Neurocomputing 228 (2017) 3–10

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 31 October 2016
0925-2312/

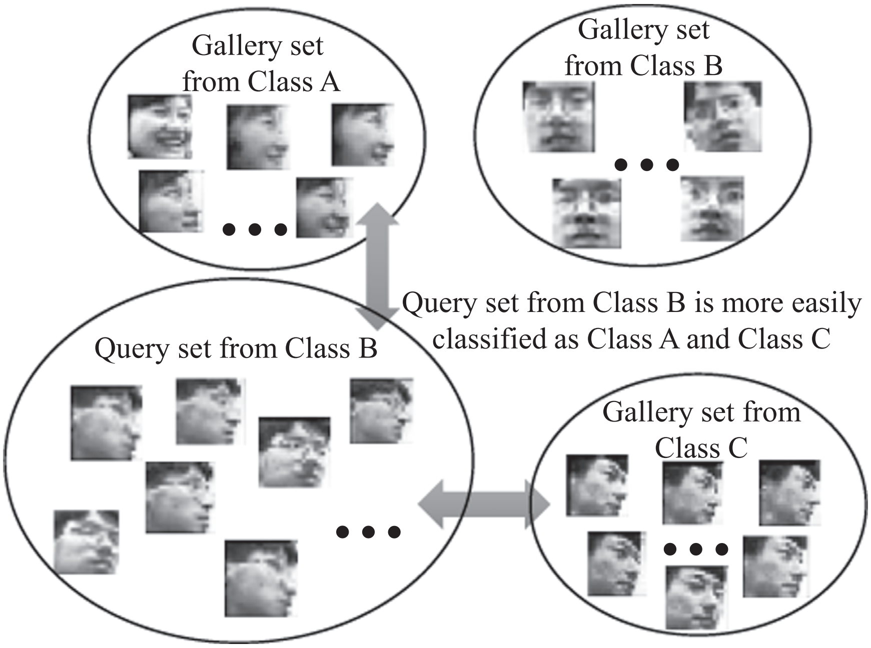


Fig. 1. Images in the query set from class B have different poses from the images in the gallery set from class B. However, their poses are quite similar to the images in the gallery set from classes A and C. The query set from class B is more likely to be misclassified as A or C.

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 similar poses while their poses are quite different from images in GB. At the test stage, a query set, comprising images actually belonging to class B, comes and is denoted by QB. In this example, images in QB have quite different poses from images in GB, but similar poses with images in GA and GC. Hence, it is highly possible that the image set QB will be misclassified as class A or class C.

In this paper, to solve such a problem, we propose a simple yet effective approach by synthesizing more samples for each image set. In this way, the variety of poses and viewpoints within an image set can be explicitly enriched. With respect to the classification framework, we resort to Reverse Training [36] since it is a concept-simple and effective approach to deal with the image set classification problem. The proposed method is named as *Synthetic Examples based Reverse Training*, SERT for short. Our method can deal with the pose and the view point mismatch in video based face recognition quite well. SERT is evaluated against the state-of-the-art image set classification methods and found to have superiority over them.

The rest of this paper is organized as follows. Section 2 discusses how to generate synthetic examples. Section 3 describes the reverse training method for image set classification and an overview of SERT. Section 4 presents the experimental results. Finally, Section 5 concludes the paper.

2. Image set feature extraction

We propose a face sample synthesizing method in which the symmetry property of the human face is fully exploited. The proposed approach is inspired by [36]. In [36], Hayat et al. pointed out that based on the manual inspection of the most challenging YouTube Celebrities dataset [38], a great amount of misclassified query image sets have a common characteristic that their head poses are not covered in the corresponding training sets. Fig. 2 shows a challenging example in Youtube Celebrities dataset. Fig. 2(a) shows images in one training image set and Fig. 2(b) shows images in the corresponding test image set. In this example, images in the test set and training set have quite different facial poses, which raises a great challenge to the set classification algorithms. To address this issue, here we present our solution.

2.1. Synthesizing examples and extracting block-wise LBP based features

We create synthetic examples to enrich the set variations, by operating directly in "data space". For each sample in an image set, we first flip the image horizontally and get another symmetric version of the original face. To determine the necessity of this flipping step, we use the Euclidean distance metric to measure the similarity between the original face and the flipped one. A threshold is empirically set. If the distance is less than the threshold, we neglect the flipped face since the original face itself has a good symmetry. Otherwise, we add the new flipped face to the image set and therefore augment the number of instances in all sets.

