Neurocomputing 341 (2019) 60–69

Contents lists available at ScienceDirect

Neurocomputing

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

Image set classification using candidate sets selection and improved reverse training

Zhenwen Ren[a,b], Bin Wu[c], Xiaoqian Zhang[b], Quansen Sun[a,*]

[a] Department of Computer Science, Nanjing University of Science and Technology, Nanjing 210094, China
[b] School of National Defence Science and Technology, Southwest University of Science and Technology, Mianyang 621010, China
[c] School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China

ARTICLE INFO

Article history:
Received 23 April 2018
Revised 18 February 2019
Accepted 8 March 2019
Available online 12 March 2019

Communicated by Dr. H. Yu

Keywords:
Image set classification
Face recognition
Object recognition
Candidate sets selection
Reverse training

ABSTRACT

Image set classification is recently a competitive technique, of which each gallery set and query set contains multiple images of a person or an object. However, strong noise and outlier could be generated from undesirable environmental conditions in real-world applications, which will result in deterioration of classification accuracy, robustness, and high time complexity of special learning algorithm. To address these problems, this paper presents a novel parameterless method called *Candidate Sets Selection and Improved Reverse Training (CSSIRT)*. It aims at enhancing noise robustness and reducing time complexity while retaining the high classification accuracy via two-stage successive cascaded framework. Specifically, an image set is modeled as a robust reduced affine hull, which provides a better representation to account for the unseen appearances of each image set. The first stage selects the most promising top-*N* candidate sets and rejects the vast majority of the impossible gallery sets quickly by measuring dissimilarity between two hulls. Since few returned candidate sets from the first stage may intersect the given test one, the second stage utilizes the returned candidate sets to address the issue on sets overlapped by using a novel *K*-means clustering and normalized frequency histograms fusion based reverse training algorithm. The method we proposed is more robust, has a lower time complexity and higher accuracy. Three kinds of extensive comparisons with the other state-of-the-art methods have corroborated the superiority of our method on a number of challenging datasets.

1. Introduction

Image set classification has been one of the most important tasks in computer vision [1–8] due to its broad applications in various areas, including multi-view visual recognition, video-based surveillance, video retrieval, dynamic scene recognition, etc., where multiple images are readily available. Different from traditional single-shot image classification [9,10] task, image set classification is more promising since it aims to effectively deal with a variety of appearance variations for improving the accuracy of discrimination and the robustness to image variations. These variations could be caused by pose, illumination, non-rigid deformation, obscuration and misalignment. For these advantages, classification based on image set has obtained significant attentions and developments in recent years. However, image set shows huge intra-class variability and large inter-class ambiguity, which also poses great challenges to make effective use of the potential supplemental information and faithfully measure the dissimilarity between image sets for accurate classification.

Generally speaking, there are two major steps involved in image set classification, which are to effectively model an image set and to define an appropriate metric to compute the dissimilarity between two sets. According to model types, relevant methods mainly fall into five categories [11]: statistical model based methods [12–14], linear subspace based methods [15], nonlinear manifold based methods [16–21], affine subspace based methods [7,13,22–24,26], compressed sensing based methods [1,19,26,27,29]. Besides, deep learning has recently gained significant success in some tasks [28], but its applications to image set classification is few, the most recent articles are [30,31].

Based on a comprehensive literature review of the line about affine hull model [7,13,19,22–26], we know that affine hull can effectively characterize the large variations in an image set, which tends to represent the unseen appearances in the set via linear combinations of all images of the set. Nevertheless, there are some limitations due to the presence of strong noises and outliers from

* Corresponding author.
E-mail addresses: sunquansen@njust.edu.cn, rzw@njust.edu.cn (Q. Sun).

https://doi.org/10.1016/j.neucom.2019.03.010
0925-2312/

undesirable environmental conditions: (1) The affine hull may be overlarge when image sets contain noises or outliers [22]. Therefore, we usually need to develop complex iteration algorithm to alleviate/remove the influence of noises and outliers, such as prototype (or nearest points) learning [5,13,25], discriminant projection (or metric) learning [23,25], sparse representation [1,19,26,27], low-rank representation [29], collaborative representation [2] etc. (2) The relevant learning method is hard to solve with the possibility of trapping in local optimum and high time complexity. To significantly reduce the computation complexity, candidate sets selection is a very effective strategy [32,33]. For example, Cevikalp and Serhan Yavuz [32] use Extended Polyhedral Conic Classifier (EPCC) to find the closest candidate sets from the gallery. Xu et al. [33] use the effect on representing the test sample of the kth class to coarsely determine a small number of candidate classes. (3) The learning methods need some arguments about data distribution assumptions, model representations, learning factors, and experimental settings. Especially the deep learning methods, they require more parameters [30,31].

Most of the previous methods of this line were developed under certain specific assumptions [13] (e.g., the image set is pure, and the data obeys a pre-specified distribution), which would not hold in practice or be shared across different applications. Unlike them, we propose a novel method called Candidate Sets Selection and Improved Reverse Training (CSSIRT), which aims at enhancing noise robustness and reducing time complexity while retaining the high classification accuracy via two-stage successive cascaded framework. Specifically, we model an image set as a robust reduced affine hull, and then resort to compute the dissimilarity between all the gallery sets to the query set respectively. Hence, the dissimilarity can be used for finding the closest candidate sets so that the first stage can return the few promising top-N candidate gallery sets and reject the majority of impossible gallery sets quickly. With respect to the demand of high classification accuracy and noise robustness in real applications, several candidate affine hulls may intersect with the given probe one, i.e., the probe set has near-zero distances to the returned gallery sets. Therefore, the second stage firstly divides the candidate gallery sets into two parts, and then the query set can be optimally classified to the same class of one gallery set by using the improved reverse training algorithm. The stage is very fast due to it does not use all the gallery sets but few candidate gallery sets, and thus better performance is achieved. It will enable us to cope with real-time and large-scale face and object recognition tasks.

The main contributions of this paper are as follows:

(1) A novel two-stage cascaded framework is proposed for robust image set classification. Our method is categorized as online method which does all computations at run time. Moreover, another strength of our method is excellent scalability, which means that new classes can easily be added.
(2) To avoid the interference from the images of impossible sets and reduce time complexity, a Top-N algorithm is introduced for selecting the most promising candidate gallery sets.
(3) To further improve the robustness, a voting strategy and histograms fusion based improved reverse training algorithm is proposed, which introduces a K-means clustering algorithm instead of randomly selecting strategy to capture the structure of a set.
(4) In order to evaluate the performance of CSSIRT, three kinds of scenarios are elaborated, including origin clean datasets, image sets contaminated by face detection errors, and image sets contaminated by images from other classes.

We continue the paper as follows. Section 2 presents our method in detail. Section 3 presents and analyzes the experimental results. Section 4 explains the reason why our method achieves superior performance. Section 5 is the conclusion of this paper.

2. Proposed method

Inspired by both affine hull model [22] and pseudo-label based reverse training method [34–36], we proposed the Candidate Sets Selection and Improved Reverse Training (CSSIRT) method for image set classification, which is a two-stage cascaded framework. In summary, our method contains three sub-algorithms: *Top-N Promising Candidate Sets Selection (CSS)* (Section 2.2), *Extraction of Subsets (EOS)* (Section 2.3.1), and *Improved Reverse Training (IRT)* (Section 2.3.2), which are visualized in (Figs. 1 and 2).

