Research Article

Improved Collaborative Representation Classifier Based on l_2-Regularized for Human Action Recognition

Shirui Huo,[1] Tianrui Hu,[2] and Ce Li[3,4]

[1]*City University of Hong Kong, Kowloon Tong, Hong Kong*
[2]*Beijing University of Posts and Telecommunications, Beijing, China*
[3]*State Key Laboratory of Coal Resources and Safe Mining, China University of Mining & Technology, Beijing, China*
[4]*University of Chinese Academy of Sciences, Beijing, China*

Correspondence should be addressed to Ce Li; licekong@gmail.com

Human action recognition is an important recent challenging task. Projecting depth images onto three depth motion maps (DMMs) and extracting deep convolutional neural network (DCNN) features are discriminant descriptor features to characterize the spatiotemporal information of a specific action from a sequence of depth images. In this paper, a unified improved collaborative representation framework is proposed in which the probability that a test sample belongs to the collaborative subspace of all classes can be well defined and calculated. The improved collaborative representation classifier (ICRC) based on l_2-regularized for human action recognition is presented to maximize the likelihood that a test sample belongs to each class, then theoretical investigation into ICRC shows that it obtains a final classification by computing the likelihood for each class. Coupled with the DMMs and DCNN features, experiments on depth image-based action recognition, including MSRAction3D and MSRGesture3D datasets, demonstrate that the proposed approach successfully using a distance-based representation classifier achieves superior performance over the state-of-the-art methods, including SRC, CRC, and SVM.

1. Introduction

Human action recognition has been studied in the computer vision community for decades, due to its applications in video surveillance [1], human computer interaction [2], and motion analysis [3]. Prior to the Microsoft Kinect, the conventional research focused on human action recognition from RGB, but Kinect sensors provide an affordable technology to capture RGB and depth (D) images in real time, which can offer better geometric cues and less sensitivity to illumination changes for action recognition. In [1], a bag of 3D points and graphical model are obtained to characterize spatial and temporal information from depth images. In [3], three depth motion maps (DMMs) are projected to capture body shape and motion, which is a discriminant feature to describe the spatiotemporal information of a specific action from a sequence of depth images. Seen from the literature review, although depth based methods appear to be compelling

toward a practical application, even if there are a few of deep-learned features for depth based action recognition, the performance is still far from satisfactory due to the large variations of the motion. In this paper, we focus on leveraging one kind of the structure of representative model to improve performance in multiclass classification with handcrafted DMMs descriptor. In [4], three channel deep convolutional neural networks are trained to extract features of depth map sequences after projecting weighted DMMs on three orthogonal planes at several temporal scales. It was verified that the method using DCNN features can achieve almost the same state-of-the-art results on the MSRAction3D and MSRGesture3D dataset. DCNNs have been demonstrated as an effective kind of models for performing state-of-the-art results in the tasks of image recognition, segmentation, detection, and retrieval. With the success of DCNN, we also take it as feature extraction and apply it in our classifier model.

As for representative models, many achievements based on space representation include image restoration [5], compressive sensing [6, 7], morphological component analysis [8], and super-resolution [9, 10]. As the advances of classifiers based on representation, several pattern recognition problems in the field of computer vision can be effectively solved by sparse coding or sparse representation methods in recent decades. In particular, the linear models can be represented as $y = A\alpha$ [11], where y, α, and A represent the data, a sparse vector, and a given matrix with the overcomplete sample set, respectively. Because of the great success of sparse coding algorithms on image processing, the sparse representation based classifiers, such as sparse representation classification (SRC) and collaborative representation classification (CRC), have gained more attention nowadays.