With respect to feature extraction, we propose to use a block-wise LBP (Local Binary Patterns [39]) based scheme. Specifically, each face image is at first uniformly partitioned into $k \times k$ blocks and then an LBP histogram is extracted from each block. After normalizing each histogram, all the normalized histograms are concatenated together as the final feature vector. Such a feature extraction procedure is quite robust to image noise since LBP is actually an ordinal feature, simply depending on the signs of pixel differences. Actually, LBP has three classical mapping table: (1) uniform LBP ('u2'), (2) rotation-invariant LBP ('ri'), and (3) uniform rotation-invariant LBP ('riu2'). Here we adopt the uniform LBP ('u2'), whose binary pattern contains at most two bitwise transitions from 0 to 1 (or 1 to 0). There are totally 2 cases for zero transition and 56 cases for 2 transitions (1 transition is impossible) when the sampling density is 8. All the non-uniform LBPs that contain more than two transitions are labeled as the 59th bin. Details of our feature extraction scheme are summarized in Table 1.

As for rotation-invariant LBP ('ri'), the number of patterns is largely filtered out and only 36 conditions/bins are remained, which is to some degree not sufficient to express the texture details of a face. For uniform rotation-invariant LBP ('riu2'), the patterns are reduced even more with only 10 conditions/bins reserved, having the same defects as 'ri'. That explains why we don't use those two mapping tables.

Since we use block-wise *LBPu2* 8,1 based feature extraction scheme which is not rotation invariant, the flipped image must have a different LBP value from its original one. An intuitive illustration can be seen in Fig. 3. Imagine the case that a training set only comprises left profile faces, while its corresponding upcoming test set only consists of right profile faces. It is obviously that the original gallery and probe set are

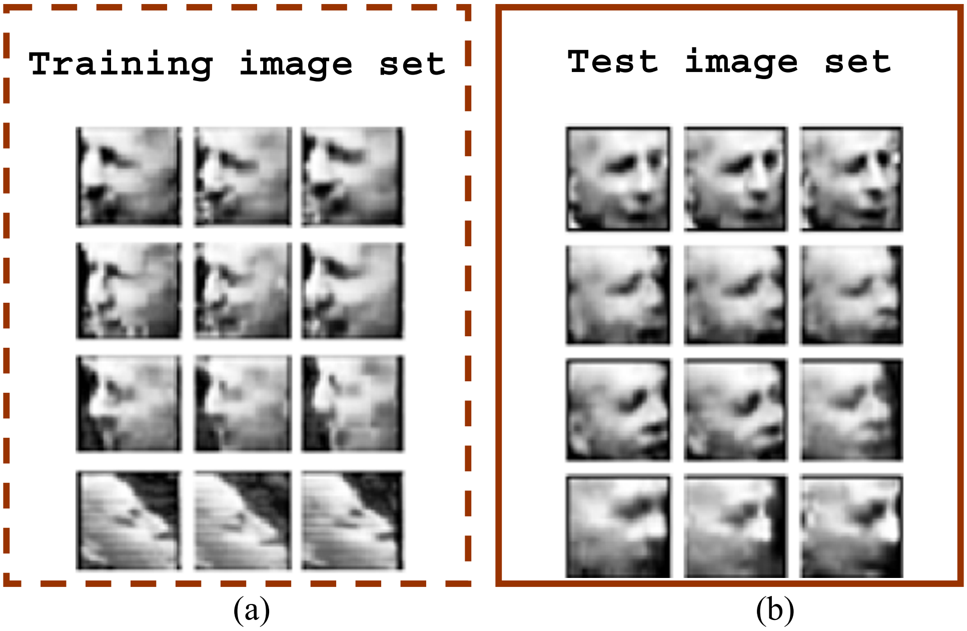


Fig. 2. A challenging example in YouTube Celebrities dataset. (a) shows images in one training set while (b) shows images in the corresponding test set.

Table 1
Feature extraction based on block-wise $LBP_{8,1}^{u2}$.

Input: A face image
1. Divide the face image into k×k non-overlapping uniformly spaced grid cells;
2. For each pixel in one cell, sample its 8 neighbors with radius 1 and map its pattern into one of the 59 cases;
3. Build the histogram over each cell, which counts the frequency of each number (1–59);
4. Normalize the histograms and concatenate them one after another (either column-wise or row-wise);

Output: A feature vector whose dimension is $59k^2$

hard to match to each other. However, after we create synthetic examples, both gallery and probe set contains left and right profile features, making later classification much easier.

2.2. SMOTE and PCA whitening

The number of instances varies a lot from set to set. Such an uneven distribution will lead to bias in the classification stage especially for those methods who do not represent the image set as a whole entity. To solve this problem, here we use the Synthetic Minority Oversampling TEchnique (SMOTE) proposed by Chawla et al. [40] to oversample the minority class and create synthetic training examples. For each training sample of the minority class, we take the difference between the sample and its nearest neighbor. Then the difference is multiplied with a random number between 0–1 and added to the original example to get the synthesized sample. In this way, we can get a synthetic sample lying on the line connecting the original sample and its nearest neighbor. The total number of required synthetic samples can be controlled by the number of nearest neighbors considered for each sample and the number of points generated on the line connecting the original sample and its nearest neighbor.