2.1. Reduced affine hull representation

In a M classes of image sets classification scenario, we have an unlabelled probe set S_Q for testing and a set of M labeled gallery image sets $D = \{S_1, S_2, \ldots, S_M\}$ with class labels $y \in [1, 2, \ldots, M]$, respectively, for training. The cth gallery image set denotes as $S_c = \{I_{ck}\}_{k=1}^{n_c}$, where $I \subset \mathbb{R}^d$ is the d-dimensional feature vector of the kth image belonging to class c. In image set classification, the goal is to predict which gallery set belongs to the same class as the probe set S_Q.

In this work, an image set is represented as a reduced affine hull [22], which provides better localization to account for the unseen appearances of each image set. The dissimilarity measure between two sets can be defined as the distance between a pair of nearest points from two hulls, respectively, which can be considered as an enhancement of nearest neighbour (NN) classifier with an attempt to reduce the sensitivity of within-class variations by artificially generating samples within the set.

In this way, the cth image set $S_c \in D$ can be approximatively represented by an affine hull H_c^{aff}, which contains all liner combinations of the elements in S_c, and can be written as:

$$H_c^{aff} = \{x\}, \forall x = \sum_{k=1}^{n_c} \omega_{ck} I_{ck}, \sum_{k=1}^{n_c} \omega_{ck} = 1, \tag{1}$$

where ω is the affine coefficient vector. Note that Eq. (1) is a constrained representation, while an unconstrained representation can be given by

$$H_c^{aff} = \{x\}, \forall x = \mu_c + U_c v_c, v_c \in \mathbb{R}^{l_c}, \tag{2}$$

where $\mu_c = \frac{1}{n_c}\sum_{k=1}^{n_c} I_{ck}$ is the mean of samples in the cth image set S_c, U_c is an orthonormal matrix which is obtained by applying thin Singular Value Decomposition (SVD) to $[I_{c1} - \mu_c, \ldots, I_{cn_c} - \mu_c]$, v_c is a vector of free parameters. It provides reduced coordinates for the points within the affine subspace, and $l_c (l_c < n_c)$ is the number of the singular vectors of U_c by discarding vectors corresponding to near-zero singular values i.e., removing spurious noise dimensions within data. In this paper, the hull H_c^{aff} is called as reduced affine hull due to $l_c < n_c$.

According to [22], the dissimilarity between two sets is defined as the distance between a pair of nearest points belonging to either reduced affine hull by

$$dist(H_i^{aff}, H_j^{aff}) = \min_{y,z} \|y - z\|_2^2, y \in H_i^{aff}, z \in H_j^{aff}, \tag{3}$$

Following Eq. (2), let $y = \mu_i + U_i v_i$ and $z = \mu_j + U_j v_j$, then the problem (3) can be rewritten as follow:

$$dist(H_i^{aff}, H_j^{aff}) = \min_{v_i, v_j} \left\| (\mu_i + U_i v_i) - (\mu_j + U_j v_j) \right\|_2^2. \tag{4}$$

By defining $U = (U_i, -U_j)$ and $v = \binom{v_i}{v_j}$, problem (4) becomes a standard least squares problem, i.e.,

$$\min_{v} \left\| Uv + \mu_i - \mu_j \right\|_2^2, \tag{5}$$

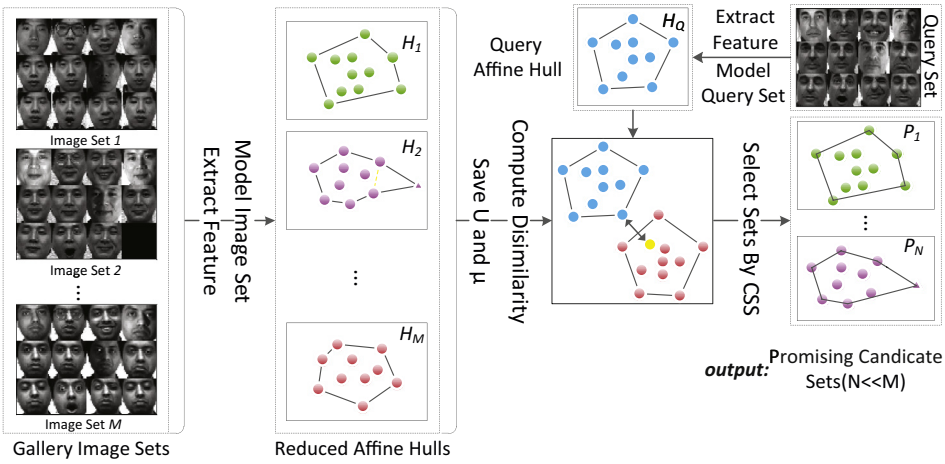


Fig. 1. Framework of the proposed CSS method for selecting candidate gallery sets (first stage). Firstly, the reduced affine hull is used for characterizing the large variations in the gallery sets and query set. Then, calculate the dissimilarity between the query set and each gallery, respectively, which is defined as the distance between a pair of nearest points belonging to either hull, respectively. Next, we select the most promising top-*N* candidate gallery sets and reject the majority of the impossible gallery sets.

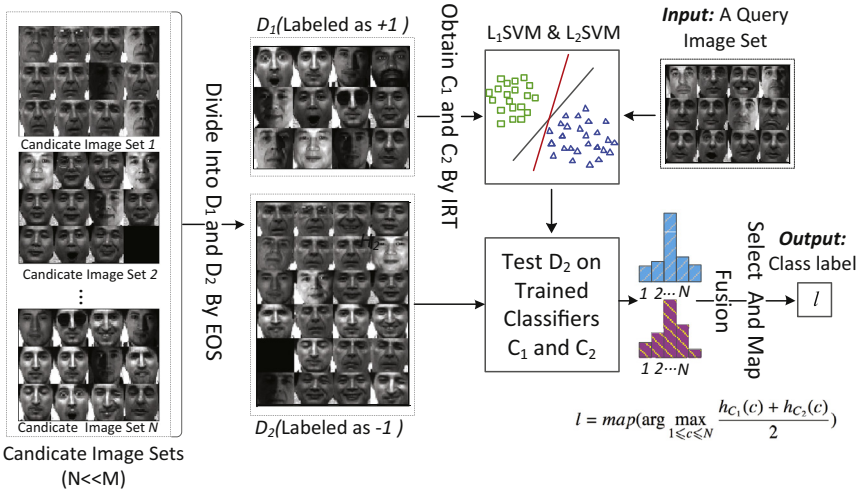


Fig. 2. Framework of the proposed EOS+IRT scheme for training and testing (second stage). Firstly, the returned candidate gallery sets are divided into two subsets D_1 and D_2 by using *K*-means clustering algorithm. Based on query set S_Q (labeled $+1$) and D_1(labeled -1), two specific linear classifiers are obtained by using improved reverse training algorithm, and then test the images of D_2 on them. The class label of S_Q is decided by using voting strategy and fused the normalized frequency histograms from the two classifiers.

whose analytical solution can be computed as follow:

$$v = (U^T U)^{-1} U^T (\mu_i - \mu_j). \tag{6}$$

Therefore, the Eq. (4) can be rewritten as

$$dist(H_i^{aff}, H_j^{aff}) = \left\| I - U(U^T U)^{-1} U^T (\mu_i - \mu_j) \right\|. \tag{7}$$

From the experimental results in [22], we know reduced affine hull model has certain robustness to the noise case especially outliers. However, when the nearest point is corrupted by strong noise (e.g., the whole features are polluted or the point is from other classes), the hull is too lose to achieve good discrimination information i.e., having overlapped area between probe hull and multiple gallery hulls.

