DEPTH HUMAN ACTION RECOGNITION BASED ON CONVOLUTION NEURAL NETWORKS AND PRINCIPAL COMPONENT ANALYSIS

Manh-Quan Bui[1], Viet-Hang Duong[1], Tzu-Chiang Tai[2], and Jia-Ching Wang[1]

[1]Dept. of Computer Science and Information Engineering, National Central University, Jhongli, Taiwan
[2]Dept. of Computer Science and Information Engineering, Providence University, Taichung, Taiwan

ABSTRACT

In this work, we address human action recognition problem under viewpoint variation. The proposed model is formulated by wisely combining convolution neural network (CNN) model with principle component analysis (PCA). In this context, we pass real depth videos through a CNN model in a frame-wise manner. The view invariant features are extracted by employing convolution layers as mid-outputs and considered as 3D nonnegative tensors. The PCA algorithm is separately imposed on view-invariant high-level space of image and video groups to seek both local and holistic hidden dynamics information. To deal with noisy data and temporal misalignment, we utilize the Fourier temporal pyramid to encode temporal and obtain the final descriptors. Our proposed framework supplies a robust discriminative representation with low dimension and computational requirements. We evaluate the proposed method on two standard multiview depth video datasets. The experimental results show that our method has superior performance compared to other approaches.

Index Terms— Human action recognition, convolution neural network, principal component analysis, view invariance, feature representation.

1. INTRODUCTION

Recently, human action recognition has been an interesting topic in computer vision because it offers various capacities involving video surveillance, video games, robotics, etc. Along with the appearance of Kinect sensor [3], a kind of real-time and low-cost depth camera, the problems of action recognition with the depth images have been figured out and attracted to research community. Depth images themselves contain some useful information, e.g., geometry, shape, and texture, which allow action understanding easier. Observationally, there are many effective methods to handle the videos captured from a fixed viewpoint [4–9]. However, dealing with the videos acquired from viewpoint changes is still a big challenge. The main reason here is that an action acquired from different viewpoints gives different shapes and poses which significantly affect action perception. To solve this issue, the view-invariant features should be extracted [10–18]. For instance, Junejo *et al*. [10] extracted histogram of oriented gradients (HOG) and histogram of optical flows (HOF) to achieve the view-invariance for RGB videos. Discovering the local spatio-temporal features from the most discriminative 3D pointclouds for depth videos was proposed by Rahmani *et al*. [14, 16]. More recently, Rahmani and Mian [18] performed view-invariant features from a CNNs-based human pose model (HPM).

Principal component analysis (PCA) is known as a powerful technique for dimensionality reduction and multivariate analysis. It was first introduced by Pearson [1] and developed independently by Hotelling [2]. Several extensions of PCA have been developed for action recognition task, particularly the models of jointing PCA with CNNs. C. Colombo *et al*. [19] explored further "one action, one eigenspace" by using PCA technique for reducing dimension and increasing discriminative training. Q. Le *et al*. [20] integrated the independent subspace analysis (ISA) and PCA to generate a deep learning model that learns invariant spatio-temporal features from unlabeled video data. S. Ji *et al*. [21] constructed a model by connecting multiple 3D-CNN models and applied PCA in one 3D-CNN model as a dimensional reduction tool in order to obtain an auxiliary output.

The key point of PCA technique is that it is able to revoke the ill correlation and extract the most relevant features. Meanwhile, the convolutional layers of CNNs can capture discriminative features from both the spatial and the temporal dimensions [21, 22].

Motivated by the aforementioned ideas, we propose a view invariant representation model for depth human action recognition based on CNN and PCA (CNN-PCA). The designed model exploits a pre-trained human pose structure [18] to obtain a discriminative data space. We map each frame in the video to a view-invariant high level space by taking the *4-th* convolution layer activations for view-invariant descriptors. The PCA algorithm is applied on each descriptor to collect local motion information. Moreover, in order to capture global dynamic information within viewpoints and between action classes, we create the overlapping descriptors, and use PCA algorithm to further

extract features and reduce dimensional space. These obtained features are then temporally aligned by the Fourier temporal pyramid algorithm to produce the final representation. The proposed framework is evaluated on the Multiview UWA3D-II [16] and the Northwestern-UCLA [23] datasets. The experiment results reveal that our method achieves higher accuracies than the state-of-the-art competitor [18].