With our feature extraction scheme, histograms from all the blocks are concatenated together as the final feature vector. Considering that there exists a strong correlation between adjacent patches of an image and consequently the LBP features are redundant, we use PCA whitening to make our input features uncorrelated with each other and have unit variance along each dimension.

3. SERT: synthetic examples based reverse training

3.1. Problem formulation

Denote $\boldsymbol{X}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ as an image set containing n face examples from a person, where $\boldsymbol{x}_i$ is the feature vector of the i^{th} single image, and is in the form of LBP. A subject can have multiple image sets. Given k training image sets $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_k$ that belong to c classes (k > =c) and their corresponding labels y={1, 2, ..., c}, when a query image set $\boldsymbol{X}_q$ comes, our task is to find out which class it belongs to.

3.2. Reverse training and the proposed framework

After the preparation for features, we use the Reverse Training algorithm proposed in [36] to do the classification work.

For better illustration, here we use a toy example. Suppose a coming query set $\boldsymbol{X}_q$ has 200 images. The 20 training sets that belong to 20 classes (multiple sets per subject are combined as a whole) are denoted by $\boldsymbol{D}=\{\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_{20}\}$. 10 images per set in $\boldsymbol{D}$ are randomly selected to form a set $\boldsymbol{D}_1$ containing 200 images and the rest of images in $\boldsymbol{D}$ form the set $\boldsymbol{D}_2$. As the name "reverse training" suggests, we treat the 200 images in $\boldsymbol{X}_q$ and images in $\boldsymbol{D}_1$ as training data and the images in $\boldsymbol{D}_2$ as test data. We train a binary classifier. Specifically, all 200 images in $\boldsymbol{X}_q$ are labeled as +1 while 200 images in $\boldsymbol{D}_1$ are labeled as −1. A binary classifier *Liblinear* [41] is trained on these 400 instances. Since images from all classes are present in $\boldsymbol{D}_1$, the classifier learns to separate images in $\boldsymbol{X}_q$ from images of other classes. Actually, $\boldsymbol{D}_1$ does have a small number of images belonging to the same class as $\boldsymbol{X}_q$. However, since the number of these images is quite small, the learned binary classifier treats them as outliers and learns to discriminate the class of the query image set from all other classes.

Then, images in $\boldsymbol{D}_2$ are tested on the learned binary classifier. Those images who are classified as +1 (same side as $\boldsymbol{X}_q$) are denoted as $\boldsymbol{D}^+{}_2$. Let $y_{D_2^+}$ denote the class labels of images in $\boldsymbol{D}^+{}_2$. A normalized frequency histogram h of class labels in$y_{D_2^+}$ is computed, which has 20 bins in our example. Intuitively, the ith bin of the histogram, h_i (i=1, 2, ..., 20), represents the percentage of images of class i in $\boldsymbol{D}_2$ which are classified as +1. Or in other words, h_i is given by the ratio of the number of images of $\boldsymbol{D}_2$ belonging to class i and classified as +1 to the total number of images of $\boldsymbol{D}_2$ belonging to class i. h_i is computed as,

$$h_i = \sum\nolimits_{y \in y_{D_2^+}} f_i(y) / \sum\nolimits_{y \in y_{D_2}} f_i(y), \quad \text{where } f_i(y) = \begin{cases} 1, & y = i \\ 0, & y \neq i \end{cases} \tag{1}$$

Finally, the label of the query set $\boldsymbol{X}_q$ can be predicted as the class in $\boldsymbol{D}_2$ with most of its images classified as +1. Therefore, the class label of $\boldsymbol{X}_q$ is assigned to,

$$y_q = \arg\max_i h_i \tag{2}$$

The reason we choose reverse training as our classification method is straightforward. Reverse Training (RT) has two main advantages: (1) it does not need offline training and can adapt to newly added training

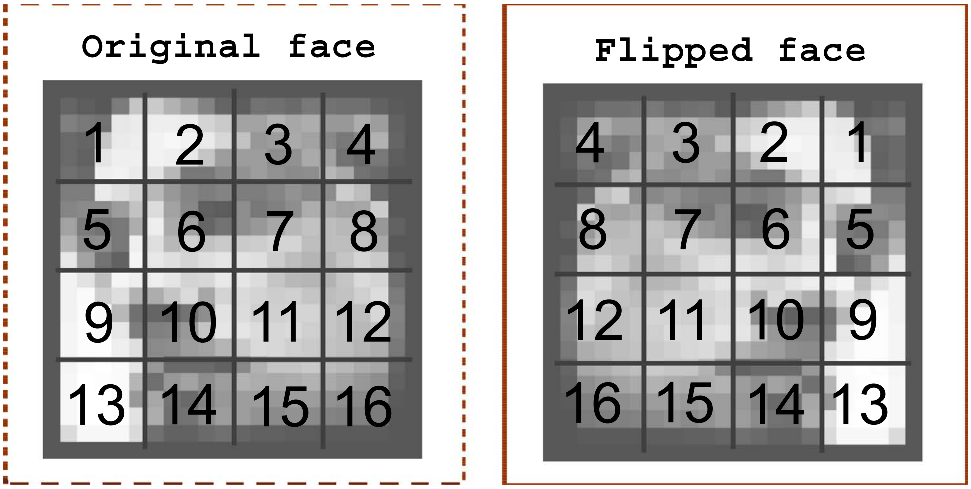


Fig. 3. A synthetic feature and its original feature.