The basic idea of SRC/CRC is to code the test sample over a set of samples with sparsity constraints, which can be calculated by l_1-minimization. In [12], Wright proposed a basic SRC model for classification by the discriminative nature of sparse representation, which is based on the theory that newly signals are recognized as the linear combinations of previously observed ones. Based on SRC, Yang and Zhang proposed a Gabor occlusion dictionary based SRC, which can significantly reduce the computational cost [13]. In [14], the authors combined sparse representation with linear pyramid matching for image classification. Rather than using the entire training set, Zhang and Li [15] proposed learned dictionary using SRC. In [16], l_1-graph is constructed by a sparse representation subspace over the other samples. Yang et al. [14] also proposed a method to preserve the l_1-graph for image classification by using a subspace to solve misalignment problems in image classification task. Besides, SRC is used for robust illumination [17], image-plane transformation [18], and so on. However, Zhang argued that the good performance of SRC should be largely attributed to the collaborative representation of a test sample by training samples across all classes and proposed more effective CRC. In summary, SRC/CRC simply uses the reconstruction error or residual by each class-specific subspace to determine the class label, and many modified models and solution algorithms to SRC/CRC are also proposed for visual recognition tasks, including Augmented Lagrange Multiplier, Proximal Gradient, Gradient Projection, Iterative Shrinkage-Thresholding, and Homotopy [19]. Recently, some researchers [20, 21] have pointed out the purpose of l_1-regularized based sparsity in pattern classification. On the contrary, using l_2-regularized based representation for classification can do a similar job to l_1-regularized but the computational cost will reduce a lot.

Motivated by the work of modifications of CRC, in this paper, we mainly present the improved collaborative representation classifier (ICRC) based on l_2-regularized for human action recognition. Based on three DMMs' descriptor feature, the ICRC approach is to jointly maximize the likelihood that a test sample belongs to each of the multiple classes, then the final classification is performed by computing the likelihood for each class. The experiments on human action classification tasks, including MSRAction3D and MSRGesture3D datasets, are demonstrated and analyzed on the superior performance of this algorithm over the state-of-the-art methods, including SRC, CRC, and SVM. The rest of the paper is organized as follows. In Section 2, we introduce related feature descriptors using DMM. Section 3 details the action classifier based on ICRC, and Section 4 shows the experimental results of our approach on relevant datasets. The conclusion and acknowledgment are drawn in Section 5 and Acknowledgments section.

2. Feature Descriptors

2.1. Using Depth Motion Maps. In this section, we explain the extracted feature descriptor using depth motion maps (DMMs) from depth images, which is generated by selecting and stacking motion energy of depth maps projected onto three orthogonal Cartesian planes, aligning with front (f), side (s), and top (t) views (i.e., p^f, p^s, and p^t, resp.). As for each projected map, its motion energy is computed by thresholding the difference between consecutive maps. The binary map of motion energy provides a strong clue of the action category being performed and indicates motion regions or where movement happens in each temporal interval. We suggest that all frames should be deployed to calculate motion information instead of selecting frames. Considering both the discriminability and robustness of feature descriptors, we use the l_1-norm of the absolute difference of a frame to define the salient information on depth sequences. Because l_1-norm is invariant to the length of a depth sequence, and l_1-norm contains more salient information than other norms (e.g., l_2), we have

$$\text{DMM}^{\{f,s,t\}} = \sum_{i=1}^{N-v} \left| p_{i+v}^{\{f,s,t\}} - p_i^{\{f,s,t\}} \right|, \tag{1}$$

where v is the frame interval, i represents the frame index, and N is the total number of frames in a depth sequence. In the case that the sum operation in (1) is only used given a threshold satisfied, the scale of v affects little the local pattern histogram on the DMMs.

2.2. Using Deep Convolutional Neural Networks. In this section, we introduce three deep convolutional neural networks (DCNNs) to train the features on three projected planes of DMMs and perform fusion of three nets by combining the softmax in fully connected layer. The layer configuration of our three CNNs is schematically shown in Figure 1, in which there are five convolutional layers and three fully connected layers in each net. The detail of our implementation is illustrated in Section 4.1.2.

3. Action Classifier Based on ICRC

Based on depth motion maps, to incorporate the feature descriptors into a powerful classifier, an improved collaborative representation classifier (ICRC) is presented for human action recognition.