2.2. *Top*-N *promising candidate sets selection*

As above, the affine hull matching fails when several gallery hulls overlap with the given probe one. In this section, we propose the candidate sets selection algorithm, which aims to obtain the most promising candidate gallery sets and rejects the vast majority of the impossible gallery sets quickly. Here, the promising candidate gallery sets is the candidate image sets returned by the CSS algorithm (i.e., the first stage of our framework), which contains the ground truth set and few deceitful ones caused by outlier and noise. The impossible gallery sets are the clearly known impossible gallery sets that can not be the target, which are discarded by the CSS algorithm. The main advantage is that it excludes the impossible sets so that significantly reduces the computational complexity of the subsequent algorithm, and to avoid interference from images of impossible sets.

Assuming the reduced affine hull of the probe image set is H_Q^{aff} (i.e., the probe hull), and the reduced affine hulls of the M gallery image sets are $H_1^{aff}, H_2^{aff}, \ldots, H_M^{aff}$ (i.e., the gallery hulls), respectively. Let $P = \{P_1, P_2, \ldots, P_N\}$ be a collection of the candidate sets that needs to be learned, and we select some of these gallery hulls according to their distance to the reference probe hull H_Q^{aff}. That is, we can get the index of the most promising candidate gallery set $\hat{k}$ by

$$\hat{k} = \arg\min_k dist(H_Q^{aff}, H_k^{aff}), k \in [1, 2, \ldots, M]. \tag{8}$$

Once we determine $\hat{k}$, the $\hat{k}$th set is the most promising one. According to Eq. (8), we select the most promising $N(s.t. N << M)$ candidate gallery sets by repeating it. Notice that the set had been

selected which should no longer be considered in the remainder of decision. The entire process can be optimized by Algorithm 1. Next,

Algorithm 1 Top-N Promising Candidate Sets Selection (CSS).

Input:
The size of input gallery sets: C.
The expected size of candidate gallery sets: N.
The gallery image sets: $S_1, S_2, \ldots, S_C$.
The reduced affine hulls: $H_1^{aff}, H_2^{aff}, \ldots, H_C^{aff}$.
The reduced affine hull of query set: H_Q^{aff}.
Output:
The candidate gallery sets: $P_1, P_2, \ldots, P_N$.
$k \leftarrow 1$,$d[M] \leftarrow inf$.
for $k = 1$ to M **do**
  %Calculate the dissimilarity between the probe hull and the kth gallery hull.
  $d[k] = dist(H_Q^{aff}, H_k^{aff})$.
end for
%Rearrange in ascending order.
$sort(d, asc)$.
%Select N promising candidate sets.
for $i = 1$ to N **do**
  $l \leftarrow$ The class label of H_i^{aff}.
  $P_i = S_l$.
end for
return $P_1, P_2, \ldots, P_N$.

the following improved reverse training algorithm can be both very fast and accurate for image set classification by utilizing learned P.

2.3. Improved reverse training for image set classification

After getting the most promising candidate gallery sets P, a novel algorithm based on the reverse training method [35] is used for classification. The reverse training method efficiently extents binary classifiers for the task of multi-class image set classification. It trains a binary classifier by all the images of the query set and a randomly sampled subset of all the gallery sets, and then evaluates on the rest of training images of the gallery sets by using the learned classifier to predict the label of the query set [35]. It has many advantages in terms of robustness, accuracy, and time complexity.

Compared with it, our improved reverse training algorithm consists of two main steps, including extraction of subsets (EOS) and improved reverse training (IRT). EOS adopts K-means clustering algorithm instead of randomly selecting samples to divide the candidate gallery sets into two subsets (i.e., the training subset and testing subset). In IRT, it trains fewer points due to only few candidate gallery sets. More specifically, The training subset is feeded into two SVM classifiers with different regularization terms. Then, two normalized frequency histograms from the two binary classifiers are fused, the label of the query set is subsequently obtained by using the voting strategy. Therefore, it is an effective method to significantly enhance the robustness to noise and significantly reduce computational complexity.

2.3.1. Extraction of subsets (EOS)

Reverse training requires to split the candidate gallery sets into two subsets, one of which is used for training two classifiers, and the other is used for testing and predicting the class label of the probe image set.

To solve the extraction of subsets (EOS) problem for reverse training, the candidate gallery sets are gathered into a single set $P = \{P_1, \ldots, P_N\}$ with class labels $y \in [1, \ldots, N]$, where N is the number of the candidate gallery sets. Note that, a function map is designed to map the class label from candidate gallery set to the origin gallery set. Then, D is divided into two subsets D_1 and D_2 by adopting K-means clustering algorithm instead of random algorithm to select samples from each candidate set. Compared with the random selection strategy, the K-means algorithm takes into account the structure and distribution information of the samples. With respect to the size of N_{D_1} and N_{D_2}, let $N_{D_1} = N_q$ and $N_{D_2} = \sum_{c=1}^{N} n_c - N_{D_1}$.

Specifically, our algorithm starts with an initial random selection of k images from each candidate gallery set as initial centroids. Then it assigns each image to its closest centroid and recomputes the centroid of each cluster until centroids do not change for each set. Next, the selected k images form a subset D_{1c}. Note that D_{1c} is a subset of S_c with a set size $k = \left\lfloor \frac{N_q}{N} \right\rfloor$, where N_q is the size of the query image set. Next, those images of subset D_{1c} are grouped into the set D_1, and the rest images are grouped into the set D_2. This is given by

$$\begin{aligned} &D_1 = D_{11} \cup D_{12} \cup \ldots \cup D_{1N}, D_2 = D \setminus D_1 \\ &y_{D_1} = \{y^{(t)} \in [1, 2, \ldots, N], t = 1, 2, \ldots, N_{D_1}\} \\ &y_{D_2} = \{y^{(t)} \in [1, 2, \ldots, N], t = 1, 2, \ldots, N_{D_2}\}, \end{aligned} \tag{9}$$

where y_{D_1} and y_{D_2} are the class labels of images in D_1 and D_2, respectively. The entire process is presented in Algorithm 2.

Algorithm 2 Extraction of Subsets (EOS).

Input:
The gallery image sets: D.
The number of candidate sets: N.
The number of images in S_Q: N_q.
Output:
The subsets D_1 and D_2.
$k \leftarrow \left\lfloor \frac{N_q}{N} \right\rfloor$, $D_1 \leftarrow \emptyset$.
for $i = 1$ to N **do**
  Select k points as initial centroids, the ith centroid is D_{1i}.
  repeat
    Assign each point to its closest centroid.
    Recomputed the centroid D_{1i} of each cluster.
  until Centroids do not change.
  $D_1 = D_1 \cup D_{1i}$, $D_2 = D \setminus D_1$.
end for
return D_1 and D_2

In doing so, D_1 contains an equal representation of images from all classes of the candidate gallery sets and the total number of images in D_1 is approximately equal the number of images of the query image set S_Q.

2.3.2. Improved Reverse Training (IRT)

We treat the images both in probe image set S_Q and subset D_1 as final training data and the images in D_2 as testing data. The proposed improved reverse training algorithm is presented in Algorithm 3 and the details are summarized below.

