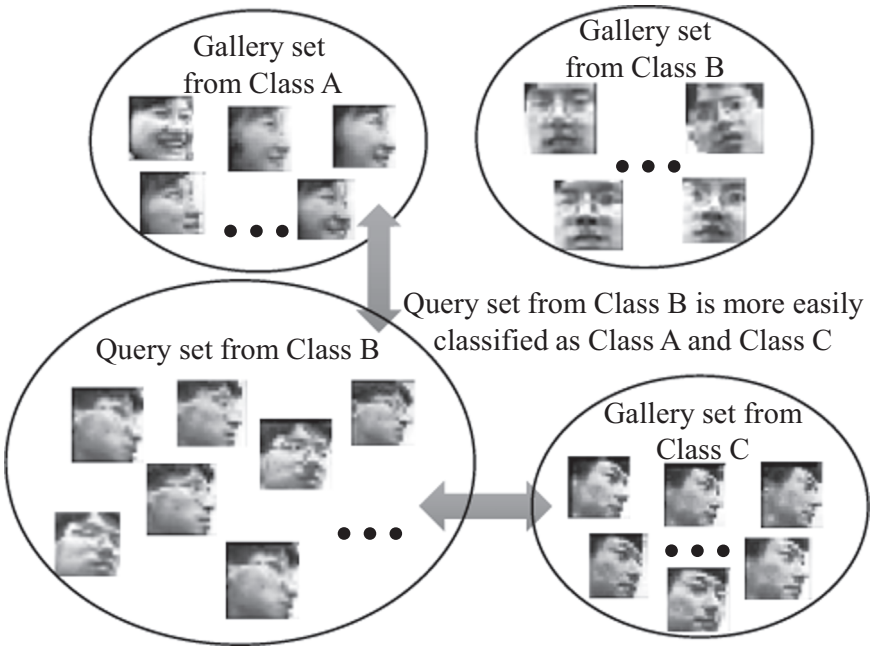

Fig. 1. Images in the query set from class B have different poses from the images in the gallery set from class B. However, their poses are quite similar to the images in the gallery set from classes A and C. The query set from class B is more likely to be misclassified as A or C.

representative exemplars, usually the Euclidean distance between the set representatives is regarded as the set-set distance. The set representatives can simply be the set mean or adaptively learned set samples [14–17]. In [14], Cevikalp and Triggs learned the representative set samples from the affine hull or convex hull models of the set images and accordingly the set-set distance is termed as Affine Hull Image Set Distance (AHISD) or Convex Hull Image Set Distance (CHISD). In Hu et al.'s approach [15], the SANPs (Sparse Approximated Nearest Points) of two sets are first determined from the mean image and the affine hull model of the two corresponding sets. After that, SANPs are sparsely approximated by the set's sample images and then the closest points between sets can be obtained. The set-set distance is computed as the Euclidean distance between two closest SANPs of the two sets. In [16], by representing the image set as a nonlinear manifold, Hadid et al. extracted exemplars from the manifold using Locally Linear Embedding (LLE) and *k*-means based clustering. In [17], Yang et al. modeled an image set as a regularized affine hull (RAH) and then two regularized nearest points (RNP), one for each RAH, are automatically computed. Then, the between-set distance was computed as the modulated distance between RNPs by the structure of image sets. One potential drawback of the set representative based methods is that their performance is highly sensitive to outliers. In addition, they are also computationally very expensive since a one-to-one match of the query set with all the gallery sets is required. Hence, these methods run very slowly when the size of the gallery set is quite large.

Different from exemplar-based methods, some other non-parametric model based approaches attempt to represent an image set by a point on a geometric surface. Using these methods, an image set can be represented by a subspace [18–22], a combination of subspaces [23–26], or a point on a complex nonlinear manifold [27–31]. For methods using a linear subspace to represent an image set, the angles between two subspaces, which mainly characterize the common modes between variations of the two subspaces, are commonly used as a similarity measure. For manifold-based image set representations, appropriate distance metrics have been developed, such as the geodesic distance [32], the projection kernel metric [33] on the Grassmann manifold, the log-map distance metric [34] on the Lie group of Riemannian manifold, or even learned by some distance metric learning techniques [35]. In order to discriminate image sets on the manifold surface, different learning strategies have been proposed, including Discriminative Canonical Correlations (DCC) [18], Manifold Discriminant Analysis (MDA) [28], Graph Embedding Discriminant Analysis (GEDA) [27], Covariance Discriminative Learning (CDL) [31]. In [36], Hayat et al. tried to keep every example independent and to remain the image set in its original form rather than seeking a global representation. They argued that whatever form you use, once you model a set as a single entity, there must be loss of information. For image set classification, they proposed a reverse training scheme. With the reverse training scheme, the classifier is trained with the images of the query set (labeled as positive) and a randomly sampled subset of the training data (labeled as negative). The trained classifier is then evaluated on rest of the training images. The class of the images with their largest percentage classified as positive is predicted as the class of the query image set. Quite recently, Hayat et al. introduced a deep learning based framework to deal with the image set classification problem [10,11]. Specifically, a Template Deep Reconstruction Model (TDRM) is defined and initialized by performing an unsupervised pre-training in a layer-wise fashion. The initialized TDRM is then separately trained for images of each class and class-specific DRMs are learned. At the testing stage, the classification is performed based on the minimum reconstruction errors from the learned class-specific models. Also based on deep learning, Shah *et al.* proposed an Iterative Deep Learning Model (IDLM) that could automatically and hierarchically learn discriminative representations from raw face and object images [37].

Based on the literature review, we found that all the aforementioned methods mainly focus on devising effective classifiers for image sets. They implicitly make an assumption that the distribution of a person's poses and view points in a probe image set are similar to those in the gallery image set. However, it is sometimes the case that there is pose or view point mismatch between the gallery and probe image sets of the same subject. In such case, the probe image set is more likely to be misclassified as the class containing images with the same head pose as the probe set but actually from a different subject. In Fig. 1, we use a vivid example to illustrate this phenomenon. We suppose that there are three classes A, B, and C in the gallery set and they are denoted by GA, GB, and GC, respectively. Suppose that images from GA and GC have