This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology

Patch Set Based Representation for Alignment-Free Image Set Classification

Shenghua Gao, Zinan Zeng, Kui Jia, Tsung-Han Chan, Jinhui Tang

***Abstract*—This paper presents a patch set based sparse representation for image set classification. Compared with image-based image set representation, our patch set based representation is alignment-free and thus has an advantage for the tasks like video-based face recognition, image set based object recognition, and video-based hand gesture recognition, where precious alignment is usually difficult or even impossible due to large variance in view angle or pose. Specifically, to bypass the alignment issue, we propose to adopt the patch based image set representation by dividing each image within each set into patches, then we cluster all the training patches into multiple clusters and classify the test patches based on the cluster centers of training patches. The labels of test patches within each cluster are inferred from a Patch Set based Sparse Representation for Classification (PS-SRC), and the labels of all test patches from all the clusters are then aggregated to predict a single label for the test set. Experimental results on video-based face recognition datasets (CMU-MoBo and Youtube Celebrities), image set based object recognition dataset (ETH-80) and video-based hand gesture recognition dataset (Kinect Hand Gesture) demonstrate that our proposed method consistently outperforms all existing ones, and the improvement is very significant on the Youtube Celebrities and Kinect Hand Gesture datasets. Moreover, we also quantitatively show the robustness of our method to misalignment on the Mutli-PIE dataset.**

***Index Terms*—image set classification; patch set based representation; alignment-free; video-based face recognition.**

I. Introduction

Image set classification has widespread real-world applications, such as video-based face recognition, video-based hand gesture recognition, and real-time object recognition for robots. In image set classification, test data are a collection of images belonging to the same class. Generally speaking, an image set provides more information about the properties of its associated class than any individual image in the set does. Therefore, set based classification usually achieves higher recognition accuracy than single image based recognition does. Because of the importance and realistic setting of image set classification, this research topic has drawn increasing attention in the vision community [1][2][3][4][5][6].

Shenghua Gao is with ShanghaiTech University, Shanghai, China. Zinan Zeng,and Tsung-Han Chan are with Advanced Digital Sciences Center, Singapore. Kui Jia is with the University of Macau, Macau, China. Jinhui Tang is with Nanjing University of Science and Technology, China.
E-mail: gaoshh@shanghaitech.edu.cn

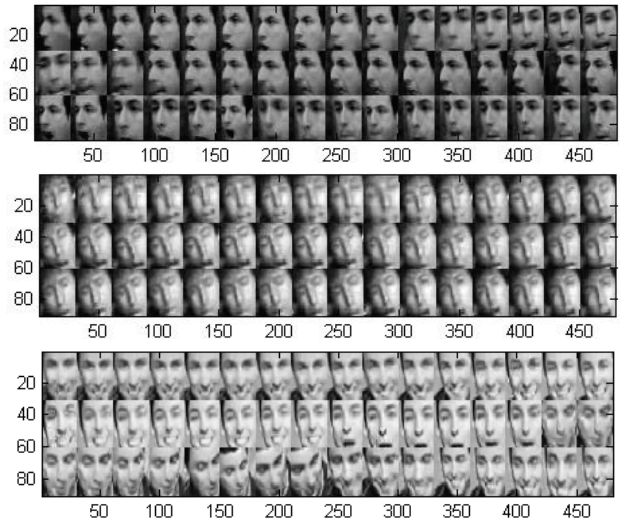
This work was supported by the Shanghai Pujiang Program under Grant 15PJ1405700.

The task of image set classification is to predict the label of the object contained in a given collection of images, where object instances usually appear with intra-class variations. For video-based recognition, such variations may include illumination, expression, occlusion, pose change, *etc.* Traditional image set classification methods usually represent each image in an image set by a single feature vector, and classify the image set based on such image-based representation [7][8][9]. However, such an image-based representation usually requires that the object instances in images are reasonably aligned. Considering that alignment of object instances with various intra-class variations is practically difficult, this requirement of alignment greatly limits the applicability of these methods. Even though there exists effort of face alignment for the video-based face recognition [10], the alignment usually does not work well for faces captured from different viewpoints. In addition, auxiliary face images are needed to help with the alignment, which restricts the stability and generalization of such methods. Fig. 1 illustrates the difficulty of aligning object images taken from different viewpoints.

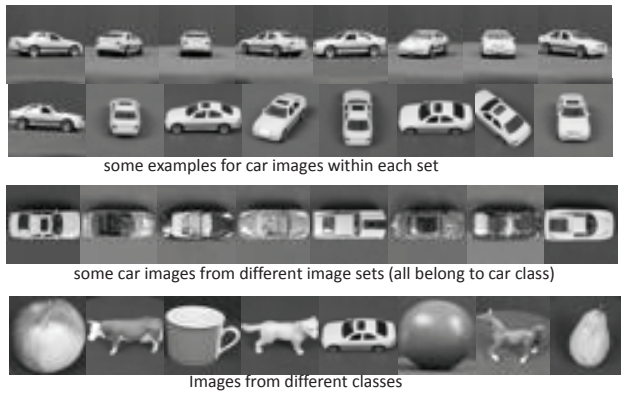
Patch-based representation has shown good performance for face recognition [11] and face verification [12] due to its robustness against misalignment. Motivated by these works, to overcome the alignment issue for image set classification, we propose a patch set based representation which represents the whole image set as a collection of patches generated from all the images within this set. We further group similar patches together so as to model the same part of the object/face/hand in the image set. Then we use a patch set based sparse representation to construct a virtual test patch and classify its label with respect to those of training data. Finally the class labels of all virtual test patches are aggregated to predict a single label of the test image set. Existing works have shown the effectiveness of sparse representation for face recognition [13] and video based face recognition [14], especially in the cases where there are unconstrained image corruptions and noises. In our cases, noises and corruptions usually exist, so we adopt the sparse representation framework for image set classification. Experiments in section V will show the effectiveness of our model, which validates the correctness of our choice.

Contributions. Our work contributes to image set classification in the following aspects: (i) A patch set based representation is proposed for alignment-free image set classification, and we have quantitatively showed the robustness of our method to the misalignment issue; (ii) A nonnegative constraint regularized Patch Set based Sparse Representation for Classification (PS-SRC) is proposed for classification of

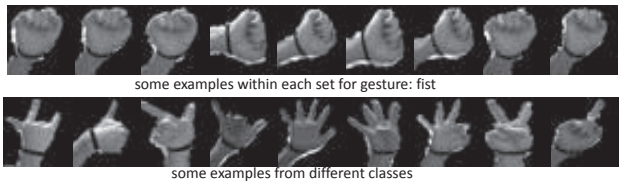
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology



(a) Some exemplary images from Youtube Celebrities. These face images are from the same person.



(b) Some exemplary images from ETH-80. The variance in view-angle makes the alignment almost impossible for images within the same same on this dataset.



(c) Some exemplary images from Kinect Hand Gesture.

Fig. 1. Some exemplary images from different datasets.

each subset of test patches, and such formulation is robust to the outlier patches within each cluster; (iii) To reduce the computational burden, we propose to cluster the training patches first and partition the test patches into different clusters based on the cluster centers of training patches. We then conduct PS-SRC for training and test patches falling into the same clusters. Extensive experimental results have verified that such a strategy greatly improves the computational efficiency and represents local information well for recognition, thereby promoting the recognition accuracy; (iv) We evaluate the effect of different regularizers in our PS-SRC formulation.

The rest of the paper is organized as follows: The works related to set based image classification are discussed in Section II. We introduce our patch-based image set representation and the image set classification method based on such representation in Section III. Then we propose an accelerating strategy for our patch-based image set classification in Section IV. We experimentally evaluate our method in Section V, and conclude our work in Section VI.

II. Related Work

Set based image classification methods can be categorized into parametric model based methods and non-parametric model based methods.

In parametric model based methods, each image set is characterized by some parametric distribution, like Gaussian [15] or Gaussian Mixture Model [7], then Kullback-Leibler Divergence is used to evaluate the similarity between two distributions (image sets). But the performance of such methods greatly relies on parameters estimation which is very difficult, especially when there aren't many images in some image sets. Moreover, the performance of such methods may drop significantly if the correlation between training and test set is weak. Therefore non-parametric models are more often used.