First of all, all images in S_Q are labelled $+1$, while the images in D_1 are labelled -1. Then we obtain the robust binary classifiers C_1 and C_2 by training all images of S_Q and D_1. Since images from all classes are present in D_1, in which a few images have the same class as of S_Q, C_1 and C_2 treat them as outliers to separate images of S_Q from the images of the other classes. For the two classifiers, C_1 is the linear Support Vector Machine (SVM) classifier with L_2 regularization and L_2 loss function, and C_2 is the SVM with L_1 regularization and L_2 loss function [37]. Specifically, given a set of training pairs $T = \{(x, y) \mid y \in [+1, -1]\}$ which represents an image with known class label information. C_1 solves the following

Algorithm 3 Improved reverse training (IRT).

Input:
The gained two subsets by EOS: D_1 and D_2.
The probe set: S_Q.
The candidate sets: P.
Output:
The class label l of S_Q.
1: S_Q labeled $+1$ and D_1 labeled -1.
2: C_1 and C_2 $\leftarrow$ Train on D_1 and S_Q via Eq. (10) and Eq. (11), respectively.
3: h_{C_1} and h_{C_1} $\leftarrow$ Test D_2 on C_1 and C_2.
4: $y_q \leftarrow \arg\max \frac{h_{C_1}+h_{C_2}}{2}$. % see Eq. (14)
5: $l \leftarrow map(y_q)$. % Map class label from P to D.
6: **return** The class label l of S_Q.

optimization problem,

$$\min_w \frac{1}{2}w^T w + C\sum_{\forall(x,y)\in T}(\max(0, 1-yw^T x))^2, \tag{10}$$

while, C_2 solves the following optimization problem,

$$\min_w \|w\|_1 + C\sum_{\forall(x,y)\in T}(\max(0, 1-yw^T x))^2. \tag{11}$$

Here, w is the coefficient vector and $C>0$ is the penalty parameter. To find the best C, LIBLINEAR (after version 2.0) provides an easy solution to find a suitable parameter, the -C option efficiently conducts cross validation several times and finds the best parameter automatically [37].

Next, all images of D_2 are tested by the classifiers C_1 and C_2, and then, after testing all the images from set D_2, two subsets $D_{2C_1} \subset D_2$ and $D_{2C_2} \subset D_2$ are obtained respectively, which are classified as the same class of S_Q (i.e., the images of D_{2C_1} and D_{2C_2} are classified as $+1$). Two normalized frequency histograms h_{C_1} and h_{C_2} of gallery class labels in $y_{D_{2C_1}}$ and $y_{D_{2C_2}}$ are computed respectively. The cth value of the one histogram, $h_C(c)$, is given by the percentage of the images of class c in $y_{D_{2C}}$ which are classified $+1$. Formally, $h_C(c)$ is given by the ratio of the number of images of D_2 belonging to class i and classified as $+1$ to the total number of D_2 belonging to class c. For each value of the one histogram, this is given by,

$$h_C(c) = \sum_{y^{(t)}\in y_{D_{2C}}} \delta_c(y^{(t)}) / \sum_{y^{(t)}\in y_{D_2}} \delta_c(y^{(t)}), \tag{12}$$

where $c \in [1, 2, \ldots, N]$ denotes the cth element of h_{C_1} and h_{C_2}, $y^{(t)}$ is the predicted class label, and δ_c is delta function, i.e.,

$$\delta_c(x) = \begin{cases} 1, & x = c \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Finally, the class in D_2 with most of its images classified as $+1$ can be predicted as the class of S_Q. The class label l of S_Q is therefore given by

$$l = map(\arg\max_{1\leqslant c\leqslant N} \frac{h_{C_1}(c) + h_{C_2}(c)}{2}), \tag{14}$$

where the function *map* (which is explained in Section 2.3.1) is used for maping the class label from the candidate set P to the origin gallery set D.

For better illustration, we take a toy example to help readers understand our method. We have a query set S_Q and 6 gallery sets with 3 possible gallery sets S_1, S_2, S_3 and 3 impossible gallery set S_4, S_5, S_6. All the data points of all the training sets and the query set are shown in Fig. 3(a). Firstly, we use CSS to obtain the most promising top-N candidate gallery sets and reject the majority of the impossible gallery sets. The three candidate gallery sets S_1, S_2, and S_3 are shown in Fig. 3(b). Then, we form D_1 and D_2 by EOS from the candidate gallery sets. Next, two linear classifiers are trained by labeling the data points of S_Q as $+1$ and D_1 as -1. Fig. 3(c) shows the data points of D_1, S_Q and classifiers. Then, we classify the data points of D_2 by the learned classifiers, the points with the same class of S_Q in D_2 as $+1$. Lastly, the class of S_Q is decided by fusing the output histograms of the two classifiers which is shown in Fig. 3(d).

2.4. Optimization tip

For large-scale and real-time face or object image set recognition tasks, a major limitation of online methods (including CSSIRT) is that all the computation is done at run-time and comparatively more memory storage is required. Apart from our elaborate method, we also suggest a tip about memory storage and speed optimization. In Section 2.1, an orthonormal basis matrix U and the sample mean μ are obtained by applying SVD, which can be computed off-line. In addition, we store them instead of all feature vectors of the set's images for measuring the dissimilarity between two sets. Now, we analyze the memory storage performance for an image set $S \in \mathbb{R}^{d\times n}$ under Matlab 2016b. If we store the set directly, it takes d*n*8 bytes. However, if we store the reduced affine hull (i.e., μ and U), the sample mean $\mu \in \mathbb{R}^d$ takes d*1*8 and the orthonormal basis matrix $U \in \mathbb{R}^{d\times l}$ only takes d*l*8, where $l << n$.

3. Experiments

In this section, we compare the results on five databases with several state-of-the-art image set classification methods.

3.1. Datasets and settings

We evaluate the experimental performance of different methods on five datasets, including Honda/UCSD (Honda) [38], CMU MoBo (MoBo) [39], YouTube Celebrities (YTC) [40], ETH-80 [41] and COX [42]. Amongst these widely used datasets, COX and YTC datasets are real-life datasets collected under unconstrained conditions, whereas Honda, MoBo and ETH-80 are relatively easy due to they are acquired in indoor lab environment under controlled conditions. For fair competition, the detected face images are histogram equalized, but no further pre-processing such as background removal or alignment. With respect to feature extraction, we use gray-level values or local binary pattern (LBP) [36] features. The brief descriptions and specific experimental settings of each of these datasets are presented in the following.

The Honda [38] dataset consists 59 video sequences of 20 different persons. Each sequence contains approximately 12–645 frames covering large variations. Face images in the dataset were detected by using the Viola Jones Face Detector [43], and we resize them to 20×20, then all video sequences are divided into two groups, 20 video sequences are used for training and the remaining 39 ones are used for testing. We do not extract LBP features for this dataset since using pixel values already achieved very high accuracies. For each experiment, we randomly select training and testing sets and limit the number of images in each set to 100.

The MoBo [39] dataset has 96 video sequences of 24 persons, which are captured from different walking situations inclined walk, and slow walk holding a ball on a treadmill. We first extract faces by using the Viola Jones Face Detector [43], and resize to 30×30. Each video is further divided into four illumination sets, the first set for training and the rest sets for testing.