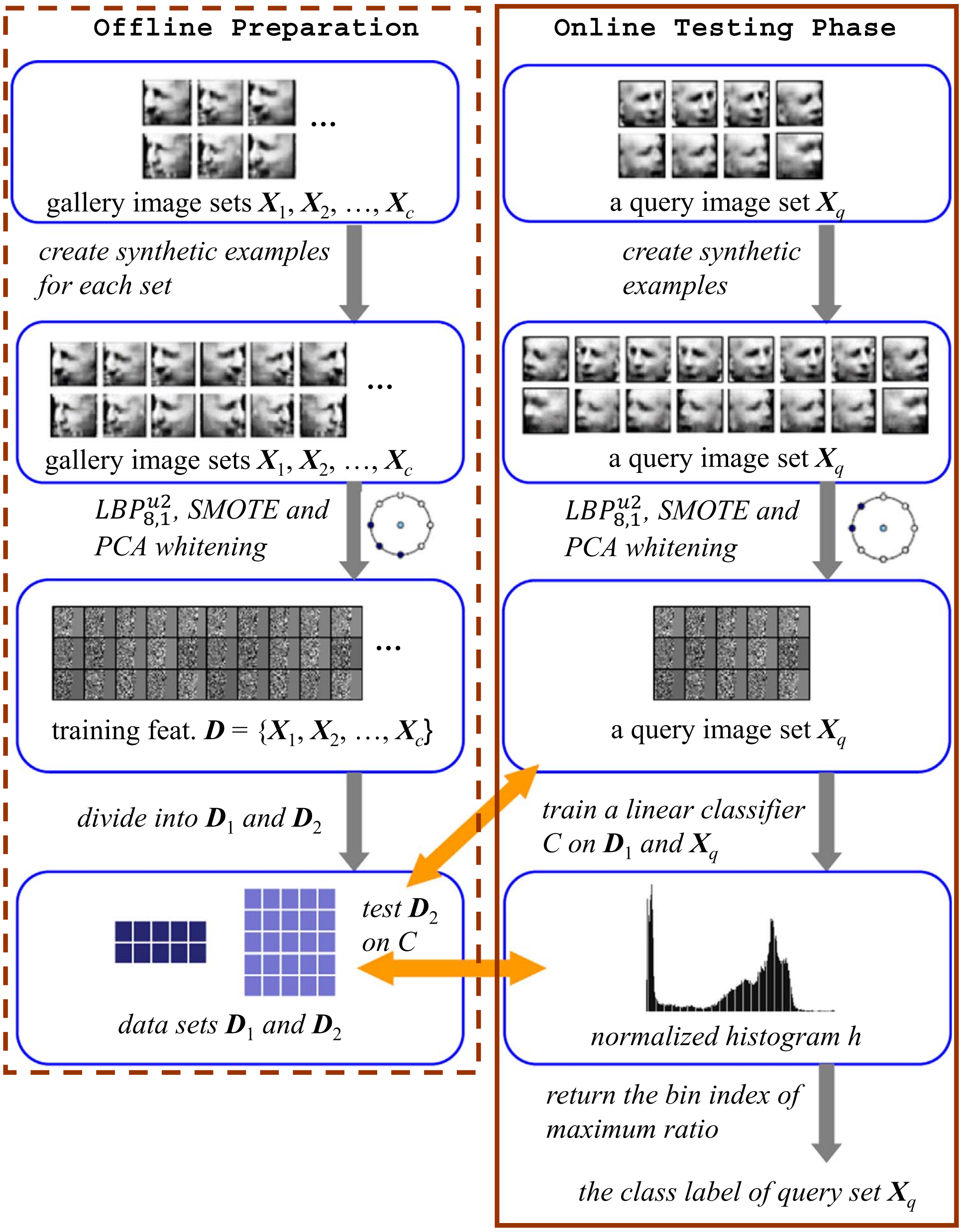

Offline Preparation
Online Testing Phase
gallery image sets $X_1, X_2, \ldots, X_c$
a query image set X_q
create synthetic examples for each set
create synthetic examples
gallery image sets $X_1, X_2, \ldots, X_c$
a query image set X_q
$LBP_{8,1}^{u2}$, SMOTE and PCA whitening
$LBP_{8,1}^{u2}$, SMOTE and PCA whitening
training feat. $D = \{X_1, X_2, \ldots, X_c\}$
a query image set X_q
divide into D_1 and D_2
train a linear classifier C on D_1 and X_q
test D_2 on C
data sets D_1 and D_2
normalized histogram h
return the bin index of maximum ratio
the class label of query set X_q

Fig. 4. Illustration for the computation process of SERT.