Non-parametric model based methods can be further categorized several categories. The first category represents each image set as a linear subspace [16] or a mixture of some linear subspaces [17][18], and set-to-set distance is usually calculated based Canonical Correlation Analysis (CCA) [16] and its variants, like Kernelized CCA [19] or Localized CCA [20]. The second category represents each image set as a nonlinear manifold [21][1][8] and set-to-set distance is calculated based on the manifold to manifold distance, which can be either calculated directly or calculated based on the subspace-to-subspace distance. The third category characterizes the distance between two sets based on the covariance matrix [22][5][23]. Such covariance based methods demonstrate robustness to the data noises and varying sizes of image set. The fourth category represents each image set as a affine or convex hull [2][24][25][9], and set-to-set distance is calculated based on the nearest virtual points between two hulls. In other words, such affine/convex hull based methods represent the test set with a virtual point and linearly reconstruct such virtual point with the training images from each class. In this sense, it is quite related to Nearest Subspace [26] method in face recognition. It is worth noting that manifold based representation and subspace based representation is not suitable for image set representation if the number of the image is small but in contrast the variance is significant. In contrast, such affine/convex hull based methods are less sensitive to the number of images in the set. However, all the aforementioned works are based on the pairwise set-to-set distance to predict the label of the test set.

Previous works [13][27] have shown that using all the training samples to collaboratively represent the test image is important for the good performance of face recognition. Therefore, recently some researchers also use all the training images from all the classes to represent the virtual point(s) in affine/convex hull based image set representation. Specifically,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology

in [28], Zhu *et al.* propose to represent each test set with only one virtual point and use all the training images to linearly reconstruct such virtual point. As a result, some important information about the test image set is lost, for example, the object images taken from different view angles also help the recognition of this object. Consequently such lost information will harm the robustness of recognition accuracy. In [14], Ortiz *et al.* propose to sparsely represent each test image with all the training images while enforcing all the test images share the same reconstruction coefficients. Such constraint is too strong especially for the case where images within each set contains severe variances. Moreover, the performance of such method can also be affected if test set contains some outlier or the images are not well aligned. In [3], Chen *et al.* propose to partition the images into different groups, and represents all the test images in the same group with the group sparse constraint. But such method is still not able to deal with the image sets containing outlier images, which means the test image pattern may neither appear in the training set nor can be inferred by the training set. Furthermore, all the previous works are based on image-based image set representation, i.e., each image in the image set is represented as one feature vector of the image set. Such image set representation requires the images are well aligned, which is usually very impractical, for example, for faces with different poses, and very different or even impossible for object images taken different view angles.

III. Patch Set Based Representation for Image Set Classification

A. Patch Set Based Sparse Representation

Patch-based image representation has been shown to be robust to misalignment to some extent [11][12]. In patch-based face recognition [11], each patch of the test image is separately represented by all the training patches [27], then the label is predicted for each patch and all the labels are aggregated to predict the label for the whole image. But for image set classification, such a strategy is problematic due to the following reasons: (i) Such a representation is very computationally expensive. Each image set contains many images, and each image contains many patches. As a result, each image set contains a large amount of patches. It would be computationally prohibited to compute a sparse representation for every patch. (ii) There are usually many similar patches among different classes, which could reduce the prediction accuracy of the image set classification. For example, if the background covers a large proportion of the image and they are similar among different classes, patches from the background would mislead the label prediction of the whole image set.

To overcome the previously stated two issues in conventional patch-based methods, we first partition the patches of the test set into different clusters. In this paper, for simplification, k-means clustering is used. After partitioning all the test patches into different clusters, patches within the same cluster are similar, and they are likely to represent the similar part(s) of the objects belonging to the same class. Motivated by the Sparse Representation based Classification (SRC) for image classification, we propose a Patch Set based SRC (PS-SRC) formulation to represent the patches within the same cluster by using all the training patches. Namely, for all the test patches within the same cluster, we first generate a virtual test patch by linearly combing all the test patches within the same cluster (the reason of generating virtual test patch will be given in the Remarks after equation 1). With SRC, we represent the virtual test patch with all the training patches sparsely.

Mathematically, we denote the test patches belonging to the i-th cluster as X^i $(i = 1, 2, \ldots, K)$ and denote all the patches from all the training sets as D. Each column in X^i and D corresponds to the feature of a patch. With test patches in the i-th cluster, we can use their linear superposition $y^i = X^i u^i$ to generate a virtual test patch.[1] A basic idea to identify the class of these test patches is to see which set of training patches could linearly represent this virtual test patch. Mathematically we aim to find reconstruction coefficients (u^i, v^i) that minimize the discrepancy between the test and training patch set:

$$\begin{aligned} \min_{u^i, v^i} \quad & \frac{1}{2}\|X^i u^i - D v^i\|^2 + \lambda \|u^i\|_1 + \gamma \|v^i\|_{\ell_q} \\ \text{s.t.} \quad & \mathbf{1}^T v^i = 1, u^i_j \geq 0, \end{aligned} \tag{1}$$

where u^i_j is the j-th entry of sparse coefficients vector u^i.

Remarks: (i) The ℓ_1 norm ($\|u^i\|_1$) reduces the role of outlier test patches in classification by promoting sparsity of coefficients of the test patches within the same cluster. Usually some outlier test patches can be introduced into a given cluster because of the following two reasons: Firstly, there are some background or non-class-specific patches, and they would be partitioned into some clusters and become the outlier patches. Secondly, the simple k-means clustering method could mis-partition certain patches to incorrect clusters. No matter how the outlier patches are generated, they may mislead the classification of test image sets. The sparse regularization on u^i helps discounting outlier test patches by preferring to set their coefficients to be 0. (ii) We enforce the linear superposition coefficients for generating the virtual test patch to be nonnegative. All the test patches within the same cluster are similar, therefore it is no need and also impossible to cancel the noises with the similar test patches within the same cluster. As demonstrated in Fig. 8, such nonnegativity usually improves the classification accuracy, which validates our conjecture. (iii) $\mathbf{1}^T v^i = 1$ avoids the trivial solution (u^i, y^i and v^i are zero vectors). (iv) We impose the ℓ_q $(q = 1, 2)$ norm on v^i. If $q = 1$, sparsity constraint is also imposed on the reconstruction coefficients of the virtual test patch, and such ℓ_1 norm is especially desirable for the case that there are many outlier patches in the training set, like background patches. When the training set are relatively clean, similar to that of Collaborative sparse Representation based Classification (CRC) [27], ℓ_2 can be used, which will accelerate the computational efficiency. (v) Our formulation can also be understood as that we seek two nearest points between two sets, and these two points are the linear combination of instances within each set with some constraints. (vi)Our work extends the SRC [13] in terms of

[1] In this paper, we use the superscript to index the cluster number, and use the subscript to index certain element in a vector or a submatrix.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology

the formulation and application. Compared with Patch-based CRC [11], our formulation is more robust the misalignment and noises within each cluster.

After solving u^i and v^i, the virtual patch of the i-th cluster is computed as $y^i = X^i u^i$. We denote the sub-dictionary corresponding to the m-th class as D_m, and subvector of v^i corresponding to D_m as v_m^i. Following the classification criteria of SRC [13], the label of y^i is predicted based on the minimum reconstruction error criteria:

$$\text{label}(y^i) = \arg\min_m \frac{1}{2}\|y^i - D_m v_m^i\|_2^2. \quad (2)$$

Finally the label of the test set is predicted by majority voting of classification results of all the virtual test patches. If the ties case is encountered in majority voting, we just assign the test set to the class with the smallest index of the class label. The same strategy is applied to the baseline methods with majority voting.

B. Optimization of PS-SRC

It can be easily proven that objective function in equation (1) is convex w.r.t. the reconstruction coefficients: (u^i, v^i). For simplification, following the commonly used methods [29], we update u^i and v^i alternatively.

1): Let $y^i = X^i u^i$ be the virtual test patch of the i-th cluster. When u^i is fixed and ℓ_1 norm is imposed on v^i, the objective w.r.t. v^i can be written as

$$\min_{v^i} \quad \frac{1}{2}\|y^i - Dv^i\|^2 + \gamma\|v^i\|_1 \quad \text{s.t.} \quad \mathbf{1}^T v^i = 1. \quad (3)$$

To optimize the objectives in equation (3), we first optimize the objective function without the constraint, then we project the solution onto the feasible region. The objective without the constraint is the standard sparse coding (LASSO) formulation and it has been well studied. Lots of solvers have been developed to optimize it. In this paper, following the work [24] we adopt the Feature-Sign-Search algorithm [29] because of its good performance and efficiency. It is worth noting that Feature-Sign-Search algorithm is quite related to the Homotopy/LARS algorithm [30] and work in [31] has shown that Homotopy method demonstrates good performance for SRC based face recognition, but Feature-Sign-Search algorithm is more efficient than LARS in terms of computational costs [29].