The YTC [40] dataset contains 1910 video clips of 47 celebrities (actors and politicians), most of the videos are low resolution

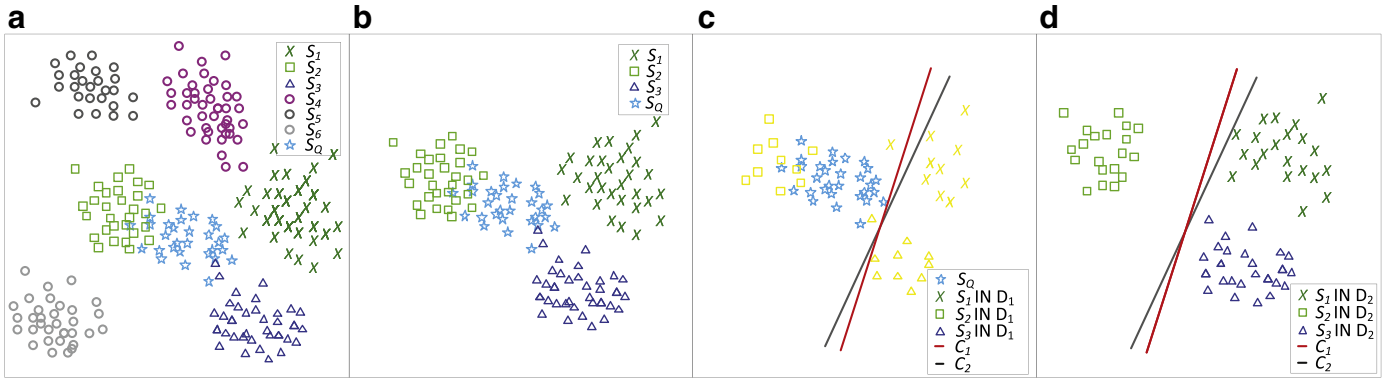


Fig. 3. Toy example to illustrate our method. (a) The data points of 6 gallery sets and 1 query set. (b) The data points of query set and candidate gallery sets S_1, S_2, S_3. (c) The data points of D_1, query set and boundaries between X_Q (labeled +1) and D_1 (labeled −1). (d) The data points of D_2 w.r.t. the learnt classifiers boundary. Since the points of S_2 in D_2 have the highest vote by fusing the output histograms of the two classifiers, therefore, the proposed method declares S_Q to have the same class of S_2.

and highly compressed, which leads to noise, low-quality image frames. Each clip contains hundreds of frames. For this dataset, we use a cascaded face tracking detector [44] to extract faces. The obtained face images are resized to 30×30 gray-scale images and then extract their LBP features. We conduct experiments by randomly choosing 3 sets for training and 6 sets for testing.

The COX Faces dataset [42] contains 1000 high resolution still images and 2000 uncontrolled low resolution video sequences of 1000 subjects and has 3 video sequences for each subject. The videos are captured inside a gymnasium with three fixed camcorders when the subjects walk around the pre-designed S-shape route. The dataset has variations without any restriction on changing illumination conditions,expression variations, different head pose rotations and resolution through this S-shape route. For this dataset, we use LBP features which are extracted from 32×40 face images detected by using a cascaded face detector [44]. We randomly choose 100 frames of each video as an image set and then one image set is held out for testing and remaining two sets are used as gallery sets.

The ETH-80 [41] dataset contains 8 categories and each category contains 10 objects with 41 still RGB images per object, for a total of 3280 images. For our experiments, we crop and resize the 128×128 images to 30×30. For each subject, we randomly choose 5 sets for training and the rest 5 objects for testing.

3.2. Comparative methods

We compare our method with several excellent methods, including Discriminant Canonical Correlation Analysis (DCC) [15], Manifold Discriminant Analysis (MDA) [46], Affine Hull based Image Set Distance (AHISD) [22], Convex Hull based Image Set Distance (CHISD) [22], Sparse Approximated Nearest Points (SANP) [26], Covariance Discriminative Learning (CDL) [12], Deep Reconstruction Models (DRM-WV) [31], Prototype Discriminative Learning(PDL) [23], Reverse training(RT) [35]. We adopt the implementations provided by the respective authors on their homepage for all methods, whose parameters are followed by the recommendations in the original references for best performance.

Specifically, for DCC, the dimensions of the embedding space are set to 100, the number of retained dimensions for a subspace is set to 10 (90% energy is preserved) and the corresponding 10 maximum canonical correlations are used for computing set-set dissimilarity. No parameter settings are required for CDL, AHISD and CHISD. For SANP, the same weight parameters as in [26] are adopted for optimization. The parameters for MDA are used from [46], the number of connected nearest neighbours to compute the geodesic distance is either set to 12 or to the number of images in the smallest image set of the dataset. For DRM, which is a deep learning method, the learning rate, l_2-weight decay, sparsity target, non-sparsity penalty term, batch size, and λ are set as in [31]. For PDL, the initialization of W, P_c, and the varying numbers of images contained in different image sets m_c as in [23] are adopted for optimization.

For our framwork, we also retain 90% energy when computing the orthonormal matrix of affine hull, then limit the number of candidate sets to 5 for each experiment, the reason of which is shown in Section. 3.4. In all the experiments, we evaluate ten-fold cross validation and then report averages and standard deviations of the resulting classification rate.

3.3. Accuracy and robustness analysis

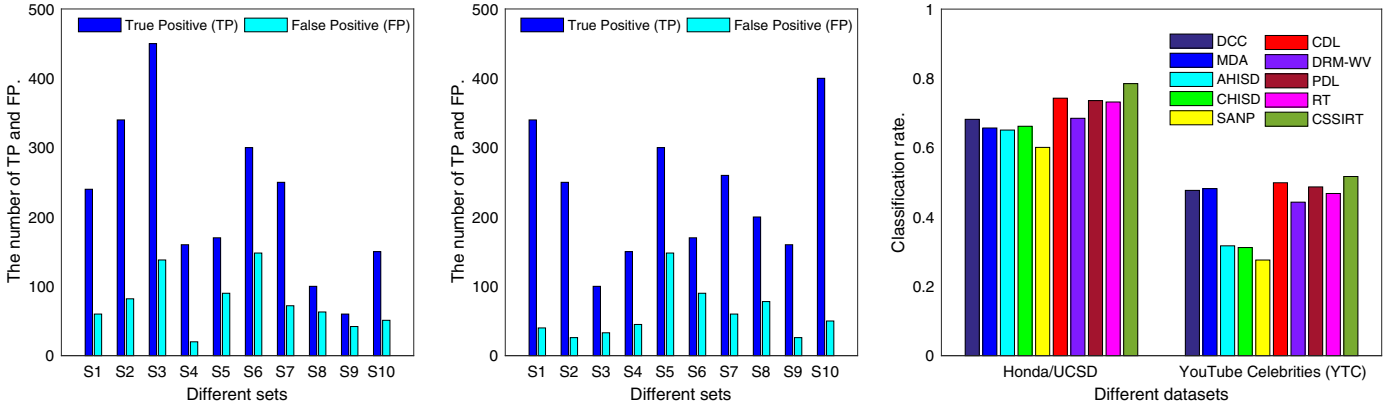
In order to demonstrate the accuracy and robustness of our method, we evaluate experimental performance under three kinds of scenarios where different kinds of noise are added artificially, including (i) use the original datasets without any modification directly, (ii) image sets contaminated by face detection errors, such as random noise, blurred images and corrupted images, and (iii) image sets contaminated by images from other classes, such as simulated group images or multiple faces in a video frame.