2. RELATED WORKS

Depth-based methods. Depth-based human action recognition techniques can be divided into two main categories including holistic and local approaches. The former are commonly used to discover global features from silhouettes and space-time volume information [5, 9, 24, 25]. Specifically, Li *et al*. [24] set up an expandable graphical model which represents the postures and dynamics of the explicit postures and obtains holistic dynamics from the contour of the silhouettes. Yang *et al*. [25] introduced a discriminative action model, depth motion maps (DMM), to capture the global actions and result in the histograms of oriented gradients (HOG) descriptor from the motion map. Oreifej and Liu proposed a new holistic method that uses a histogram of oriented 4D surface normal (HON4D) to capture the complex and jointed structure and motion within the sequence videos [5]. Yand and Tian [9] extended HON4D and proposed super normal vector (SNV) that splits a depth video into a set of space-time grids based on adaptive spatio-temporal pyramid.

For the local perspective, the set of interest points is significantly considered [14, 16, 26–28]. Filtering the noise received from depth sensor [28] or exploiting histogram of oriented principal components (HOPC) to represent spatio-temporal interest points [14, 16] are classical models.

Deep learning methods: Convolutional neural networks have had an impressive success in large-scale image and video recognition. As a type of deep models, CNNs are able to achieve superior performance on visual object recognition tasks. Moreover, CNNs have been shown to be invariant to challenges of image and video processing such as pose variations, lighting conditions, background clutter, and camera viewpoint changes [29]. For action recognition, several deep network models have been proposed such as the convolutional gated RBMs [30], the 3D CNNs [21]. Noticeably, the single-frame model can perform equally well as the multi-frame model [31] and a single non-linear virtual path between all actions and all camera viewing directions can be learnt through a deep network named non-linear knowledge transfer mode (NKTM) [15]. When the NKTM is learned, dense trajectories of synthetic points fitted to mocap data are extracted and a large corpus of action video training data is required. However, both of them are not reliable and available in the case of depth videos. In the context of unseen poses, Rahmani and Mian [18] proposed an effective depth image representation which is robust to depth noise and temporal misalignment.

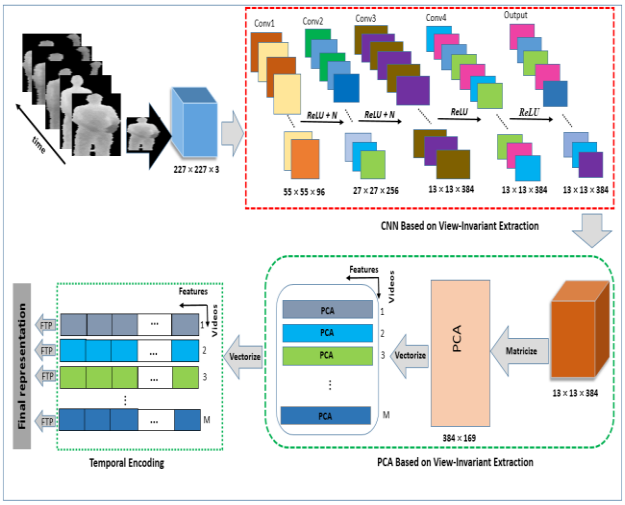

Fig.1. The proposed model for feature extraction

3. THE PROPOSED MODEL

3.1. Model architecture

Our proposed model is based on HPM model [18] whose architecture is similar to AlexNet [32]. When feeding input data into CNNs, we do not use any fully connected layer. Instead, we exploit the convolutional layer activations to capture the information from local neighborhoods of a single image. Assume that $\boldsymbol{C}(m, n, s)$ denotes a convolutional layer with kernel size $m \times m$, n filters and a stride of s. $\boldsymbol{P}(m, s)$ is a max pooling layer of kernel size $m \times m$ and a stride of s. We denote $\boldsymbol{N}$ as a normalization layer, and $\boldsymbol{ReLU}$ as a rectified linear unit. Our proposed CNN structure is as follows:

$\boldsymbol{C}(11,96,4) \rightarrow \boldsymbol{ReLU} \rightarrow \boldsymbol{P}(3,2) \rightarrow \boldsymbol{N} \rightarrow \boldsymbol{C}(5,256,1) \rightarrow \boldsymbol{ReLU} \rightarrow \boldsymbol{P}(3,2) \rightarrow \boldsymbol{N} \rightarrow \boldsymbol{C}(3,384,1) \rightarrow \boldsymbol{ReLU} \rightarrow \boldsymbol{C}(3,384,1) \rightarrow \boldsymbol{ReLU}$.