2): If ℓ_2 norm is imposed on v^i, the objective w.r.t. v^i can be written as

$$\min_{v^i} \quad \frac{1}{2}\|y^i - Dv^i\|^2 + \gamma\|v^i\|_2^2 \quad \text{s.t.} \quad \mathbf{1}^T v^i = 1. \quad (4)$$

With the Lagrange multiplier, we can easily get the closed form solution corresponding to the objective function in equation (4).

3): When v^i is fixed, the objective w.r.t. u^i can be written as

$$\min_{u^i} \quad \frac{1}{2}\|X^i u^i - Dv^i\|^2 + \lambda\|u^i\|_1 \quad \text{s.t.} \quad u_j^i \geq 0. \quad (5)$$

We use the same optimization method as that of equation (3) to optimize u^i.

4): Convergence of the algorithm. We alternative optimize each variable, and do the projection in each subroutine. Such strategy is similar to the work "Projected gradient methods for linearly constraint problems [32]" where the gradient of the objective is first calculated without considering the constraint. Then gradient decent is used to update the solution, and at the end of each step, the solution is projected to the feasible region. Such idea is also frequently used in the problems with the nonnegative constraint (eg. nonnegative sparse coding [33]), or the solution is on the probabilistic simplex [34]. Tough the solutions these methods achieved are not the exact solution of the problems, they are actually the solutions that satisfy the constraints, and experiments have demonstrated the good performance of these solutions. Fig. 2 also shows that the objective function usually converges after a few iterations with this optimization method.

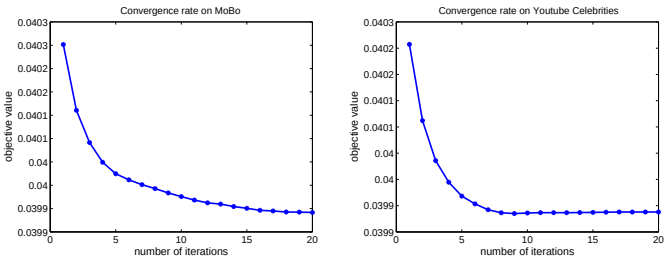


Fig. 2. The change of the objective value with respect to the number of iterations on the MoBo and Youtube Celebrities datasets.

IV. Cluster Training Patches for Efficient Image Set Classification

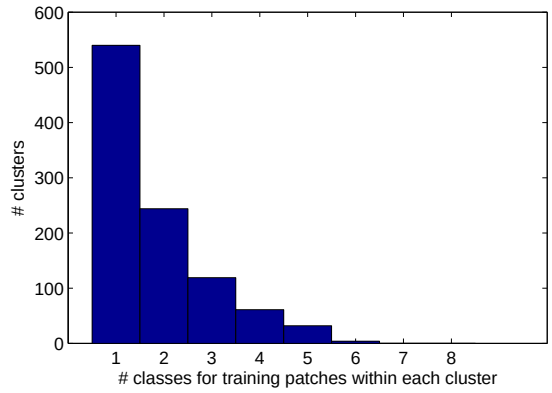
In previous section, we propose a patch set based representation for image set classification based on the PS-SRC formulation, and all the patches of all the training sets are used as the dictionary to sparsely reconstruct the virtual test patch. Consequently, the resultant dictionary D would probably be extremely large. Taking the image set classification on the MoBo dataset [35] which contains faces of 24 persons as an example, if 50 images are used as training faces for each person, and each image contains about 50 patches, then the number of the atoms in the whole dictionary would approximately be 60,000. The optimization of the sparse reconstruction coefficients w.r.t. such a large dictionary would be extremely time consuming [31]. However, in practical applications of larger datasets, the dictionary could be even larger. As a result, the image set classification directly based on equation (1) is usually infeasible.

To address this issue, we only select a relevant subset of training patches to reconstruct the virtual patches. A natural idea is to cluster the training patches and the test patches into smaller subsets and apply the PS-SRC formulation within each subset. In the following sections, we propose two strategies and discuss their feasibilities, respectively.

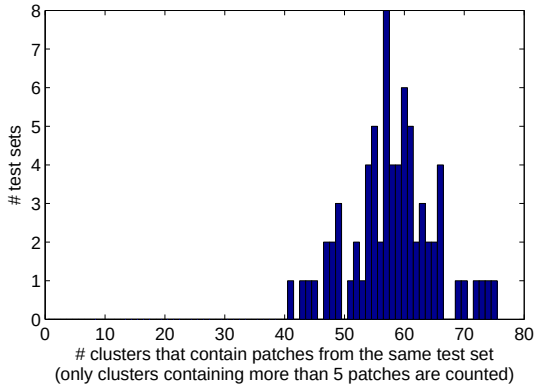
A. Cluster Test Patches First and Partition Training Patches Based on Their Distances to Cluster Centers of Test Patches? NO!

The first strategy is that we cluster the test patches first, and all the training patches are partitioned into different clusters

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology



(a) The distribution of the training patches.



(b) The distribution of all the test sets.

Fig. 3. Some statistics about the distribution of the patches on the ETH-80 dataset ($K = 1000$). (a) shows that patches within most clusters are only from a few classes, and more than a half of clusters only contain patches from one class. (b) shows that for most test sets, patches of the same set mostly falls into a very few clusters, which is typically less than 90 (in our experiments, we do not count or consider clusters which contain less than 6 test patches from the same set).

based on their distances to the cluster centers of test patches. Then we apply PS-SRC to training and test patches falling into the same cluster and predict the label of virtual test patch within that cluster. However, such strategy is not so scalable to large datasets. For example, given a test set, since all these images belong to the same class, it is not reasonable to set the number of the clusters to be so large. Consequently the number of the training patches falling into each test cluster is still very large for large dataset. Besides, we need to conduct a new clustering for each test set, computational cost at testing time is very significant.

B. Cluster Training Patches and Partition Test Patches Based on Their Distances to Cluster Centers of Training Patches? YES!

The second strategy is that we cluster the training patches first, then we partition all the test patches based on their distances to the cluster centers of training patches. Similar to previous strategy, the training and test patches falling into the same cluster are also used to conduct PS-SRC, learn the virtual test patch, and predict its label. Such strategy has several advantages. Firstly, it reduces the computational cost in PS-SRC. On the one hand, training patches within each cluster are similar, therefore it is reasonable that most clusters only contain patches from a few classes.[2] For example, more than a half of clusters only contain patches from one class on the ETH-80 dataset (Fig. 3). Therefore the dictionary size is not very large in PS-SRC. On the other hand, test patches from the same set probably only fall into a small fraction of all the clusters which contain training patches that have the same class label as that of the test patches (as shown in Fig. 3). Therefore, the number of PS-SRC objectives needed to be solved and the size of the dictionary in each objective will not be so large. Secondly, this strategy is scalable to large scale datasets. Based on the number of training classes and the content complexity of patches, we can choose a number of the clusters (K) according to the scale of the dataset. For example, if there are many training classes and the patches are more diverse, we can increase K accordingly. Therefore the number of training patches falling into one cluster can always be kept to be small. Thirdly, we use the patches falling into the same clusters, which are probably similar, to conduct PS-SRC. Therefore the locality information of all patches are preserved. As demonstrated in [36] that such locality information helps to improve the classification accuracy.

C. Summary.

In real applications, rather than using all the training patches as the dictionary, we cluster the training patches with k-means first. Based on the cluster centers of training patches, we partition the test patches, and do the PS-SRC for training and test patches falling into the same cluster. Mathematically, for the i-th cluster which contains both training patches and test patches, the objective of PS-SRC is given as follows:

$$\begin{aligned} \min_{u^i, v^i} \quad & \frac{1}{2}\|X^i u^i - D_i v^i\|^2 + \lambda\|u^i\|_1 + \gamma\|v^i\|_{\ell_q} \\ \text{s.t.} \quad & \mathbf{1}^T v^i = 1, u^i_j \geq 0. \end{aligned} \tag{6}$$

Here X^i and D_i are the test and training patches falling into the i-th cluster, respectively. Then the label of virtual test patch is also predicted based on the minimum reconstruction error criteria, and the label of the test set is still based on the majority voting of all the virtual test patches. We illustrate the pipeline of our patch set based representation for image set classification in Fig. 4. We summarize the contribution of each components as follows: i) By representing each image with a collection of patches with different sizes, the misalignment can be overcome. ii) By clustering training patches and partitioning test patches based on their distances to cluster centers of training patches, we can reduce the computational costs and make patch-based sparse representation for image set classification feasible. iii) Sparse representation based classification has demonstrate good performance for image classification. By

[2]For similar patches among all classes, like background, they probably be clustered into only one or a few clusters, or clustered into other clusters as outliers. For the first two cases, the prediction of test set won't be affected because of the majority voting. For the last case, the sparsity regularizer can also eliminate the effect of the outliers. Therefore our algorithm also handles the case where similar background covers a large portion of image.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology

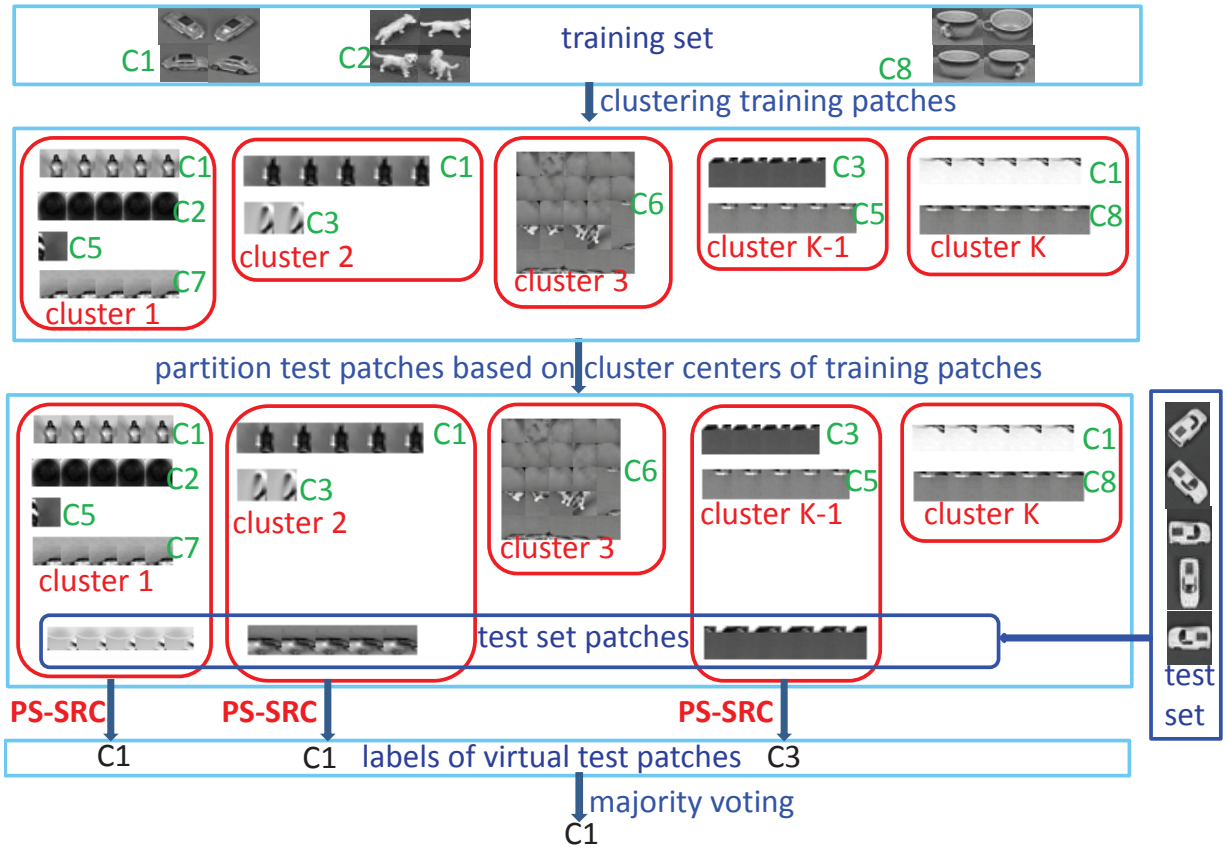


Fig. 4. The flowchart of patch set based representation for image set classification.

performing sparse representation based classification within each patch cluster, the performance image set classification is guaranteed. iv) The majority voting procedure gets rid of the effect of those outlier patches, and makes the image set classification more robust.

V. Experiments

In this section, we experimentally evaluate the proposed method on the CMU-MoBo dataset [35], the Youtube Celebrities dataset [37], the ETH-80 dataset [38], and the Kinect Hand Gesture Recognition dataset [39]. These 4 datasets cover 3 possible image set classification scenarios: video-based face recognition, object recognition for robot, and video-based hand gesture recognition.Moreover, we name our method as PS-SRC (ℓ_1-ℓ_1) if ℓ_1 norm is imposed on u^i and v^i, respectively, and we name our method as PS-SRC (ℓ_1-ℓ_2) if ℓ_1 norm is imposed on u^i and ℓ_2 norm is imposed on v^i.

A. Dataset Description

The CMU-MoBo (Motion of Body) Dataset [35] contains videos of 24 persons walking on the treadmill, and each person contains 4 different walking videos. Following the work [1][22][10], we use the Viola-Jones's cascade face detector [40] to automatically detect the face for each video frame, and all the faces are converted to gray-scale images and resized to 30×30 pixels. Then histogram equalization is used to overcome the illumination effect. Faces from the same video forms a image set. Besides the image patches which are detected by false positives and the alignment issues, faces within each set also contain variances in pose, illumination, and expression. Following the commonly used setting, for each person, one image set is randomly selected as training set and the remaining 3 videos as used test sets for each person. 10-fold validation experiments are conducted on this dataset. Same with [28], the performance of different methods are evaluated under three settings: we set the number of images from each set to be 50, 100, and 200, respectively (If the total image number is less than this number, all the images are used) for both training and test sets. Such setting is practical because in real applications, like video-based face recognition, it is possible that we only have a short video clique for training and test videos.

The Youtube Celebrities Dataset [37] is a very challenging video-based face tracking and recognition dataset in terms of dataset scale, the variances of the faces, and the low resolution and high compression of the videos in this dataset. Specifically, it contains 1910 video sequences of 47 persons, and all the videos are downloaded from the Youtube. Same with [1][22][10], we also use the Viola-Jones's cascade face detector [40] to detect the faces and resize faces to 30×30. All faces are also converted to gray-scale images. Histogram equalization is also used. Faces within each video sequence form an image set. Besides the false positive patches, the faces

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology

of the same person contains large variances in illumination, expression, pose, occlusion. We also use the commonly used setting on this dataset[1][24][10], i.e., 3 video sequences are used for training and 6 video sequences are used for test. 5-fold validation experiments are conducted on this dataset.

The ETH-80 Dataset [38] contains 8 classes, and each class contains 10 instances/objects of the same class. 41 images are taken under different view angle for each instance, and they form an image set. We use the 32×32 gray scale images for classification. For this dataset, we use two settings. i) setting S1: We use 5 objects as the training set for each class, and use the images for the other 5 objects as test sets. So the number of test sets is 8×5=40; ii) Setting S2: Same with [9], we sequentially choose one set as training set and use the rest 9 sets as test tests for each classes.

The Kinect Hand Gesture Dataset [41] contains both depth images and RGB images corresponding to 10 gestures taken under less controlled environments with Kinect. Each gesture contains the same gestures made by 10 different persons, which corresponds to the variances in the hands. Moreover, the same gestures made by the same person also contains 10 different images in terms of the direction and location w.r.t. the cameras of the Kinect. For each gesture, images done by the each person form an image set, therefore each set contains 10 images. Similar to [41], we use the depth information to detect and segment the hand first, then the detected hands are converted to gray-scale images and resized to 32×32 pixels. We randomly choose 2 image sets as the training set and use the remaining 8 image sets as test sets for each gesture.

We show some randomly sampled images from the Youtube Celebrities, ETH-80, and Kinect Hand Gesture Datasets in Fig. 1. For Youtube Celebrities, we show the images of the same person from different image sets. It can be easily seen that these images contain variance in pose, expression, and illumination. For example, on ETH-80 and Kinect Hand Gesture, there exist significant variance in view-angle for images within each set, which makes the alignment difficult. Consequently, the alignment issue affects the performance of the image-based image set representation for image set classification.

B. Experimental Setup

For simplification, we set the $\lambda = \gamma = 0.01$ on all the dataset for the PS-SRC (ℓ_1-ℓ_1). For the PS-SRC (ℓ_1-ℓ_2), we set $\lambda = 0.1$ on all datasets, and set γ=1e-5,1e-5,1e-3,1e-3, on ETH-80, Kinect Hand Gesture, MoBo, and Youtube Celebrities, respectively. As for the number of clusters (K in k-means), we set $K = 500$ on the Kinect Hand Gesture dataset, and set $K = 1000$ on the CMU-MoBo, Youtube Celebrities, and ETH-80 datasets [3]. For simplicity, we directly use pixel values in the patches as features, and all features are normalized by their ℓ_2 norm. We fix the patch size and the distance between two neighboring patches to be 8 and 4, respectively, on CMU-MoBo, Youtube Celebrities, and ETH-80. For the Kinect Hand Gesture dataset, we fix the patch size and the distance between two neighboring patches to be 20 and 1 because the patches covering a larger hand/finger region is more meaningful for the recognition of different gestures.