(1) *Comparison results on the original datasets without any modification:* The average recognition rates with standard deviation of different state-of-the-art methods on Honda, MoBo, YTC, ETH-80 and COX datasets are summarized in Table 1. The results demonstrate that our method consistently achieves superior performance by producing the highest average classification accuracy on all of the evaluated datasets. Moreover, it has the lowest standard deviation on all datasets, which means it is more stable than other methods. In particular, the achieved performance on Honda, MoBo, ETH-80, YTC, COX are $100.0 \pm 0.0\%$, $98.5 \pm 0.2\%$, $96.3 \pm 1.6\%$, $79.5 \pm 4.2\%$, $78.1 \pm 8.7\%$, respectively. Amongst the recent methods, CDL, DRM, PDL, RT show a comparable performance.

(2) *Comparison results on sets contaminated by detection errors:* To validate the robustness of our proposed method to noisy data with detection errors, we follow the approach reported in [13] to construct the noisy data on the two popular datasets Honda and YTC. As done in [13], we use the fast multi-pose face detection system [43] to detect faces in frames of the two datasets, and thus the image sets contain both real faces and false alarms. These noise or outlier images of the two contaminated sets are quite commonly encountered in practical conditions. Moreover, it should be noted that Honda is acquired in pure environment under controlled conditions, but YTC dataset is a real-life dataset collected under unconstrained conditions. They are two very typical databases, this

Table 1
Average classification rates(%) with standard deviation of our method versus different state-of-the-art methods on five datasets.

Method/Year	Honda	MoBo	ETH-80	YTC	COX
DCC[15]/2007	89.3 ± 2.5	92.8 ± 1.2	86.0 ± 6.5	66.8 ± 3.5	43.3 ± 12.1
MDA[45]/2009	91.8 ± 1.6	81.0 ± 12.3	77.3 ± 5.5	67.0 ± 4.6	73.1 ± 10.4
AHISD[22]/2010	92.8 ± 2.2	85.8 ± 4.2	78.8 ± 5.3	63.7 ± 2.1	64.1 ± 11.3
CHISD[22]/2010	92.3 ± 1.8	96.4 ± 0.8	79.5 ± 5.3	66.5 ± 2.0	63.1 ± 10.4
SANP[26]/2012	94.9 ± 2.6	97.5 ± 1.8	77.8 ± 7.3	68.4 ± 2.8	66.2 ± 13.4
CDL[12]/2012	95.4 ± 2.2	90.0 ± 5.0	77.8 ± 4.2	69.7 ± 1.0	56.1 ± 16.3
DRM-WV[31]/2015	**100.0 ± 0.0**	97.6 ± 0.9	95.9 ± 2.2	75.8 ± 4.9	69.4 ± 13.2
PDL[23]/2017	94.7 ± 1.6	93.5 ± 1.4	94.6 ± 5.4	74.3 ± 2.5	**78.5 ± 10.0**
RT[35]/2017	99.6 ± 0.4	98.2 ± 0.7	96.1 ± 1.8	77.4 ± 3.5	74.7 ± 10.2
CSSIRT/2019	**100.0 ± 0.0**	**98.5 ± 0.2**	**96.3 ± 1.6**	**79.5 ± 4.2**	78.1 ± 8.7



(a) TP and FP of contaminated Honda (b) TP and FP of contaminated YTC (c) Classification rates of contaminated sets.

Fig. 4. The settings and results on the sets contaminated by detection errors. (a) and (b) are the numbers of True Positive (TP) and False Positive (FP) face images in 10 randomly picked sets from Honda and YTC respectively, which are contaminated by detection errors. (c) Classification rates of different methods on those contaminated sets.

is why we chose them. We randomly pick 10 sets from Honda and YTC respectively. The number of True Positive (TP) and False Positive (FP) face images of each set are shown in Fig. 4(a) and (b), which show that the average portion of FP are as high as about 25.6% and 20% of the total number of face detector output (TP + FP), respectively.

The classification results are shown in Fig. 4(c). More precisely, we also achieve higher classification rates of 78.6% and 51.7% on contaminated Honda and YTC, respectively. Compared with the deep learning method DRM, DRM has poor robustness. Compared with MDA, AHISD, CHISD, and SANP, they can't cope well with the problem of the gallery affine hulls may intersected with the probe one, so that the performances drop sharply. Moreover, our method, as a parameterless one, has also achieved comparable performance with CDL due to it takes into account the statistical structure of the image sets, so it achieves good robustness. Hence, our method is suitable for the challenging problem of real-life applications with detection noisy data.

(3) *Comparison results on sets contaminated by images from other classes:* In order to verify the robustness on the sets contaminated by images from other classes, we still perform experiments on unclean Honda and YTC datasets, which are corrupted by adding images from other classes, as done in [12], i.e., the specified number of samples of an image set are replaced by samples from other classes.

For the better description of the experiments, we give some notations here, "NG" is the experimental scenario where only gallery set is contaminated, "NP" denotes that only probe set is contaminated, and "NGP" denotes both are contaminated. The gallery and/or probe sets are contaminated by 1, 2 or 3 images from each of all the other classes.

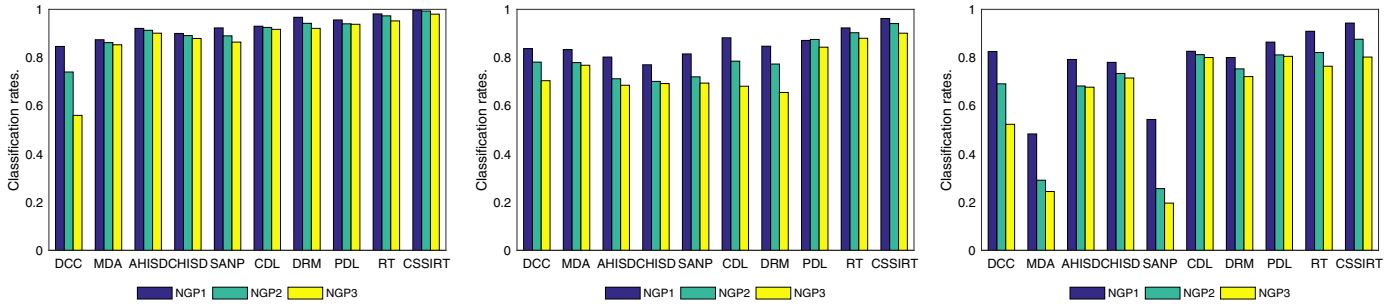
Figs. 5 and 6 show the recognition rates of our method versus different state-of-the-art methods on the scenario of data contaminated by images from other classes. It proves that our method is robust w.r.t. some numbers of images are contaminated by cross-contamination for real-life applications. Relevant analysises are the same as the second scenario.

3.4. Parameter analysis

For evaluating the performance of our method versus the number of candidate sets, we evaluate on five challengeable datasets, including Honda, MoBo, YTC, ETH-80 and COX datasets. From Fig. 7, we can ensure that our method almost always obtains a best and stable classification accuracy when the number of candidate sets larger than 6. This clearly shows that our CSS algorithm discard the gallery sets "far" from the query set is really reasonable.