3.1. l_2-Regularized Collaborative Representation Classifier. The basic idea of SRC is to get a test sample by sparsely choosing a small number of atoms from an overcomplete dictionary
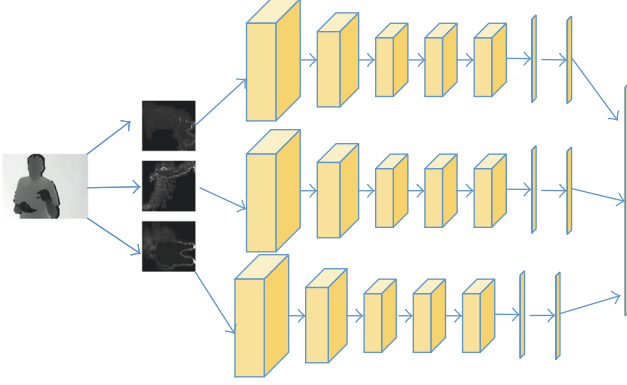
FIGURE 1: Three DCNNs architecture for a depth action sequence to extract features.

that contain all training samples [12]. Denoted by $A_j \in \mathbb{R}^{d \times n}$, the set of training samples form class j, and suppose we have C class of subjects. So $A = [A_1, A_2, \ldots, A_C] \in \mathbb{R}^{d \times n}$ involves many samples from all classes, and A_j $(j = 1, 2, \ldots, C)$ is the individual class of training samples, n is the total number of training samples, and d is the dimension of training samples. A query sample $y \in \mathbb{R}^d$ can be presented by $y = A\alpha$, where y, α, and A represent the data, a sparse vector, and a given matrix with the overcomplete training samples, respectively.

To be specific, in the mechanism of collaborative representation classifier (CRC), each data point in the collaborative subspace can be represented as a linear combination of samples in A, where $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_C]$ is an $n \times 1$ representation vector associated with training sample and α_j $(j = 1, 2, \ldots, C)$ is the subvector corresponding to A_j. Generally, it is formulized as a l_1-norm minimization problem with a convex objective and solved by

$$\min_{\alpha} \quad \left\{ \|y - A\alpha\|_2^2 + \theta \|\alpha\|_1 \right\}, \tag{2}$$
$$\text{s.t.} \quad y = A\alpha,$$

where θ is a positive scalar to balance the sparsity term and the residual. The residual can be computed as

$$e_j = \left\| y - A_j \alpha_j \right\|_2, \tag{3}$$

where α_j is the coefficient vector corresponding to class j. And then the output of the identity of y can be obtained by the lowest residual as

$$\text{class}(y) = \arg \min_j \left\{ e_j \right\}. \tag{4}$$

For more details of SRC/CRC, one can refer to [12]. Because of the computational time consuming in l_1-regularized minimization, (1) is approximated as

$$\min_{\alpha} \quad \left\{ \|y - A\alpha\|_2^2 + \lambda \|L\alpha\|_2^2 \right\}, \tag{5}$$
$$\text{s.t.} \quad y = A\alpha,$$

where $y \in \mathbb{R}^d$ is representative by α and A if the l_2-norm of α is smaller. In (5), L is the Tikhonov regularization [27] to calculate the coefficient vector, and λ is the regularization parameter. $\|L\alpha\|_2$ is the l_2-regularization term to add a certain amount of sparsity to α, which is weaker than l_1-norm minimization. The diagonal matrix L and the coefficient vector α are calculated as follows [21]:

$$L = \begin{pmatrix} \|y - h_1\|_2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \|y - h_k\|_2 \end{pmatrix},$$
$$\alpha = Py, \tag{6}$$

where $P = (A^T A + \lambda L^T L)^{-1} A^T$ is independent of y and precalculated. With (3) and (4), the data y is assigned different identities based on α.