Baseline methods: We compare our method with four types of methods:[4]

- Manifold based methods, including Manifold-Manifold Distance (MMD) [1], and Manifold Discriminant Analysis (MDA) [8];
- Affine/Convex hull based methods, including Affine Hull based Image Set Distance (AHISD) [2], Convex Hull based Image Set Distance (CHISD) [2], and Sparse Approximated Nearest Points (SANP) [24];
- Collaborative representation based methods, including Collaborative Representation based Classification (CRC) [27], Sparse Representation based Classification (SRC) [13], Image Set based Collaborative Representation for Classification (ISCRC) with ℓ_1 norm constraint [28] and Mean Sequence Sparse Representation-based Classification (MSSRC) formulation [14]. For CRC and SRC, following the work of Zhu *et al.* [28], we use the average representation residual of query set for classification.
- Besides the existing method, we also propose a baseline method by regularizing with u^i and v^i with ℓ_2 norm respectively in equation (6), i.e., we use the following objective function to solve the (u^i, v^i), and denote the baseline method based on this objective function as PS-SRC (ℓ_2-ℓ_2).

$$\begin{aligned} \min_{u^i,v^i} \quad & \frac{1}{2}\|X^i u^i - D_i v^i\|^2 + \lambda\|u^i\|_2^2 + \gamma\|v^i\|_2^2 \\ \text{s.t.} \quad & \mathbf{1}^T v^i = 1, u^i_j \geq 0. \end{aligned} \tag{7}$$

- Patch-based CRC (PCRC)[11]. We divide the images in each set into patches by using the same way with our method, and use the PCRC formulation to predict each patch's label, then majority voting is used to predict the label of the test set.
- Patch based kNN (P-kNN). We divide the images in each set into patches by using the same way with our method, and use the kNN to predict each patch's label, then majority voting is used to predict the label of the test set. In our experiments, we set k=5.[5]

C. Performance Evaluation

1) Comparison with Baseline Methods: We list the performance of different methods on CMU-MoBo, Youtube Celebrities, ETH-80 and Kinect Hand Gesture in Table I, Table II, and Table III, respectively. We see that our method, despite its simplicity, achieves the best performance on all

[3]Though k-means converges to local minima, we repeat the experiments on the same test sets by 10 times, and the standard deviation is 0.70% on ETH-80 and 0.37% on Hand-Gesture, respectively. We can see that the variance is not significant.

[4]We don't compare our method with kernel based methods, but our method can be easily extended to the kernel formulation which can easily be done in our future work.

[5]We also tried k=3, 7, but the performance is worse than the cases where k=3.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I
PERFORMANCE COMPARISON BETWEEN DIFFERENT METHODS ON CMU-MOBO (%)

	50	100	200
CRC [27]	89.6±1.8	92.4±3.7	96.4±2.8
SRC [13]	91.0±2.1	91.8±2.7	96.5±2.5
MMD [1]	73.8±5.1	76.4±4.1	75.6±5.2
MDA [8]	79.7±3.8	83.9±2.5	86.0±2.3
AHISD [2]	93.2±1.9	93.9±2.5	94.3±2.5
CHISD [2]	91.7±2.4	92.5±3.3	93.1±2.8
SANP [24]	88.8±3.7	89.4±3.7	88.1±2.4
ISCRC [28]	96.1±2.6	96.2±2.6	96.9±2.1
MSSRC [14]	94.7±3.1	95.1±2.6	95.0±2.8
PCRC[11]	91.8± 3.2	90.7±2.9	NA
P-kNN	96.2±2.5	96.5±2.4	96.7±2.5
PS-SRC(ℓ_1-ℓ_1)	96.8±2.3	**97.4±2.1**	**98.1± 1.9**
PS-SRC(ℓ_1-ℓ_2)	**96.9±2.2**	97.0±2.0	93.3± 4.0
PS-SRC(ℓ_2-ℓ_2)	96.4±1.9	96.8±2.3	96.9± 2.5

TABLE II
PERFORMANCE COMPARISON BETWEEN DIFFERENT METHODS ON YOUTUBE CELEBRITIES (%)

	50	100	200
CRC [27]	56.7±7.4	59.5±6.6	61.4±6.4
SRC [13]	61.5±6.9	64.4±6.8	66.0±6.7
MMD [1]	57.8±6.6	62.8± 6.2	64.7±6.3
MDA [8]	58.5±6.2	63.3±6.1	65.4±6.6
AHISD [2]	57.5±7.9	59.7±7.2	57.0±5.5
CHISD [2]	58.0±8.2	62.8±8.1	64.8±7.1
SANP [24]	57.8±7.2	63.1±8.0	65.6±7.9
ISCRC [28]	62.3±6.2	65.6±6.7	66.7±6.4
MSSRC [14]	70.9±3.9	70.9±3.8	70.7±3.6
PS-SRC(ℓ_1-ℓ_1)	**73.5±3.3**	**74.3±3.3**	**74.5 ± 3.1**
PS-SRC(ℓ_1-ℓ_2)	60.5±3.1	61.2±1.7	60.9±3.0
PS-SRC(ℓ_2-ℓ_2)	60.5±2.4	60.5±3.3	61.3 ± 3.0

TABLE III
PERFORMANCE COMPARISON BETWEEN DIFFERENT METHODS ON ETH-80 AND KINECT HAND GESTURE (%)

	ETH-80 (S1)	ETH-80 (S2)	Hand Gesture
CRC [27]	83.8±8.6	61.9±6.2	70.5±4.2
SRC [13]	88.0±8.4	65.8±7.2	72.0±3.9
MMD [1]	82.3±5.6	72.1±7.1	77.4±5.7
AHISD [2]	78.0±5.2	57.8±5.7	87.0±4.8
CHISD [2]	81.3±6.7	58.6±5.3	84.8±3.2
SANP [24]	75.5±5.3	57.9±6.3	66.0±8.6
ISCRC [28]	83.0±3.7	61.3±4.4	79.0±4.6
MSSRC [14]	83.75±4.6	62.9±4.5	76.5±4.5
PCRC[11]	69.3±5.4	71.4±7.2	72.1±4.9
P-kNN	74.3±6.4	72.4±5.4	89.1±4.3
PS-SRC(ℓ_1-ℓ_1)	**89.1±5.8**	74.6±6.5	93.1±3.0
PS-SRC(ℓ_1-ℓ_2)	85.0±3.5	**76.3±7.1**	**93.9±3.5**
PS-SRC(ℓ_2-ℓ_2)	74.2±7.2	72.9±5.3	92.3±3.5

the datasets under all the settings. Specifically, most existing methods perform reasonably well on MoBo, which is relatively easy. Compared with MoBo, Youtube Celebrities contains false positives and severe misalignments caused by the large pose variance, expression, occlusion, it is more challenging. Therefore most methods will have difficulty on this dataset. Thanks to the robustness of our method (PS-SRC (ℓ_1-ℓ_1)) against outliers and misalignment, our method is superior over the previous methods by a large margin, including PCRC. Moverover, our method is also better than the patch-based kNN. The good performance of our method on ETH-80 and Hand Gesture also results from the robustness of our method to the alignment issue which is also serious on these two datasets because of severe view angle variances.[6] Moreover, the performance of our method can be further improved by carefully tuning the patch size, the distance between two neighboring patches, and the number of clusters (K) based on the data in the task we are dealing with. For instance, when patches of different size (4, 6, and 8 respectively) are used on ETH-80, the classification accuracy of our method moves up to 77.1% (please refer to section V-E).

It is also worth noting that the performance of our method (200 images per set) is comparable with or better than that of the image alignment based method [10] whose accuracy is 95.0% and 74.6% on MoBo and Youtube Celebrities, respectively, even though [10] uses more training images for each set and an auxiliary dataset is used for face alignment in [10]. Moreover, as shown in Table I, more images in each set usually help improve the accuracy. If all the images are used in each set on the MoBo dataset, the accuracy of PS-SRC (L1-L1) will reach 98.8±1.8%, which is even higher than the accuracy of 200 images per set (98.1%). Besides the classification accuracy of our method on different datasets, we also show the patch label prediction confusion matrix on ETH-80 (S1 setting) in Fig. 5. We can see that the percentage of the correctly predicted patches is much higher than that of misclassified patches. Then with the the majority voting, each image set would probably be correctly classified.