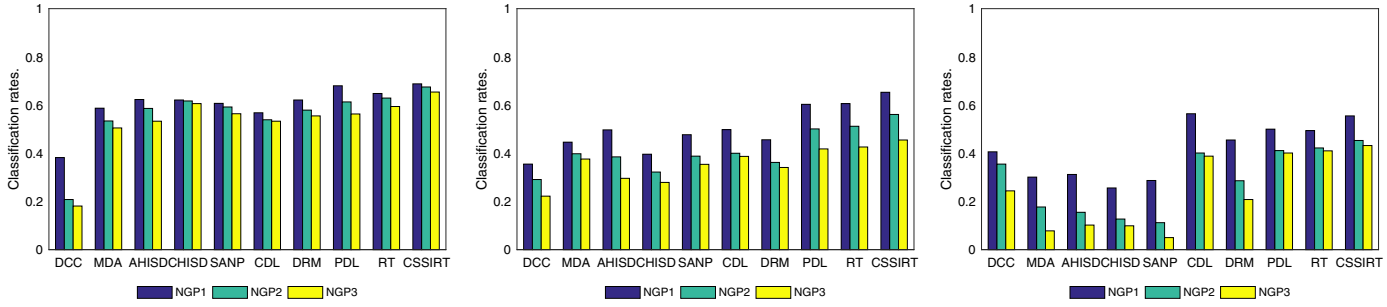
3.5. Timing analysis

Lastly, we compare the training and testing time of different methods on MacBook Pro 2017 with an Intel Core i5 (2.3GHz) CPU. Our implementation of CSSIRT is based on Matlab. Time in seconds required for offline training and online testing for total subjects (50 images per set) respectively on the YTC dataset are listed in Table 2, where 'N/A' means that the method does not perform any offline training. The results in Table 2 suggest that our method is almost much faster than other methods in both training and testing. Note that, the computational time of the deep learning method DRM far exceeds other methods up to 570.4*s* in training stage, but our method only requires 0.56*s* in total. Compared with RT, our method uses the candidate sets selection strategy. Compared with



(a) Gallery sets are contaminated. (b) Probe set is contaminated. (c) Both gallery and probe are contaminated.

Fig. 5. Classification rates on Honda database contaminated by images from other classes, where noisy datasets are called as "NG" (only gallery sets are contaminated), "NP" (only probe set is contaminated), and "NGP" (both gallery and probe sets are contaminated), respectively, and the number of contaminated images are suffixed by "1/2/3".



(a) Gallery sets are contaminated. (b) Probe set is contaminated. (c) Both gallery and probe are contaminated.

Fig. 6. Classification rates on YTC database contaminated by images from other classes. Here, all the notations are the same as that in Fig. 5.

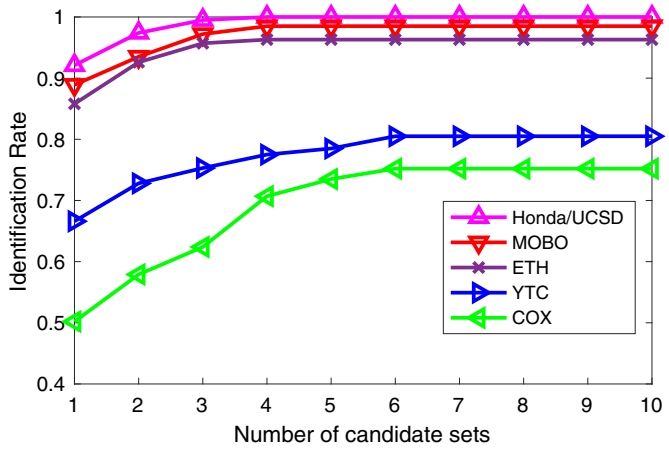


Fig. 7. Evaluate the influences of different number of candidate sets on five datasets. In consideration of the performance and time complexity, keeping 6 candidate sets is the best choice.

other iterative learning methods, our method avoids solving complex object function. Moreover, some optimization tips are given in Section 2.4. Therefore, our method is suitable for online real-life large-scale scenarios such as security, surveillance systems and multi-view camera networks.

4. Discussion

We have analyzed our method under three different experimental scenarios in Section 3.3. The experimental results clearly indicate that the proposed method achieves superior performance on robustness, classification accuracy and time complexity. The main reasons are as follows: (1) The Top-*N* promising candidate sets algorithm (CSS) is used for selecting few candidate sets instead of all the gallery sets. It avoids the interference from images of impossible sets to improve robustness. Moreover, it significantly reduces the computation complexity and memory storage requirement. (2) The exaction of subsets from candidate sets algorithm (EOS) adopts *K*-means clustering algorithm instead of randomly select algorithm, which can better take into account the statistical structure of each image set to enhance robustness. (3) The improved reverse training algorithm (IRT) uses two robust binary classifiers, and fused their corresponding normalized frequency histograms, which also helps increase accuracy and robustness of our method.

Table 2
Time cost of different methods on YTC (50 images per set) for training and testing.

Method/Year	Train	Test
DCC[15]/2007	13.7	0.2
MDA[45]/2009	4.3	0.10
AHISD[22]/2010	N/A	1.64
CHISD[22]/2010	N/A	2.31
SANP[26]/2012	N/A	22.4
CDL[12]/2012	230.6	3.50
DRM-WV[31]/2015	570.4	0.38
PDL[23]/2017	82	1.13
RT[35]/2017	N/A	6.50
CSSIRT/2019	N/A	**0.56**

5. Conclusions

In this paper, we have proposed a very promising two-stage cascaded framework CSSIRT for image set classification. In CSSIRT,

the first stage selects the most promising top-N candidate sets by CSS. The second stage utilizes the returned candidate sets and input query set for training and testing by using a K-means clustering and histogram fusion based improved reverse training algorithm. CSSIRT has evaluated on five benchmark image set datasets under three kinds of situations (i.e., origin clean data, data contaminated by face detection errors, and data contaminated by images from other classes). The experimental results have corroborated that the superiority of our method on robustness, accuracy, time complexity. Therefore, we strongly believe that our method will be more notable for real-life large-scale image set classification tasks.

However, how to deal with the linearly inseparable points in the second stage is still one issue worthy of further study.

Acknowledgment

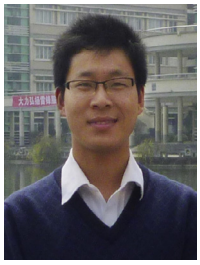
This work was supported by National Natural Science Foundation of China (Grant no. 61673220) and State Administration of Science Project of China (Grant nos. JCKY2017209B010, JCKY2018209B001).