3.2. The Proposed ICRC Method.

Based on the training sample set, we propose an improved collaborative representation classifier based on l_2-regularized term, which assigns the data points with different probabilities based on α by adding a term $\sum_j^C \|A\alpha - A_j \alpha_j\|_2^2$ that attempts to find a point $A_j \alpha_j$ close to the common point y inside each subspace of class j. The first two terms $\|y - A\alpha\|_2^2 + \lambda \|L\alpha\|_2^2$ still form a l_2-regularized collaborative representation term, which encourages to find a point $A\alpha$ close to y in the collaborative subspace. Therefore, (5) is rewritten as

$$\min_{\alpha} \quad \left\{ \|y - A\alpha\|_2^2 + \lambda \|L\alpha\|_2^2 + \frac{\gamma}{C} \sum_j^C \|A\alpha - A_j \alpha_j\|_2^2 \right\}, \tag{7}$$
$$\text{s.t.} \quad y = A\alpha.$$

Obviously, the parameters λ and γ balance three terms, which can be set from the training data. Accordingly, a new solution of representative vector α is obtained from (7).

In the condition of $\gamma = 0$, (7) will degenerate to CRC with the first two terms, and $\|y - A\alpha\|_2^2 + \lambda \|L\alpha\|_2^2$ will play an important role in determining α. When $\gamma > 0$, these two terms $\|y - A\alpha\|_2^2 + \lambda \|L\alpha\|_2^2$ will be the same for all classes, and thus the term $\sum_j^C \|A\alpha - A_j \alpha_j\|_2^2$ will be dominant to further fine-tune α_j by A_j yielding to a precise α. That is, the last newly added term is introduced to further adjust α_j by A_j, resulting in a more stable solution to representative vector α.

We can omit the first two same terms for all classes, make the classifier rule by the last term, and formulize it as a probability exponent:

$$\text{class}(y) = \arg \max_j \left\{ \exp \left(-\|A\alpha - A_j \alpha_j\|_2^2 \right) \right\}. \tag{8}$$

The proposed l_2-regularized method for human action recognition is presented to maximize the likelihood that a test sample belongs to each class, then the experiments in the following section show that it obtains a final classification by checking which class has the maximum likelihood. So far, the abovementioned classifier model in (7) and (8) is named as the improved collaboration representation classifier (ICRC).
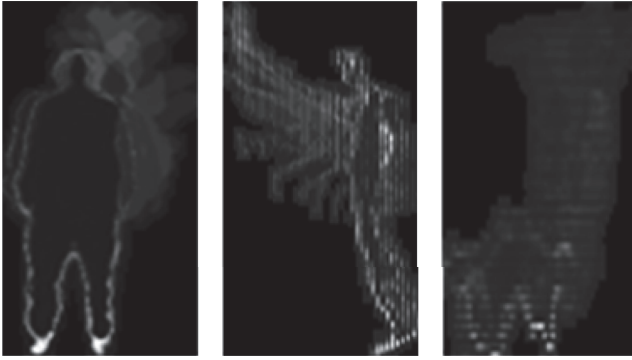
FIGURE 2: Three DMMs of a depth action sequence "ASL Z" from the front (f) view, side (s) view, and top (t) view, respectively.

FIGURE 3: Three DMMs of a depth action sequence "Swipe left" from the front (f) view, side (s) view, and top (t) view, respectively.

4. Experimental Results

Based on depth motion maps, to incorporate the feature descriptors into a powerful classifier, an ICRC is presented for human action recognition. To verify the effectiveness of the proposed ICRC algorithm on action recognition applications using DMM descriptors of depth sequences, we carry out experiments on challenging depth based action datasets MSRAction3D [1] and MSRGesture3D [1] for human action recognition.

4.1. Feature Descriptors

4.1.1. DMMs. The MSRAction3D [1] dataset is composed of depth images sequence captured by the Microsoft Kinect-V1 camera. It includes 20 actions performed by 10 subjects facing the camera. Each subject performed each action 2 or 3 times. There are 20 action types: high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw X, draw tick, draw circle, hand clap, two-hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, and pick up and throw. The size of each depth image is 240 × 320 pixels. The background information has been removed in the depth data.