2) Different Regularizers: We notice that on both the MoBo and Youtube Celebrities dataset, the ℓ_1 norm regularized PS-SRC, i.e. PS-SRC (ℓ_1-ℓ_1), usually achieves the best performance. On the ETH-80 and Kinect Hand Gesture datasets, PS-SRC (ℓ_1-ℓ_2) achieves the best performance, while the performance of PS-SRC (ℓ_1-ℓ_1) is still not bad. On all the datasets, PS-SRC (ℓ_2-ℓ_2) usually achieves the worst performance. The reason may be that the ℓ_1 norm on the v^i helps remove the noisy test patches, which is important for the good performance of image set classification. On the MoBo, ETH-80, Kinect Hand Gesture datasets, most patches are relatively clean, except for the background patches which are prone to be clustered in one cluster, therefore the performance of ℓ_1 norm and that of ℓ_2 norm on u^i are similar, which agrees with the fact that CRC [27] achieves similar performance with SRC [13] for face recognition. On the Youtube Celebrities dataset, as the variance, like occlusion and pose variance, is significant, many patches are very noisy, and randomly distributed in different clusters. These outlier patches can easily affect the prediction if ℓ_2 norm regularizers are used, and reduce the recognition accuracy significantly. In conclusion,

[6]The performance of MDA on the ETH-80 and Kinect Hand Gesture datasets is lower than 50% because the limited training images with large variance within each set (The number of training images is 41 and 20 on ETH-80 and Hand Gesture, respectively.), which constrains the performance of MDA. Similar phenomenon is also observed in paper [9] on the ETH-80 dataset. Therefore we don't include its performance on these two dataset in this paper.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology

	car	apple	dog	cow	cup	horse	pear	tomato
car	42.54	4.97	9.10	5.40	7.30	13.12	5.61	11.96
apple	2.90	50.73	3.67	2.56	3.33	2.99	11.53	22.29
dog	4.96	12.95	31.57	7.83	4.56	14.15	10.87	13.11
cow	5.20	11.67	15.35	27.82	3.12	14.23	8.95	13.67
cup	8.15	14.39	5.44	5.36	40.61	8.55	7.91	9.59
horse	5.60	9.67	14.79	12.31	8.95	29.98	10.31	8.39
pear	3.20	8.95	5.44	2.72	5.12	4.96	66.11	3.52
tomato	1.44	17.59	4.24	1.52	0.88	4.96	13.19	56.20

Fig. 5. Patch label prediction confusion matrix on ETH-80 (S1 setting). In patch prediction confusion matrix, the element locates at (i,j) is the percentage of the patches from the i^{th} class be classified to the j^{th} class.

as suggested by the Wilcoxon sign-ranks test,[7] we can either choose L1-L1 formulation or L1-L2 formulation when we have no prior about data. Experiments on YouTube celebrities also suggests that once we know that the data is noisy, L1-L1 formulation is probably a better choice, which agrees with findings in previous works [43][14].

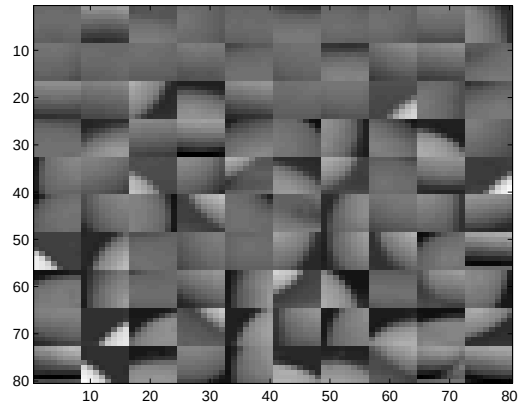
3) Visualization of Virtual Patches: We also show some reconstructed virtual test patches for test sets from car, dog, and apple in Fig 6. It can be seen that some virtual test patches are very class-discriminative, like the legs of dog, the shape of apple, *etc.*, therefore these virtual patches help the prediction of the test set.

D. The Effect of Alignment

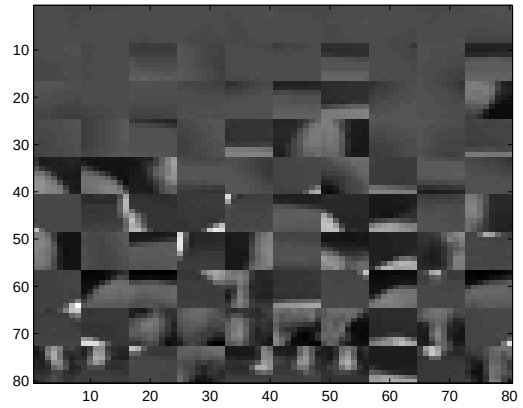
To further demonstrate the effectiveness of our method to misalignment, we test different methods on the Multi-PIE dataset [44] by using well-aligned images and images with misalignment. Specifically, Multi-PIE contains images of 337 persons taken under the four sessions over the span of 5 months. For each persons, images are taken under 15 different view angles and 19 different illuminations while displaying different facial expressions. In our experiments, the 249 persons in session 1 are used as the evaluation. Specifically, for each person, the frontal face images with neutral expression and different illuminations are used as the training set, and all the images with the same pose (15° or -15°) and different illuminations are used as the test sets. Therefore we have 2 test sets for each person (Please refer to Fig. 7 to see the training/test sets.). We use two sets of data for evaluation. The first one is aligned with manually annotated eye corners, and each set contains 20 images. The comparison data are obtained by resizing the faces detected with OpenCV face detector [8]. Then we resize all the faces to 30×30. We use the same parameters as that on Youtube Celebrities in our experiments. The performance of different methods on this datasets is listed in Table IV. As there are outliers, misalignment for the images in the same set, misalignment between images in test and training set (0° vs. 15°, 0° vs. -15°), and the number of images within each set is very limited (only around 10 images per set)

[7]We have conducted the Wilcoxon sign-ranks test [42] to compare the performance of l1-l1 and l1-l2 formulation. In Wilcoxon sign-ranks test, the sum of ranks for the positive differences is $R^+ = 37$, and the sum of ranks for the negative differences is $R^- = 8$. According to the table of exact critical values for the Wilcoxon's test, for a confidence level of α = 0.05 and N=9 datasets, the difference between the classifiers is significant if the smaller of the sum is equal or less than 5. We therefore don't reject the null-hypothesis.

[8]About 50% face images are not detected.



(a) Some virtual test patches for one test set which is from the class of apple.



(b) Some virtual test patches for one test set which is from the class of dog.

Fig. 6. Visualization of virtual test patches (i.e., $X^i u^i$) from ETH-80. We randomly sample some virtual patches $X^i u^i$ and reshape them 8×8 patches. We can see these virtual test patches corresponding to some typical patches in each class, which helps the image set classification. These patches demonstrate that the learnt virtual patches are meaningful.

which breaks the basis of manifold based methods, therefore many conventional methods perform extremely poor on this dataset if the data with misalignment are used. However, our method still achieves reasonable performance on the directly cropped data because patch-based representation overcomes the alignment issue to some extent and ℓ_1 norm also helps remove the outliers in image set representation.

E. Evaluation of Important Parameters

1) Patch Size: We illustrate the performance of our method with patches of different size in Fig. 8 on the ETH-80 dataset. Need to mention that in real applications, the patch size should be determined based on the specific task. For example, for hand gesture recognition, larger patches cover larger region which is meaningful for hand gesture recognition. Patches at different sizes covers object parts at different scales, therefore besides the single scale patches, we can also use patches

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology

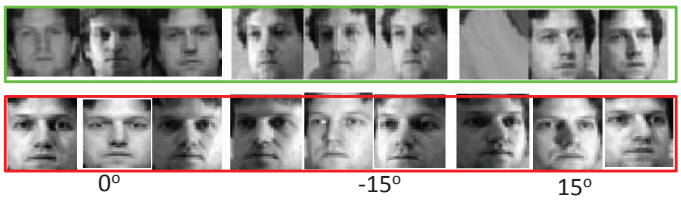


Fig. 7. Some images from the Multi-PIE dataset. Images in the first row are obtained by resizing the faces detected with OpenCV face detector, therefore there are some outliers and misalignment for these images. Images in the second row are aligned by using the manually annotated eye corners.

TABLE IV
PERFORMANCE COMPARISON BETWEEN DIFFERENT METHODS ON MULTI-PIE UNDER DIFFERENT SETTINGS (%)

	aligned images		unaligned images	
	-15°	15°	-15°	15°
SRC [13]	72.7	77.5	35.7	21.7
MMD [1]	76.7	78.3	9.6	8.0
AHISD [2]	77.5	79.2	17.7	11.7
CHISD [2]	77.1	79.9	10.4	6.4
SANP [24]	77.1	79.9	14.9	7.2
ISCRC [28]	72.7	79.1	28.9	16.5
MSSRC [14]	70.7	75.9	28.1	35.3
PS-SRC(ℓ_1-ℓ_1)	99.2	98.8	94.8	92.4

extracted from multiple scales for image classification based on our method, that it we generate virtual test patches of different size and do the majority voting with all these virtual test patches for image set label prediction. For example, on ETH-80, when we use patches of two different size (6 and 8), the accuracy reaches to (75.4±6.2)%. If we use patches of three different size (4, 6, 8), the accuracy reaches to (77.1±6.6)%. Furthermore, in this figure, it can be clearly seen that nonnegative constraint regularized PS-SRC achieves better performance than that without the nonnegative constraint, which validates the correctness of our formulation.