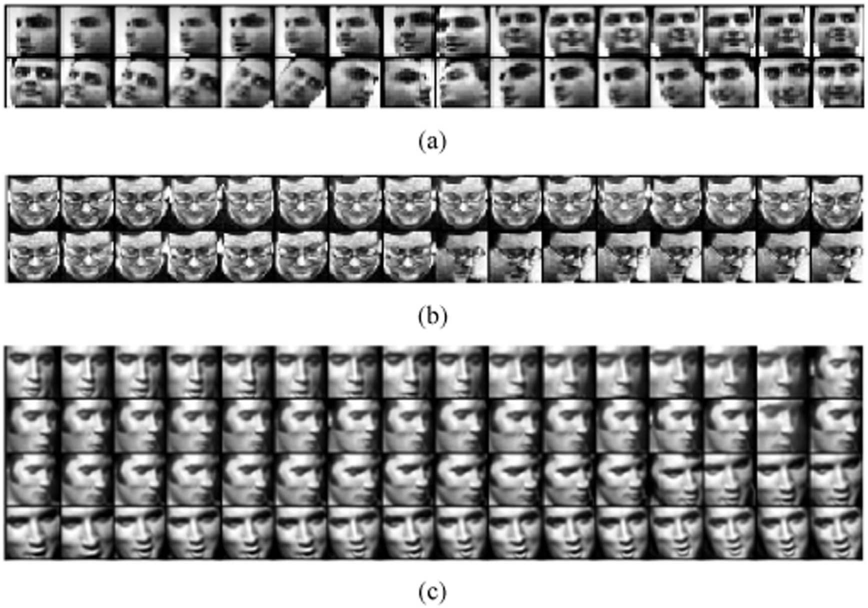
sets easily; (2) it can greatly reduce the number of binary classifiers and the number of images used for training. As for the second advantage, RT trains only one binary classifier; however, traditional multiclass classification strategies such as *one-vs-one* and *one-vs-all* train $c(c-1)/2$ and c binary classifiers respectively, which is far more than that of RT. The flowchart of the proposed SERT approach is presented in Fig. 4.

4. Experimental results and discussions

4.1. Datasets and settings

In order to demonstrate the superiority of the proposed method over the other competitors, experiments need to be performed on benchmark datasets. In the field of image set classification, there exists several different datasets and they are designed for various application scenarios. Specifically, the Honda/UCSD dataset [42], the CMU Mobo dataset [43], and the YouTube Celebrities dataset [38] are used for testing video-based face classification algorithms while the ETH-80 dataset [44] is commonly used as an object recognition benchmark. Since our proposed method SERT is designed mainly for classification of facial images, we conducted experiments on Honda/UCSD, CMU Mobo, and YouTube Celebrities. Below, we will first give a brief description of each of these three datasets followed by the adopted experimental configurations. Then, we will present a performance comparison of the proposed method with the other competitors.

The Honda/UCSD dataset [42] contains 59 video sequences involving 20 different persons. The number of frames for each video ranges from 12 to 645. The face in each frame is first automatically extracted using Viola and Jones face detection algorithm [45] and then resized to the size of 20×20. In our experiment, one video is considered as an image set. Specifically, each person has one image set as the gallery and the remaining sets as the probes (in other words, 20 sequences for training and 39 for testing). Histogram equalization is the only preprocessing procedure we made in all three datasets. k is set as 4 for LBP block division. We repeat our experiment 10 times with randomly selected training and testing combinations. Some examples of an image set from the Honda/UCSD dataset are shown in Fig. 5(a).


(a)

(b)

(c)

Fig. 5. Examples of (a) Honda/UCSD, (b) CMU Mobo, and (c) YouTube Celebrities datasets.

The CMU Mobo (Motion of Body) dataset [43] consists of 96 video sequences of 24 different subjects walking on a treadmill. Each subject has 4 sequences corresponding to four different walking patterns: slow walk, fast walk, incline walk and walking with a ball. The number of frames for each video is about 300. Similar to the Honda, the faces are detected using [45] and resized to 40×40. As a convention, we consider one video as an image set and select one set per person for training and the rest sets for testing (24 sets for training and 72 for testing). We set k=5 for LBP block division. We conduct ten-fold experiments by randomly selected gallery/probe combinations. Fig. 5(b) shows part of one person's video sequences from the CMU Mobo dataset.