The MSRGesture3D [1] dataset is for continuous online human action recognition from a Kinect device. It consists of 12 gestures defined by American Sign Language (ASL). Each person performs each gesture 2 or 3 times. There are 333 depth sequences. For action recognition on the MSRAction3D and MSRGesture3D dataset, we use the feature computed from the DMMs, and each depth action sequence generates three DMMs corresponding to three projection views. The DMMs of high arm wave class from the MSRAction dataset are shown in Figure 2, and the DMMs of ASL Z class from the MSRGesture3D dataset are shown in Figure 3.

4.1.2. DCNNs. Furthermore, our implementation of DCNN features is based on the publicly available MatConvNet toolbox [28] using one Nvidia Titan X card. The network weights are learned by mini-batch stochastic gradient descent. Similar to [4], the momentum is set to 0.9 and weight decay is set to 0.0005, and all hidden weight layers use the rectification activation function. At each iteration, 256 samples in each
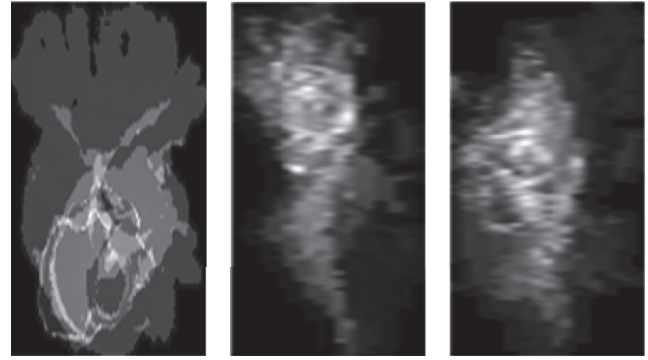
batch are constructed and resized to 256 × 256, then 224 × 224 patches are randomly cropped from the center of the selected image to artificial data augmentation. The dropout regularization ratio is 0.5 in the nets. Besides, the initial learning rate is set to 0.01 with pretrained model on ILSVRC-2012 to fine-tune our model, and the learning rate decreases every 20 epochs. Finally, we concatenate three 4096 dimensional feature vectors in 7th fully connected layer to input the subsequent classifier.

4.2. Experiment Setting. The same experimental setup in [1] was adopted, and the actions in MSRAction3D dataset were divided into three subsets as follows: AS1: horizontal wave, hammer, forward punch, high throw, hand clap, bend, tennis serve, and pickup throw; AS2: high wave, hand catch, draw x, draw tick, draw circle, two-hand wave, forward kick, and side boxing; AS3: high throw, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, and pickup throw. We performed three experiments with 2/3 training samples and 1/3 testing samples in AS1, AS2, and AS3, respectively. Thus, the performance on MSRAction3D is evaluated by the average accuracy (Accu., unit: %) on three subsets. On the other hand, the same experimental setting reported in [26, 29, 30] was followed. 12 gestures were tested by leave-one-subject-out cross-validation to evaluate the performance of the proposed method.

4.3. Recognition Results with DMMs and ICRC. We concatenate the sign, magnitude, and center features to form the feature based on DMMs as the final feature representation. The compared methods are similar to [29, 30]. The same parameters reported in [26] were used here for the sizes of SI and block. A total of 20 actions are employed and one-half of the subjects (1, 3, 5, 7, and 9) are used for training and the remaining subjects are used for testing. The recognition performance of our method and existing approaches are listed in Table 1. It is clear that our method achieves better performance than other competing methods.

To show the outcome of our method, Figures 4 and 5 illustrate the recognition rates of each class in two datasets. It is stated that there are 14 classes obtaining 100% recognition rates in the MSRAction3D dataset, and the performance of 3

TABLE 1: Recognition accuracies (unit: %) comparison on the MSRAction3D dataset and MSRGesture3D dataset.