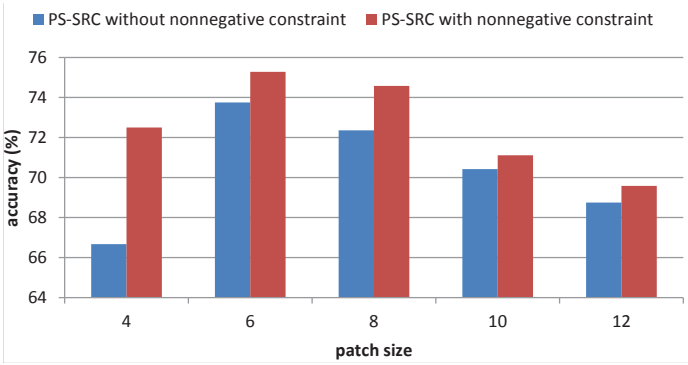


Fig. 8. Effect of patches of different size on the performance of PS-SRC (ℓ_1-ℓ_1) with or without non-negativity constraints.

2) The number of clusters (K in k-means): K is an important parameter in our method. We show the performance of different K on ETH-80 and MoBo in Fig. 9. On ETH-80, the structure of the object is diverse, though the number of class is small, the content of patches are complex, therefore larger K corresponds to better performance. Though structure of face is relatively easy, MoBo has many persons, whose faces are different, the performance corresponding to larger K (800-1200) is better. In summary, in real applications, we should determine K based on the contents complexity of the patches, which is probably proportional to the increase of classes. Moreover, this experiment also hints that our method is scalable to large scale dataset because we can use a larger K for a larger dataset.

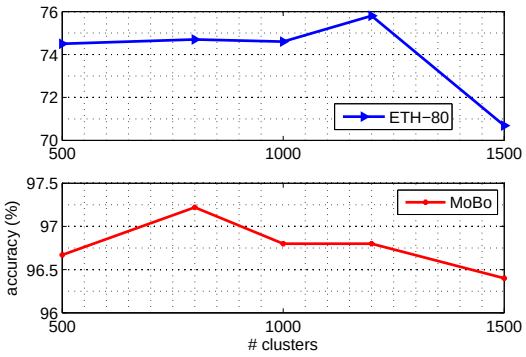


Fig. 9. Effect of different K in k-means on image set classification on ETH-80 and MoBo. The evaluation is based on the PS-SRC (ℓ_1-ℓ_1) formulation.

3) PS-SRC tuning parameters: λ and γ: λ and γ are two parameters in our PS-SRC formulation. In this paper, we set them to the same value for simplification. We show the classification accuracy and time cost with different λ (γ) in Fig. 11. We can see that the accuracy is relatively stable when they are between 0.0001 and 0.01. But it is more efficient when $\lambda = \gamma = 0.01$. So for simplification, we just set them to be 0.01 in PS-SRC (ℓ_1-ℓ_1) on all the datasets. Moreover, we also show the different combination of λ and γ in different PS-SRC formulations on the ETH-80 and MoBo datasets in Fig. 10. Fig. 10 shows that by setting λ and γ with different values, the performance of PS-SRC (ℓ_1-ℓ_1) can be further boosted. Moreover, Fig. 10 also shows that the performance of PS-SRC (ℓ_2-ℓ_2) is worse than both PS-SRC (ℓ_1-ℓ_1) and PS-SRC (ℓ_1-ℓ_2).

F. Computational Costs

We also list the computational costs of different methods in Table V. The test bed is a dual cores PC within Intel Xeon CPU (2.53GHz) and 14G RAM, and implementation is based on the Matlab. The reported time costs of our method in Table V include the whole pipeline for predicting a test set, including patch partitioning, partitioning all the test patches into their corresponding clusters based on the centers of training samples, PS-SRC, and label prediction. The preprocessing time (patches partitioning and partitioning all the patches into their corresponding clusters) is about 0.25 second for Youtube Celebrities with 50 images per set, and the time costs for optimization and prediction are about 3.9 and 1.7 seconds for $\ell_1-\ell_1$ formulation and $\ell_1-\ell_2$ formulation, respectively.

In our method, suppose there are N classes in all, each class contains p images, each image contains q patches (q is around 50), and we partition the training set into K clusters, and on average patches from each test sets only fall into

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology

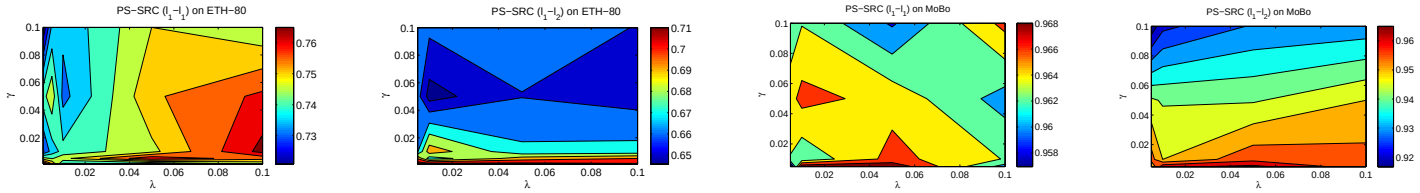


Fig. 10. Accuracy with different λ and γ in PS-SRC on image set classification on ETH-80 and MoBo. Best viewed in color.

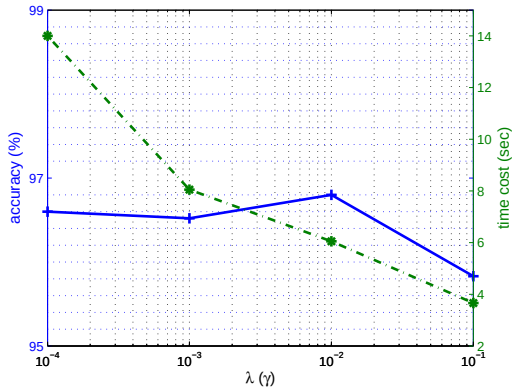


Fig. 11. Effect of different λ (or γ) on the performance of PS-SRC (ℓ_1-ℓ_1) on MoBo. For simplification, we set $\lambda = \gamma$. The dash line corresponds to the time cost. Best viewed in color.

M clusters (M is around 50-100 based on observation in Fig. 3). Therefore the dictionary size for test patches within each cluster is around $\frac{Npq}{K}$. The time cost for patches within each cluster is a function parameterized with $\frac{Npq}{K}$: $T(\frac{Npq}{K})$. Then the total time of our method is around $MT(\frac{Npq}{K})$. But for ISCRC, the dictionary size is Np. If $K = 1000$ and q = 50, then the dictionary size of our method is only $\frac{1}{5}$ of that of ISCRC. Moreover, it is also worthy noting that the feature dimensionality of our method is much smaller than that in ISCRC. As the computational cost increases significantly with the dictionary size and dimensionality of the features to be encoded, therefore though we have many clusters, the computational cost of our method is only slightly slower than that of ISCRC for ℓ_1-ℓ_1 formulation. As for the ℓ_1-ℓ_2 formulation, it is even faster than that of ISCRC. Need to mention that as the number of classes increases, we can increase K accordingly, therefore the dictionary size of our method is relatively stable. But for ISCRC, the dictionary size would be very huge. Therefore though the computational cost of our method is not the fastest, it is scalable to large scale datasets. It is also worth noting that metaface [45] can be used in ISCRC and our method to compress the training set, which will further accelerate the ISCRC and our algorithm.

We also try to cluster the test patches first and partition the training patches based on the cluster centers of test patches on the Youtube Celebrities (50 images per set), the time cost is 113.67 sec per test set, and the accuracy is 73.40%, which is almost the same with our current method. But when training image is 100, the program has run for 3 days without any results. So it proves that it is infeasible to cluster test patches first and partition the training patches based on the cluster center of test patches. Moreover, the main bottleneck of our method lies on the optimization of sparse coding. By using more advanced sparse coding solver, the algorithm can be further sped up.