Note that the outputs herein are convolution layers whose elements are non-negative and obtained by a max rectified linear unit function. In this stage, each action depth image is divided into 384 patches of size 13×13.

3.2. Feature extraction

In order to match the input dimension of CNN-based model, we cropped and resized each depth video frame to 227×227 and then passed its subtracted mean through the HPM network to extract the viewpoint invariant features. We considered each output from CNN as a tensor of size 13×13×384 which is believed to contain almost view-invariant features from local neighborhoods of an individual depth image. Let M denote the number of total action videos in the dataset and f denote the number of frames in a given

TABLE I. COMPARISON OF ACTION RECOGNITION ACCURACY (%) ON THE UWA3D-II DATASET

Training views	V1 & V2		V1 & V3		V1 & V4		V2 & V3		V2 & V4		V3 & V4		Mean (%)
Test view	V3	V4	V2	V4	V2	V3	V1	V4	V1	V3	V1	V2	
Input: Depth images													
CCD [34]	10.5	13.6	10.3	12.8	11.1	8.3	10.0	7.7	13.1	13.0	12.9	10.8	11.2
HON4D [5]	31.1	23.0	21.9	10.0	36.6	32.6	47.0	22.7	36.6	16.5	41.4	26.8	28.9
SNV [9]	31.9	25.7	23.0	13.1	38.4	34.0	43.3	24.2	36.9	20.3	38.6	29.0	29.9
DVV [35]	23.5	25.9	23.6	26.9	22.3	20.2	22.1	24.5	24.9	23.1	28.3	23.8	24.1
CVP [36]	25.0	25.6	25.5	28.2	24.7	24.0	23.0	24.5	26.6	23.3	30.3	26.8	25.6
HOPC [14]	52.7	51.8	59.0	57.5	42.8	44.2	58.1	38.4	63.2	43.8	66.3	48.0	52.2
HPM+TM [18]	80.6	80.5	75.2	82.0	65.4	72.0	77.3	67.0	83.6	81.0	83.6	74.1	76.9
Our method	**83.6**	**82.8**	**83.5**	**88.4**	**76.3**	**81.7**	**80.7**	**83.9**	**85.1**	**85.8**	**85.9**	**82.0**	**83.3**

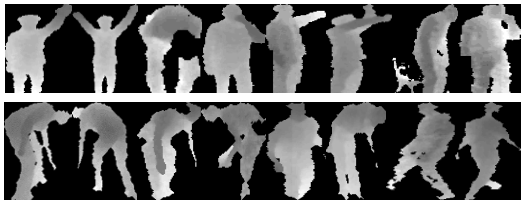

Fig.2. Video frame samples from the Multiview UWA3D-II dataset [16] (the top row) and from the Northwestern-UCLA dataset [23] (the bottom row)

video. Assume the output of the *t-th* image in the *k-th* video sample is $x_t^{(k)}$, where $x_t^{(k)} \in \mathbb{R}_+^{13\times13\times384}$, $1 \le t \le f$, and $1 \le k \le M$. For each frame, the representative tensor $x_t^{(k)}$ is normalized to $P_t^{(k)} = \frac{x_t^{(k)}}{\left\|x_t^{(k)}\right\|_F}$, and $P_t^{(k)}$ is then matricized to $\mathbf{X}_t^{(k)} \in \mathbb{R}_+^{N\times R}$ in 3-mode, where N=384, and R=169. The empirical covariance matrix of $\mathbf{X}_t^{(k)}$, denoted as $C_t^{(k)}$, is defined as follows

$$C_t^{(k)} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{X}_t^{(k)} - \mu_t^{(k)})(\mathbf{X}_t^{(k)} - \mu_t^{(k)})^T \qquad (1)$$

where $\mu_t^{(k)} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{X}_t^{(k)}$.