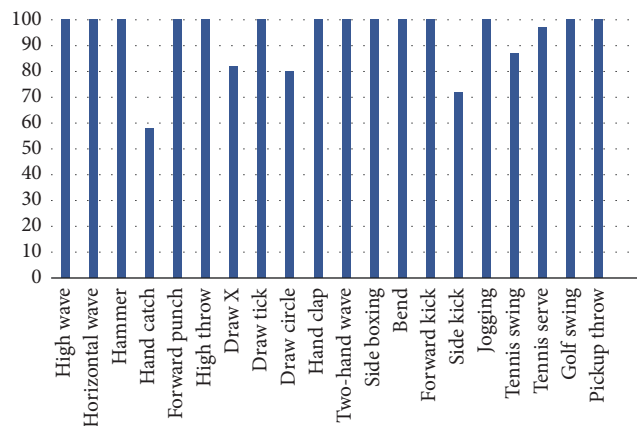| Method	Accu. (%) on two datasets	
	MSRAction3D	MSRGesture3D
DMM-HOG [3]	85.5	89.2
Random Occupancy [22]	86.5	88.5
Actionlet Ensemble [23]	88.2	88.2
Depth Cuboid [24]	89.3	90.5
Vemulapalli [25]	89.5	87.7
DMM + CRC [26]	92.3	92.5
DMM + ICRC (proposed)	*93.8*	*94.8*

FIGURE 4: Recognition rates (unit: %) of 20 classes in MSRAction3D dataset (average results of three subsets).
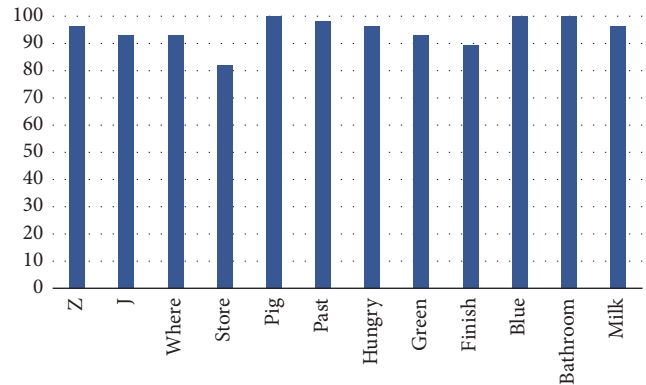

FIGURE 5: Recognition rates (unit: %) of 12 classes in MSRGesture3D dataset.

TABLE 2: Recognition accuracies (unit: %) comparison of DMM + ICRC and DCNN + ICRC on the MSRAction3D dataset and MSRGesture3D dataset.

Method	Accu. (%) on two datasets	
	MSRAction3D	MSRGesture3D
DMM + ICRC (proposed)	93.8	94.8
DCNN + ICRC (proposed)	99.99	100.0

classes reaches up to best in the MSRGesture3D dataset. All experiments are carried out using MATLAB 2016b on an Intel i7-6500U desktop with 8 GB RAM, and the average time of video processing gets about 26 frames per second, meeting a real-time processing demand basically.

4.4. Comparison with DCNN Features and ICRC. Furthermore, in order to evaluate our proposed classifier method, we also extract the deep features by the abovementioned conventional CNN model and then input the 12288 dimensional vectors to the proposed ICRC for action recognition. Table 2 shows that DCNN algorithm indeed has advances as good as in other popular tasks of image classification and object detection, and it can improve the accuracy greatly up to 6% in MSRAction3D and MSRGesture3D. This would also explain the importance of effective feature to ICRC classifier.

5. Conclusion

In this paper, we propose improved collaborative representation classifier (ICRC) based on l_2-regularized for human action recognition. The DMMs and DCNN feature descriptors are involved as an effective action representation. For the action classifier, ICRC is proposed based on collaborative representation with the additional regularization term. The new insight focuses on a subspace constraints on the solution. The experimental results on MSRAction3D and MSRGesture3D

show that the proposed algorithm performs favorably against the state-of-the-art methods, including SRC, CRC, and SVM. Future work will focus on involving the deep-learned network in the depth image representation and evaluating more complex datasets such as MSR3DActivity, UTKinect-Action, and NTU RGB+D, for the action recognition task.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '10)*, pp. 9–14, San Francisco, Calif, USA, June 2010.