TABLE V
TIME COST OF DIFFERENT METHODS ON YOUTUBE CELEBRITIES (50 IMAGES PER SET)

method	MMD	MDA	AHISD	CHISD
time(sec)	0.33	0.15	1.22	42.27
SANP	ISCRC	MSSRC	PS-SRC(ℓ_1-ℓ_1)	PS-SRC(ℓ_1-ℓ_2)
213.23	2.92	0.42	4.15	1.95

VI. CONCLUSION

In this paper we propose a simple patch set based representation method for image set classification. To improve the computational efficiency, we propose to cluster the training patches and partition the test patches based on the clusters of the training patches. Accordingly, a nonnegative constraint regularized PS-SRC is used to predict the labels for virtual test patch within each cluster and the labels of all virtual test patches are aggregated with majority voting to predict the label of the whole test set. Moreover, we also show that sparse constraint is important for getting rid of the noisy patches in our formulation. Experimental results on MoBo, Youtube Celebrities, ETH-80, and Kinect Hand Gesture validate the effectiveness of our method for image set classification.

REFERENCES

[1] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[2] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[3] Y.-C. Chen, V. M. Patel, S. Shekhar, R. Chellappa, and P. J. Phillips, "Video-based face recognition via joint sparse representation," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2013.

[4] H. Mobahi, C. Liu, and W. T. Freeman, "A compositional model for low-dimensional image set representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[5] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Kernel methods on the Riemannian manifold of symmetric positive definite matrices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 73–80.

[6] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face recognition from video," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 766–779.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology

[7] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[8] R. Wang and X. Chen, "Manifold discriminant analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[9] S. Chen, C. Sanderson, M. T. Harandi, and B. C. Lovell, "Improved image set classification via joint sparse approximated nearest subspaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[10] Z. Cui, S. Shan, H. Zhang, S. Lao, and X. Chen, "Image sets alignment for video-based face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[11] P. Zhu, L. Zhang, Q. Hu, and S. C. Shiu, "Multi-scale patch based collaborative representation for face recognition with margin distribution optimization," in *Proceedings of the European Conference on Computer Vision*, 2012.

[12] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[13] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[14] E. G. Ortiz, A. Wright, and M. Shah, "Face recognition in movie trailers via mean sequence sparse representation-based classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[15] G. Shakhnarovich, J. W. Fisher, and T. Darrell, "Face recognition from long-term observations," in *Proceedings of the European Conference on Computer Vision*, 2002.

[16] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, 2007.

[17] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara, and O. Yamaguchi, "Recognizing faces of moving people by hierarchical image-set matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[18] W. Fan and D.-Y. Yeung, "Locally linear models on face appearance manifolds with application to dual-subspace based classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[19] L. Wolf and A. Shashua, "Learning over sets using kernel principal angles," *Journal on Machine Learning Research*, vol. 4, pp. 913–931, 2003.

[20] X. Li, K. Fukui, and N. Zheng, "Image-set based face recognition using boosted global and local principal angles," in *Proceedings of the Asian Conference on Computer Vision*, 2010.

[21] A. W. Fitzgibbon and A. Zisserman, "Joint manifold distance: a new approach to appearance based clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[22] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[23] R. Caseiro, P. Martins, J. F. Henriques, F. S. Leite, and J. Batista, "Rolling riemannian manifolds to solve the multi-class classification problem," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 41–48.

[24] Y. Hu, A. S. Mian, and R. Owens, "Sparse approximated nearest points for image set classification," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[25] M. Yang, P. Zhu, L. V. Gool, and L. Zhang, "Face recognition based on regularized nearest points between image sets," in *Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition*, 2013.

[26] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[27] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proceedings of the International Conference on Computer Vision*, 2011.

[28] P. Zhu, W. Zuo, L. Zhang, S. C. K. Shiu, and D. Zhang, "Image set based collaborative representation for face recognition," eprint arXiv:1308.6687, Aug. 2013.

[29] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," in *Proceedings of the Neural Information Processing Systems*, 2006.

[30] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[31] A. Y. Yang, S. S. Sastry, A. Ganesh, and Y. Ma, "Fast ℓ1-minimization algorithms and an application in robust face recognition: A review," in *Proceedings of the IEEE International Conference on Image Processing*, 2010.

[32] P. H. Calamai and J. J. Moré, "Projected gradient methods for linearly constrained problems," *Mathematical programming*, vol. 39, no. 1, pp. 93–116, 1987.

[33] P. O. Hoyer, "Non-negative sparse coding," in *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*. IEEE, 2002, pp. 557–565.

[34] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l 1-ball for learning in high dimensions," in *Proceedings of the International Conference on Machine Learning*, 2008.

[35] R. Gross and J. Shi, "The CMU Motion of Body (MoBo) database," Robotics Institute, Carnegie Mellon University, Tech. Report CMU-RI-TR-01-18, 2001.

[36] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[37] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[38] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[39] Z. Ren, J. Meng, J. Yuan, and Z. Zhang, "Robust hand gesture recognition with kinect sensor," in *Proceedings of ACM international conference on Multimedia*, 2011.

[40] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[41] Z. Ren, J. Yuan, C. Li, and W. Liu, "Minimum near-convex decomposition for robust shape representation," in *Proceedings of the International Conference on Computer Vision*, 2011.

[42] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[43] L. Zhang, M. Yang, X. Feng, Y. Ma, and D. Zhang, "Collaborative representation based classification for face recognition," *arXiv preprint arXiv:1204.2358*, 2012.

[44] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

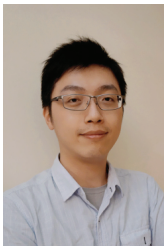
[45] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 1601–1604.

Shenghua Gao is an assistant professor in ShanghaiTech University, China. He received the B.E. degree from the University of Science and Technology of China in 2008 (outstanding graduates), and received the Ph.D. degree from the Nanyang Technological University in 2012. From Jun 2012 to Aug 2014, he worked as a research scientist in Advanced Digital Sciences Center, Singapore. His research interests include computer vision and machine learning. He has published more than 20 papers on object and face recognition related topics in many international conferences and journals, including IEEE T-PAMI,IJCV, IEEE TIP, IEEE TNNLS, IEEE TMM, IEEE TCSVT, CVPR, ECCV, etc. He was awarded the Microsoft Research Fellowship in 2010, and ACM Shanghai Young Research Scientist Award in 2015.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2015.2469571, IEEE Transactions on Circuits and Systems for Video Technology

Zinan Zeng received the master degree and B.E. degree with First Honour from the School of Computer Engineering, Nanyang Technological University, Singapore. He is now a senior software engineer in Advanced Digital Sciences Center, Singapore. His research interests include statistical learning, optimization with application in computer vision.

Kui Jia received the B.Eng. degree in marine engineering from Northwestern Polytechnic University, China, in 2001, the M.Eng. degree in electrical and computer engineering from National University of Singapore in 2003, and the Ph.D. degree in computer science from Queen Mary, University of London, London, U.K., in 2007. He is currently a Visiting Assistant Professor at University of Macau, Macau SAR, China. He is also holding a Research Scientist position at Advanced Digital Sciences Center, Singapore. His research interests are in computer vision, machine learning, and image processing.



Tsung-Han Chan received the B.S. degree from the Department of Electrical Engineering, Yuan Ze University, Taiwan, in 2004 and the Ph.D. degree from the Institute of Communications Engineering, National Tsing Hua University, Taiwan, in 2009. He is currently working as a Project Lead R&D Engineer with Sunplus Technology Co., Hsinchu, Taiwan. His research interests are in image processing and convex optimization, with a recent emphasis on computer vision and hyperspectral remote sensing.

Jinhui Tang is currently a Professor of School of Computer Science and Engineering, Nanjing University of Science and Technology. He received his B.E. and Ph.D. degrees in July 2003 and July 2008 respectively, both from the University of Science and Technology of China (USTC). From July 2008 to Dec. 2010, he worked as a research fellow in School of Computing, National University of Singapore. During that period, he visited School of Information and Computer Science, UC Irvine, from Jan. 2010 to Apr. 2010, as a visiting research scientist. From Sept. 2011 to Mar. 2012, he visited Microsoft Research Asia, as a Visiting Researcher. His current research interests include large-scale multimedia search, social media mining, and computer vision. He has authored over 80 journal and conference papers in these areas. He serves as a editorial board member of Pattern Analysis and Applications, Multimedia Tools and Applications, Information Sciences, Neurocomputing, a Technical Committee Member for about 30 international conferences, and a reviewer for about 30 prestigious international journals. Prof. Tang is a co-recipient of the Best Paper Award in ACM Multimedia 2007, PCM 2011 and ICIMCS 2011. He is a member of ACM, IEEE, and CCF.