We aim at finding out a projective basis that can filter out the noise and reveal some hidden dynamics of each action. Thus, we seek the linear mappings $\mathbf{Q}_t^{(k)}$ that maximize the amount of variance in the corresponding matrices $\mathbf{X}_t^{(k)}$ by applying the embedded PCA technique (EPCA) in [33]. The problem turns into solving an optimization problem as follows:

$$\max_{\mathbf{Q}_t^{(k)}} trace(\mathbf{Q}_t^{(k)T} C_t^{(k)} \mathbf{Q}_t^{(k)}) \text{ s.t } \left\|q_{:j}^{(k)}\right\|_F^2 = 1 \qquad (2)$$

where $q_{:j}^{(k)}$ denotes the *j-th* column of the matrix $\mathbf{Q}_t^{(k)}$. The matrix $\mathbf{X}_t^{(k)}$ is then mapped onto the linear basis $\mathbf{Q}_t^{(k)}$ to obtain the low-dimensional matrix $\mathbf{Y}_t^{(k)} = \mathbf{X}_t^{(k)}\mathbf{Q}_t^{(k)}$. By denoting $\mathbf{z}_t^{(k)} = vect(\mathbf{Y}_t^{(k)})$, we obtain the first representation from the *k-th* video sample as $\mathbf{Z}^{(k)} = [\mathbf{z}_1^{(k)}\mathbf{z}_2^{(k)}...\mathbf{z}_f^{(k)}]$.

TABLE II. COMPARISON OF ACTION RECOGNITION ACCURACY (%) ON THE NORTHWESTERN-UCLA DATASET

Method	Recognition accuracy (%)
Input: Depth images	
CCD [34]	34.4
HON4D [5]	39.9
SNV [9]	42.8
DVV [35]	52.1
CVP [36]	53.5
HOPC [14]	80.0
HPM+TM [18]	92.0
Our method	**93.93**

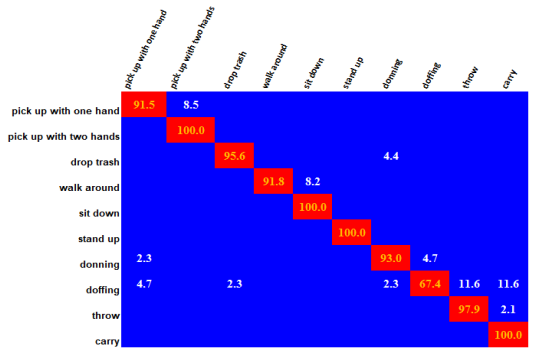

Fig. 3 The confusion matrix of the proposed method on Northwestern-UCLA dataset

Let a sub-matrix of $\mathbf{Z}^{(k)}$ be denoted as $\mathbf{Z}_{sub}^{(k)}$, which contains a random number (but not equal) of columns of the matrix $\mathbf{Z}^{(k)}$. We concatenate the matrix $\mathbf{Z}^{(k)}$ and its neighborhood matrices to obtain the overlapping matrix $\mathbf{P}^{(k)} = [\mathbf{Z}^{(k)}\mathbf{Z}_{sub}^{(k+1)}...\mathbf{Z}_{sub}^{(k+i)}]$, where $1 \le i \le M$, and $k < k+i < M$. Obviously, these matrices contain the dynamic information shared by the viewpoints, intra-action classes, and inter-action classes. In order to harvest the global information of the *k-th* video, we again impose EPCA algorithm on each matrix $\mathbf{P}^{(k)}$. The matrix $\mathbf{P}^{(k)}$ is then mapped onto its linear basis to produce the new features, which are denoted as $\mathbf{V}^{(k)}$. However, these descriptors are not strongly discriminative and even contain error sources. Therefore, we need to seek a robust representation to encode the temporal and convey much more holistic information.

3.3. Temporal encoding and classification

To obtain the discriminative descriptor and encode the temporal, we further extract global information of the entire video. The matries $\mathbf{V}^{(k)}$ are firstly vectorized to form vectors $\mathbf{v}^{(k)}$. This procedure is to isolate the video position. Then we employ the Fourier temporal pyramid (FTP) [13] on the vectors $\mathbf{v}^{(k)}$. More specifically, we perform the fast Fourier transform (FFT) on all elements of vectors $\mathbf{v}^{(k)}$ to find the first q low frequency components. Finally, these coefficients are concatenated into a vector $\mathbf{f}^{(k)}$ to form the final descriptor for the *k-th* video. Figure 1 shows our proposed framework for feature extraction.

In the matching scheme, we perform one-vs-all strategy on the extracted feature vectors using linear support vector machines (SVM) [39].