References

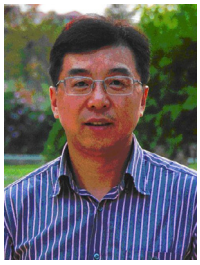
[1] Z. Chen, B. Jiang, J. Tang, B. Luo, Image set representation and classification with attributed covariate-relation graph model and graph sparse representation classification, Neurocomputing 226 (2017) 262–268.
[2] P. Zhu, W. Zuo, L. Zhang, S.C.-K. Shiu, D. Zhang, Image set-based collaborative representation for face recognition, IEEE Trans. Inf. Forensics Secur. 9 (7) (2014) 1120–1132.
[3] A. Mahmood, A. Mian, R. Owens, Semi-supervised spectral clustering for image set classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 121–128.
[4] S. Elaiwat, M. Bennamoun, F. Boussaid, A semantic RBM-based model for image set classification, Neurocomputing 205 (2016) 507–518.
[5] M. Ma, M. Shao, X. Zhao, Y. Fu, Prototype based feature learning for face image set classification, in: Proceedings of the Tenth IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–6.
[6] M. Hayat, M. Bennamoun, A.A. El-Sallam, An RGB-D based image set classification for robust face recognition from Kinect data, Neurocomputing 171 (2016) 889–900.
[7] S. Chen, A. Wiliem, C. Sanderson, B.C. Lovell, Matching image sets via adaptive multi convex hull, in: Proceedings of the 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2014, pp. 1074–1081.
[8] M. Shao, D. Tang, Y. Liu, T.-K. Kim, A comparative study of video-based object recognition from an egocentric viewpoint, Neurocomputing 171 (2016) 982–990.
[9] Y. Guo, R. He, W.-S. Zheng, X. Kong, Z. He, Robust spectral regression for face recognition, Neurocomputing 118 (2013) 33–40.
[10] J. Lu, V.E. Liong, X. Zhou, J. Zhou, Learning compact binary face descriptor for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2041–2056.
[11] Z.-Q. Zhao, S.-T. Xu, D. Liu, W.-D. Tian, Z.-D. Jiang, A review of image set classification, Neurocomputing (2018), doi:10.1016/j.neucom.2018.09.090.
[12] 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 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2496–2503.
[13] W. Wang, R. Wang, S. Shan, X. Chen, Probabilistic nearest neighbor search for robust classification of face image sets, in: Proceedings of the Eleventh IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 1, IEEE, 2015, pp. 1–7.
[14] R. Vemulapalli, J.K. Pillai, R. Chellappa, Kernel learning for extrinsic classification of manifold features, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 1782–1789.
[15] 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.
[16] 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 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 2705–2712.
[17] M. Hayat, M. Bennamoun, S. An, Learning non-linear reconstruction models for image set classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1907–1914.
[18] Z. Cui, S. Shan, H. Zhang, S. Lao, X. Chen, Image sets alignment for video-based face recognition, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2626–2633.
[19] S. Chen, C. Sanderson, M.T. Harandi, B.C. Lovell, Improved image set classification via joint sparse approximated nearest subspaces, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 452–459.
[20] R. Wang, S. Shan, X. Chen, Q. Dai, W. Gao, Manifold–manifold distance and its application to face recognition with image sets, IEEE Trans. Image Process. 21 (10) (2012) 4466–4479.
[21] R. Wang, S. Shan, X. Chen, W. Gao, Manifold-manifold distance with application to face recognition based on image set, in: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2008, pp. 1–8.
[22] H. Cevikalp, B. Triggs, Face recognition based on image sets, in: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 2567–2573.
[23] W. Wang, R. Wang, S. Shan, X. Chen, Prototype discriminative learning for image set classification, IEEE Signal Process. Lett. 24 (9) (2017) 1318–1322.
[24] M. Yang, P. Zhu, L. Van Gool, L. Zhang, Face recognition based on regularized nearest points between image sets, in: Proceedings of the Tenth IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–7.
[25] M. Leng, P. Moutafis, I.A. Kakadiaris, Joint prototype and metric learning for image set classification: Application to video face identification, Image Vis. Comput. 58 (2017) 204–213.
[26] 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.
[27] Z. Cui, H. Chang, S. Shan, B. Ma, X. Chen, Joint sparse representation for video-based face recognition, Neurocomputing 135 (2014) 306–312.
[28] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[29] L. Xuan, Z. Wang, W. Zhao, Y. Liu, Image set classification based on low-rank representation, J. Tongji Univ. (Natural Science) 2 (2013) 022.
[30] S.A.A. Shah, M. Bennamoun, F. Boussaid, Iterative deep learning for image set based face and object recognition, Neurocomputing 174 (2016) 866–874.
[31] M. Hayat, M. Bennamoun, S. An, Deep reconstruction models for image set classification, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2015) 713–727.
[32] H. Cevikalp, H. Serhan Yavuz, Fast and accurate face recognition with image sets, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1564–1572.
[33] Y. Xu, X. Zhu, Z. Li, G. Liu, Y. Lu, H. Liu, Using the original and 'symmetrical face' training samples to perform representation based two-step face recognition, Pattern Recognit. 46 (4) (2013) 1151–1158.
[34] D. Das, C.G. Lee, Graph matching and pseudo-label guided deep unsupervised domain adaptation, in: Proceedings of the International Conference on Artificial Neural Networks, Springer, 2018, pp. 342–352.
[35] M. Hayat, S.H. Khan, M. Bennamoun, Empowering simple binary classifiers for image set based face recognition, Int. J. Comput. Vis. 123 (3) (2017) 479–498.
[36] L. Zhang, Q. Liang, Y. Shen, M. Yang, F. Liu, Image set classification based on synthetic examples and reverse training, Neurocomputing 228 (2017) 3–10.
[37] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: a library for large linear classification, J. Mach. Learn. Res. 9 (Aug) (2008) 1871–1874.
[38] K.-C. Lee, J. Ho, M.-H. Yang, D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, in: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, IEEE, 2003. I–I.
[39] R. Gross, J. Shi, The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Carnegie Mellon University, 2001.
[40] M. Kim, S. Kumar, V. Pavlovic, H. Rowley, Face tracking and recognition with visual constraints in real-world videos, in: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2008, pp. 1–8.
[41] B. Leibe, B. Schiele, Analyzing appearance and contour based methods for object categorization, in: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2, IEEE, 2003, pp. II–409.
[42] Z. Huang, S. Shan, H. Zhang, S. Lao, A. Kuerban, X. Chen, Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on cox-s2v dataset, in: Proceedings of the Asian Conference on Computer Vision, Springer, 2012, pp. 589–600.
[43] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[44] D.A. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, Int. J. Comput. Vis. 77 (1–3) (2008) 125–141.
[45] R. Wang, X. Chen, Manifold discriminant analysis, in: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 429–436.
[46] J. Hamm, D.D. Lee, Grassmann discriminant analysis: a unifying view on subspace-based learning, in: Proceedings of the Twenty-fifth International Conference on Machine Learning, ACM, 2008, pp. 376–383.

Zhenwen Ren received the M.S. degree in Communication and Information System at Southwest University of science and Technology (SWUST), Mianyang, China. He is currently with the Department of School of National Defence Science and Technology of SWUST, and pursuing the Ph.D. degree in control science and engineering at Nanjing University of Science and Technology (NUST), Nanjing, China. His research interests include image set classification, sparse representation, low-rank representation, deep learning and their applications in machine learning.

Bin Wu received the Ph.D. degree in Control Theory and Control Engineering from Beijing University of Science and Technology (USTB), China, in 1999. He is a professor in the Department of Information Engineering at Southwest University of Science and Technology (SWUST). His current interests include pattern recognition, control theory.



Xiaoqian Zhang is currently with the School of Information Engineering, Southwest University of Science and Technology. His research interest covers subspace clustering, sparse representation, low-rank representation and their applications in image processing.



Quansen Sun received the Ph.D. degree in Pattern Recognition and Intelligence System from Nanjing University of Science and Technology (NUST), China, in 2006. He is a professor in the Department of Computer Science at NUST. He visited the Department of Computer Science and Engineering, The Chinese University of Hong Kong in 2004 and 2005, respectively. His current interests include pattern recognition, image processing, remote sensing information system, medicine image analysis.