The YouTube Celebrities [38] has 1910 video clips of 47 celebrities. This dataset is collected from YouTube and the videos are acquired under real-life scenarios. Consequently, the faces in this dataset exhibit a wide range of diversity and appearance variations in the form of changing illumination conditions, different head poses, and expression variations. The algorithm in [45] fails to detect the correct face in many frames due to its poor resolution and large pose and expression variations. We utilize the method in [46] to track the face region across the entire video, in which the face bounding boxes in initial frame is manually marked and provided along with the dataset. The cropped face region is then resized to 30×30. k=4 for LBP grid division. As a baseline performance measure, we treat the frames in each video as an image set and conduct the five-fold cross validation experiments similar to [15,23,28,36]. Specifically, we divided the whole dataset into five equal folds with minimal overlapping. From the aspect of fold division, for subjects who have more than 45 videos, we randomly select 45 from them. As for subjects who don’t have 45 videos, some videos are selected more than once. Then we divide 45 videos per person into 5 fold. Each fold covers all 47 persons and each person has 9 image sets in one fold. 3 image sets per person are randomly selected for training and the remaining 6 are used for testing (141 sets for training and 282 sets for testing). Fig. 5(c) shows some examples from one video clip.

4.2. Comparisons with existing methods

We compared our proposed framework with several recently proposed state-of-the-art methods which included Discriminant Canonical Correlation analysis (DCC) [18], Manifold-to-Manifold Distance (MMD) [23], Manifold Discriminant Analysis (MDA) [28], Affine Hull based Image Set Distance (AHISD) [14], Convex Hull based Image Set Distance (CHISD) [14], Sparse Approximated Nearest Point (SANP) [15], Covariance Discriminative Learning (CDL) [31] and Reverse Training (RT) [36]. We used the implementations provided by the respective authors for all these methods. Table 2 tabulates the recognition results for our approach and all the other methods listed above on the three datasets. In Table 2, the numbers in each field indicate the average classification accuracy and the standard deviation obtained in multifold cross-validation experiments. The experimental results clearly demonstrate that the proposed approach performs consistently better than the other state-of-the-art methods. Actually, the proposed method SERT is an extension of RT. Compared with RT, the novelty of SERT mainly lies in that unseen facial examples are synthesized to enrich the variety of the image set. The classification accuracy of SERT is higher than RT, especially on CMU Mobo and Youtube datasets, indicating that the proposed sample synthesizing strategy is quite effective to deal with the image set based face classification problem.

In order to compare the computational complexity of different methods, the time cost consumed for one classification operation by each method was also evaluated. CMU Mobo was used for this experiment. Results were obtained on a workstation with an Intel i7-5960X CPU and 64 G RAM. The software platform was Matlab2015a. Results are summarized in Table 3. From Table 3 it can be seen that with respect to the running speed, RT [36] runs the fastest while the proposed method SERT can rank the second. Both RT and SERT can run much faster than the other competitors.

Table 2
Average recognition rates (%) with standard deviation of different methods on the three benchmark datasets.

Method	*Honda/UCSD*	*CMU Mobo*	*YouTube*
DCC [18]	92.6 ± 2.3	88.9 ± 2.5	64.8 ± 2.1
MMD [23]	92.1 ± 2.3	92.5 ± 2.9	62.9 ± 1.8
MDA [28]	94.4 ± 3.4	90.3 ± 2.6	66.5 ± 1.1
AHISD [14]	91.3 ± 1.8	88.5 ± 3.3	64.4 ± 2.4
CHISD [14]	93.6 ± 1.6	95.7 ± 1.0	63.4 ± 2.9
SANP [15]	95.1 ± 3.1	95.6 ± 0.9	65.6 ± 2.4
CDL [31]	98.9 ± 1.3	88.7 ± 2.2	68.5 ± 3.3
RT [36]	100 ± 0.0	97.3 ± 0.6	76.9 ± 2.0
SERT	**100 ± 0.0**	**98.2 ± 1.1**	**80.5 ± 2.4**

Table 3
Time cost (seconds) of a classification operation by different methods on CMU Mobo.

Method	Time cost
DCC [18]	3.730
MMD [23]	0.948
MDA [28]	0.727
AHISD [14]	8.071
CHISD [14]	27.775
SANP [15]	8.674
CDL [31]	0.663
RT [36]	0.418
SERT	0.599