[2] S. Wang, Z. Ma, Y. Yang, X. Li, C. Pang, and A. G. Hauptmann, "Semi-supervised multiple feature analysis for action recognition," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 289–298, 2014.

[3] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients,"

in *Proceedings of the 20th ACM International Conference on Multimedia, MM 2012*, pp. 1057–1060, jpn, November 2012.

[4] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Deep convolutional neural networks for action recognition using depth map sequences," *Computer Vision and Pattern Recognition*, arXiv preprint, https://arxiv.org/abs/1501.04686, 2015.

[5] W. E. Vinje and J. L. Gallant, "Sparse coding and decorrelation in primary visual cortex during natural vision," *Science*, vol. 287, no. 5456, pp. 1273–1276, 2000.

[6] E. Candès, "Compressive sampling," in *Proceedings of the International Congress of Mathematics*, vol. 3, pp. 1433–1452, 2006.

[7] N. Guan, D. Tao, Z. Luo, and B. Yuan, "NeNMF: an optimal gradient method for nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2882–2898, 2012.

[8] J.-L. Starck, M. Elad, and D. Donoho, "Redundant multiscale transforms and their application for morphological component separation," *Advances in Imaging and Electron Physics*, vol. 132, pp. 287–348, 2004.

[9] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

[10] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 1838–1857, 2011.

[11] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Proceedings of the NIPS*, pp. 609–616, Vancouver, Canada, December 2006.

[12] 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.

[13] M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with gabor occlusion dictionary," in *Proceedings of the 11th European Conference on Computer Vision (ECCV '10)*, pp. 448–461, Springer, Crete, Greece, 2010.

[14] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1794–1801, June 2009.

[15] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2691–2698, June 2010.

[16] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with l1-graph for image analysis," *IEEE Transactions on Image Processing*, vol. 19, no. 4, pp. 858–866, 2010.

[17] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Towards a practical face recognition system: Robust registration and illumination by sparse representation," in *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, pp. 597–604, usa, June 2009.

[18] J. Z. Huang, X. L. Huang, and D. Metaxas, "Simultaneous image transformation and sparse representation recovery," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, Anchorage, Alaska, USA, June 2008.

[19] A. Y. Yang, A. Genesh, Z. Zhou, S. Sastry, and Y. Ma, "A Review of Fast L1-Minimization Algorithms for Robust Face Recognition," Defense Technical Information Center, 2010.

[20] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 471–478, Barcelona, Spain, November 2011.

[21] P. Berkes, B. L. White, and J. Fiser, "No evidence for active sparsification in the visual cortex," in *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems, NIPS 2009*, pp. 108–116, can, December 2009.

[22] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part II*, pp. 872–885, Springer, Berlin, Germany, 2012.

[23] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 1290–1297, Providence, RI, USA, June 2012.

[24] L. Xia and J. K. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 2834–2841, IEEE, Portland, Ore, USA, June 2013.

[25] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 588–595, Columbus, Ohio, USA, June 2014.

[26] C. Chen, M. Liu, B. Zhang, J. Han, J. Junjun, and H. Liu, "3D action recognition using multi-temporal depth motion maps and fisher vector," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016*, pp. 3331–3337, usa, July 2016.

[27] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, John Wiley & Sons, New York, NY, USA, 1977.

[28] http://www.vlfeat.org/matconvnet/.

[29] Y. Yang, B. Zhang, L. Yang, C. Chen, and W. Yang, "Action recognition using completed local binary patterns and multiple-class boosting classifier," in *Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015*, pp. 336–340, mys, November 2016.

[30] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. M. Campos, "On the improvement of human action recognition from depth map sequences using space-time occupancy patterns," *Pattern Recognition Letters*, vol. 36, no. 1, pp. 221–227, 2014.