4. EXPERIMENT

This section evaluates the proposed model on two well-known depth datasets: UWA3D Multiview Activity II [16] and Northwestern-UCLA Multiview Action3D [23]. The baseline results are obtained by using publicly available implementations of [5, 9, 14, 18, 34-36] or from the original papers. We used the MatConvNet toolbox [37] and tensor toolbox [38] to implement convolutional neural networks and tensor representation, respectively. In our experiments, we set the number of Fourier pyramid levels $l = 1$, and the number of the first low frequency Fourier coefficients $q = 4$. In classification phase, we imposed the sparse constraint on both training and testing sets.

4.1. UWA3DII dataset [16]

This dataset consists of 30 human actions that are performed by 10 subjects and captured from various viewpoints with the following styles: (1) one hand waving, (2) one hand punching, (3) two hand waving, (4) two hand punching, (5) sitting down, (6) standing up, (7) vibrating, (8) falling down, (9) holding chest, (10) holding head, (11) holding back, (12) walking, (13) irregular walking, (14) lying down, (15) turning around, (16) drinking, (17) phone answering, (18) bending, (19) jumping jack, (20) running, (21) picking up, (22) putting down, (23) kicking, (24) jumping, (25) dancing, (26) moping floor, (27) sneezing, (28) sitting down (chair), (29) squatting, and (30) coughing. This is a challenging dataset because many actions are highly similar to each other among action classes. Furthermore, the videos were received at different times from varying viewpoints and the data may contain self-occlusions. Figure 2 shows some samples of UWA3D Multiview Activity II [16] on the top row and some samples of Northwestern-UCLA Multiview Action3D [23] on the bottom row.

In the experiments, each time we used two views for training and the rest two views for testing, as in [18]. Table I compares the results of our method and other approaches. It can be observed that our method improves performance significantly to compare with the state-of-the-art work in [18]. Moreover, our method achieves the highest recognition accuracies on all views. The experiments reveal that the proposed model is robust to depth videos captured from multiple viewpoints.

4.2. Northwestern-UCLA dataset [23]

This dataset was captured simultaneously by three Kinect cameras from different views. It contains RGB, depth and human skeleton of the following ten actions: (1) pick up with one hand, (2) pick up with two hands, (3) drop trash, (4) walk around, (5) sit down, (6) stand up, (7) donning, (8) doffing, (9) throw, and (10) carry. Each action was performed one to six times by ten different subjects. This dataset is challenging because of the following reasons: (1) there are many similar actions, e.g., “pick up with one hand” and “pick up with two hands”, “doffing” and “throw”; (2) the same information, e.g., “walking”, are shared by various actions.

To firm the robustness of the introduced descriptor, we used only the samples from one camera for training and the samples from another camera for testing. This is different from the works in [18] and [23] that took the samples from two cameras for training. Table II compares the results of our method and other depth-based human action recognition approaches.

It deserves particular notice that our method utilized less training samples than other approaches but it obtained the highest recognition rate. The confusion matrix was computed and described in Figure 3. As our expectation, the proposed approach performs a pretty high recognition accuracy. Specially, it recognizes one hundred percent accuracy rate for four actions including “pick up with two hands”, “sit down”, “stand up”, and “carry”. The method gets the lowest accuracy rate for the “doffing” action because it is easily confused with “pick up one hand”, “drop trash”, “throw” and “carry”.

5. CONCLUSION

We introduce a robust descriptor for depth human action recognition in the context of viewpoint changes. The depth action images are fed forward frame by frame into the CNN model to obtain spatio-temporal features. The *4-th* convolution layer activations are exploited as the CNN outputs and then projected on the discriminative space encoded by PCA. The pyramidal Fourier coefficients are found to align temporal and form the global representation of the video. The experimental results on two multiview benchmark datasets show that our approach significantly outperforms the existing state-of-the-art methods.

6. REFERENCES

[1] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Phil. Mag*, vol. 2, no. 6, pp. 559-572, 1901.

[2] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *JEP*, vol. 24, pp. 417-441, 1993.

[3] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *ACM on Communication*, vol. 56, no. 1, pp. 116-124, Jan. 2013.

[4] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. ACM ICM*, Nov. 2012, pp. 1057-1060.

[5] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proc. IEEE CVPR*, 2013, pp. 716-723.

[6] H. Rahmani, D. Q. Huynh, A. Mahmood, and A. Mian, "Discriminative human action classification using locality constrained linear coding," *Pattern Recognition Letters*, Elsevier, vol. 72, pp. 62-71, Mar. 2016.

[7] H. Rahmani, A. Mahmood, A. Mian, and D. Huynh, "Real time action recognition using histograms of depth gradients and random decision forests," in *Proc. IEEE WACV*, 2014.