5. Conclusions

One challenging problem in image set based classification is that the poses or the viewpoints of the query set are not covered by the corresponding training image set. To solve this issue, in this paper, we proposed to synthesize multiple virtual instances so as to enrich the variety of the dataset. With this simple technique, the recognition rate of image set based face recognition can be enhanced. With respect to the classification scheme, we resort to reverse training. The proposed approach is named as Synthetic Examples based Reverse Training, SERT for short. SERT was evaluated on three benchmark image set datasets designed for video-based face recognition applications and the experimental results indicate that it can yield better performance than the other competitors.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant no. 61201394, the CCF-Tencent Open Research Fund under Grant no. CCF-Tencent RAGR20150112, and the Shanghai Pujiang Program under Grant no. 13PJ1433200.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Good practice in large-scale learning for image classification, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 507–520.
[2] X. Shi, Z. Guo, Z. Lai, Y. Yang, Z. Bao, D. Zhang, A framework of joint graph embedding and sparse regression for dimensionality reduction, IEEE Trans. Image Process. 24 (4) (2015) 1341–1355.
[3] F. Liu, D. Zhang, L. Shen, Study on novel curvature features for 3D fingerprint recognition, Neurocomputing 168 (1) (2015) 599–608.
[4] X. Wang, D. Huang, A novel density-based clustering framework by using level set method, IEEE Trans. Knowl. Data Eng. 21 (11) (2009) 1515–1531.
[5] X. Wang, D. Huang, H. Xu, An efficient local Chan-Vese model for image segmentation, Pattern Recognit. 43 (3) (2010) 603–618.
[6] Z. Zhao, D. Huang, B. Sun, Human face recognition based on multiple features using neural networks committee, Pattern Recognit. Lett. 25 (12) (2004) 1351–1358.
[7] B. Li, D. Huang, Locally linear discriminant embedding: an efficient method for face recognition, Pattern Recognit. 41 (12) (2008) 3813–3821.
[8] X. Shi, Z. Guo, Z. Lai, Face recognition by sparse discriminant analysis via joint $L_{2,1}$-norm minimization, Pattern Recognit. 47 (7) (2014) 2447–2453.
[9] M. Yang, P. Zhu, F. Liu, L. Shen, Joint representation and pattern learning for robust face recognition, Neurocomputing 168 (1) (2015) 70–80.
[10] M. Hayat, M. Bennamoun, S. An, Deep reconstruction models for image set classification, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2015) 713–727.
[11] M. Hayat, M. Bennamoun, S. An, Learning non-linear reconstruction models for image set classification, in: Proceedings of the CVPR, 201(4), pp. 1915–1922.
[12] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, T. Darrell, Face recognition with image sets using manifold density divergence, in: Proceedings of the CVPR, 2005, pp. 581–588.
[13] G. Shakharovich, J.W. Fisher, T. Darrell, Face recognition from long-term observations, in: Proceedings of the ECCV, 2002, pp. 851–865.
[14] H. Cevikalp, B. Triggs, Face recognition based on image sets, in: Proceedings of the CVPR, 2010, pp. 2567–2573.
[15] Y. Hu, A.S. Mian, R. Owens, Face recognition using sparse approximated nearest points between image sets, IEEE Trans. Pattern Anal. Mach. Intell. 34 (10) (2012) 1992–2004.
[16] A. Hadid, M. Pietikainen, From still image to video-based face recognition: an experimental analysis, in: Proceedings of the International'l Conference Automatic Face and Gesture Recognition, 2004, pp. 813–818.
[17] M. Yang, P. Zhu, L. Gool, L. Zhang, Face recognition based on regularized nearest points between image sets, in: Proceedings of the International'l Conference Automatic Face and Gesture Recognition, 2013, pp. 1–7.
[18] T.K. Kim, J. Kittler, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 1005–1018.
[19] O. Yamaguchi, K. Fukui, K. Maeda, Face recognition using temporal image sequence, in: Proceedings of the International'l Conference Automatic Face and Gesture Recognition, 1998, pp. 318–323.
[20] L. Wolf, A. Shashua, Learning over sets using kernel principal angles, J. Mach. Learn. Res. 4 (10) (2003) 913–931.
[21] K. Lee, J. Yamaguchi, The kernel orthogonal mutual subspace method and its application to 3D object recognition, in: Proceedings of the ACCV, 2007, pp. 467–476.
[22] M. Nishiyama, O. Yamaguchi, K. Fukui, Face recognition with the multiple constrained mutual subspace method, in: Proceedings of the AVBPA, 2005, pp. 71–80.
[23] R. Wang, S. Shan, X. Chen, W. Gao, Manifold-manifold distance with application to face recognition based on image set, in: Proceedings of the CVPR, 2008, pp. 1–8.
[24] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara, O. Yamaguchi, Recognizing faces of moving people by hierarchical image-set matching, in: Proceedings of the CVPR, 2007, pp. 1–8.
[25] T. Kim, J. Kittler, R. Cipollar, Incremental learning of locally orthogonal subspaces for set-based object recognition, in: Proceedings of the BMVC, 2006.
[26] W. Fan, D. Yeung, Locally linear models on face appearance manifolds with application to dual-subspace based classification, in: Proceedings of the CVPR, 2006, pp. 1384–1390.
[27] M.T. Harandi, C. Sanderson, S. Shirazi, B.C. Lovell, Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching, in: Proceedings of the CVPR, 2010, pp. 2705–2712.
[28] R. Wang, X. Chen, X, Manifold discriminant analysis, in: Proceedings of the CVPR, 2009, pp. 429–436.
[29] T. Wang, P. Shi, Kernel Grassmannian distances and discriminant analysis for face recognition from image sets, Pattern Recognit. Lett. 30 (13) (2009) 1161–1165.
[30] A.W. Fitzgibbon, A. Zisserman, Joint manifold distance: a new approach to appearance based clustering, in: Proceedings of the CVPR, 2003, pp. 26–33.
[31] R. Wang, H. Guo, L.S. Davis, Q. Dai, Covariance discriminative learning: a natural and efficient approach to image set classification, in: Proceedings of the CVPR, 2012, pp. 2496–2503.
[32] P. Turaga, A. Veeraraghavan, A. Srivastava, R. Chellappa, Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2273–2286.
[33] J. Hamm, D. Lee, Grassmann discriminant analysis: a unifying view on subspace-based learning, in: Proceedings of the ICML, 2008, pp. 376–383.
[34] M.T. Harandi, C. Sanderson, A. Wiliem, B.C. Lovell, Kernel analysis over Riemannian manifolds for visual recognition of actions, pedestrians and textures, Proc. IEEE Workshop Appl. Comput. Vis. (2012) 433–439.
[35] P. Zhu, L. Zhang, W. Zuo, D. Zhang, From point to set: extend the learning of distance metrics, in: Proceedings of the ICCV, 2013, pp. 2664–2671.
[36] M. Hayat, M. Bennamoun, S. An, Reverse training: an efficient approach for image set classification, in: Proceedings of the ECCV, 2014, pp. 784–799.
[37] S.A.A. Shah, M. Bennamoun, F. Boussaid, Iterative deep learning for image set based face and object recognition, Neurocomputing 174 (1) (2016) 866–874.
[38] M. Kim, S. Kumar, V. Pavlovic, H. Rowley, Face tracking and recognition with visual constraints in real-world videos, in: Proceedings of the CVPR, 2008, pp. 1–8.
[39] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[40] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (1) (2002) 321–357.
[41] R. Fan, K. Chang, C. Hsieh, X. Wang, C. Lin, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[42] K. Lee, J. Ho, M. Yang, D. Kriegman, Video based face recognition using probabilistic appearance manifolds, in: Proceedings of the CVPR, 2003, pp. 313–320.
[43] R. Gross, J. Shi, The CMU motion of body (MOBO) database, Technical Report CMU-RI-TR-01-18, 2001.
[44] B. Leibe, B. Schiele, Analyzing appearance and contour based methods for object categorization, in: Proceedings of the CVPR, 2003, pp. 409–415.
[45] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[46] D.A. Ross, J. Lim, R. Lin, M. Yang, Incremental learning for robust visual tracking, Int. J. Comput. Vis. 77 (1) (2008) 125–141.