[8] A. Shahroudy, T.T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE PAMI*, vol. 38, no. 10, pp. 2123-2129, 2016.

[9] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *Proc. IEEE CVPR*, 2014, pp. 804-811.

[10] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities," *IEEE PAMI*, vol. 33, no. 1, pp. 172–185, Jan. 2011.

[11] A. Gupta, J. Martinez, J. J. Little, and R. J. Woodham, "3D pose from motion for cross-view action recognition via nonlinear circulant temporal encoding," in *Proc. IEEE CVPR*, 2014, pp. 2601-2608.

[12] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," in *Proc. IEE CVPR*, 2014, pp. 588-595.

[13] J. Wang, Z. Liu, and Y. Wu, *Learning Actionlet Ensemble for 3D Human Action Recognition*, Springer, chapter 2, pp. 11-40, Jan. 2014.

[14] H. Rahmani, A. Mahmood, D. Q Huynh, and A. Mian, "HOPC: Histogram of oriented principal components of 3d pointclouds for action recognition," in *Proc. ECCV*, 2014, pp. 742-757.

[15] H. Rahmani and A. Mian, "Learning a non-linear knowledge transfer model for cross-view action recognition," in *Proc. IEEE CVPR*, 2015, pp. 2458-2466.

[16] H. Rahmani, A. Mahmood, D. Q Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE Trans. PAMI*, vol. 38, no. 12, pp. 2430-2443, 2016.

[17] H. Rahmani, A. Mian, and M. Shah, "Learning a deep model for human action recognition from novel viewpoints," *IEEE Trans. PAMI*, vol. PP, no. 99, pp. 1-1, 2016.

[18] H. Rahmani and A. Mian, "3D action recognition from novel viewpoints," in *Proc. IEEE CVPR*, 2016, pp. 1506-1515.

[19] C. Colombo, D. Comanducci, and A. Del Bimbo, "Compact representation and probabilistic classification of human actions in videos," in *Proc. IEEE AVSS*, 2007, pp. 342-346.

[20] Q. Le, W. Zou, S. Yeung, and A. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. IEEE CVPR*, 2011, pp. 3361-3368.

[21] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," IEEE *Trans. PAMI*, vol. 35, no.1, pp. 221-231, 2013.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE ICCV*, 2015, pp. 4489-4497.

[23] J. Wang, X. Nie, Y. Xia, Y. Wu, and S. Zhu, "Cross-view action modeling, learning and recognition" in *Proc. IEEE CVPR*, 2014, pp. 2649-2656.

[24] W. Li, Z. Zhang, and Z. Liu, "Expandable data-driven graphical modeling of human actions based on salient postures," *IEEE Trans. CSVT*, vol. 18, no. 11, pp. 499–1510, 2008.

[25] X. Yang, C. Zhang, Y.L. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. ACM ICM*, 2012, pp. 1057-1060.

[26] J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *Proc. ECCV*, 2012, pp. 872-885.

[27] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Proc. IEEE CVPR*, 2013, pp. 2834-2841.

[28] J. Wang, Z. Liu, Y. Wu, J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE CVPR*, 2012.

[29] Y. LeCun, F.J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proc. IEEE CVPR*, 2004.

[30] G.W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Proc. ECCV*, Springer, 2010, pp.140-153.

[31] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE CVPR*, 2014, pp. 1725-1732.

[32] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012.

[33] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. "Dimensionality reduction: a comparative review," Tilburg University Technical Report, TiCC-TR 2009-005, 2009, pp. 1-3.

[34] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *Proc. ECCVW*, Springer, 2012, pp. 52-61.

[35] R. Li and T. Zickler, "Discriminative virtual views for cross-view action recognition," in *Proc. IEEE CVPR*, Jun. 2012.

[36] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, and C. Shi, "Cross-view action recognition via a continuous virtual path," in *Proc. IEEE CVPR*, 2013, pp. 2690-2697.

[37] A. Vedaldi and K. Lenc, "MatConvNet-Convolutional Neural Networks for MATLAB," in *Proc. ACM ICM*, 2015.

[38] B. Bader and T.G. Kolda, "Tensor Toolbox Version 2.5," [Online]. Available: http://www.sandia.gov/~tgkolda/TensorToolbox/.

[39] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin, "LIBLINEAR: a library for large linear classification," *Journal of Machine Learning Research*, pp. 1871-1874, Aug. 2008.