Lin Zhang received the B.Sc. and M.Sc. degrees from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2003 and 2006, respectively. He received the Ph.D. degree from the Department of Computing, the Hong Kong Polytechnic University, Hong Kong, in 2011. From March 2011 to August 2011, he was a Research Assistant with the Department of Computing, the Hong Kong Polytechnic University. In Aug. 2011, he joined the School of Software Engineering, Tongji University, Shanghai, China, where he is currently an Associate Professor. His current research interests include biometrics, pattern recognition, computer vision, and perceptual image/video quality assessment.

Qingjun Liang received the B.S. degree from the School of Software Engineering, Tongji University in 2013. He is now pursuing her M.S. degree at the same department. Her research interests are biometrics and machine learning.

Ying Shen received the B.S. and M.S. degrees from the Software School, Shanghai Jiao Tong University, Shanghai, China, in 2006 and 2009, respectively. She received the Ph.D. degree from the Department of Computer Science, City University of Hong Kong, Hong Kong, in 2012. In 2013, she joined the School of Software Engineering, Tongji University, Shanghai, China, and currently is an assistant professor. Her research interests include bioinformatics and pattern recognition.

Meng Yang is currently an associate professor at School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He received his Ph.D. degree from The Hong Kong Polytechnic University in 2012. Before joining Shenzhen University, he has been working as a Postdoctoral fellow in the Computer Vision Lab of ETH Zurich. His research interest includes sparse coding, dictionary learning, object recognition and machine learning. He has published 9 AAAI/CVPR/ICCV/ECCV papers and several IJCV, IEEE TNNLS and TIP journal papers.

Feng Liu currently is an assistant professor at School of Computer Science and Software Engineering, Shenzhen University. She received her Ph.D. degree from the Hong Kong Polytechnic University in 2014. Her research interests include pattern recognition and image processing, especially focus on their applications to fingerprints.