2018 IEEE International Conference on Image Processing

Depth Human Action Recognition Based on Convolution Neural Networks and Principle Component Analysis

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

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

Introduction

➢PCA technique 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.

➢In this paper, 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).

➢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.

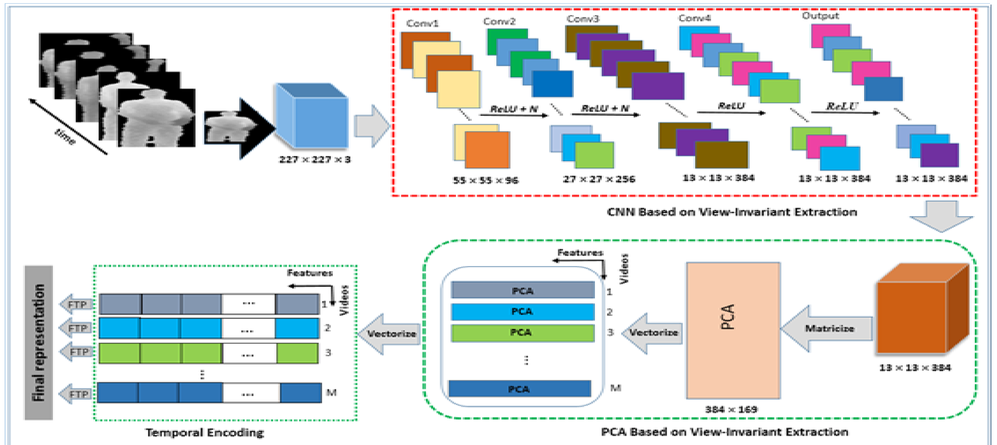
➢To deal with noisy data and temporal misalignment, we utilize the Fourier temporal pyramid (FTP) to encode temporal and obtain the final descriptors.

Related works

➢Depth-based methods: Depth-based human action recognition techniques can be divided into two main categories including holistic and local approaches. E.g. depth motion maps (DMM), histograms of oriented gradients (HOG), histogram of oriented 4D surface normal (HON4D), and histogram of oriented principal components (HOPC).

➢Deep learning methods: 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]. 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.

The proposed method



- The proposed model is based on HPM model [18] whose architecture is similar to AlexNet [32].
- The designed model exploits a pre-trained human pose structure 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.
- Applying the embedded PCA technique (EPCA) in [33] by solving the optimization problem:

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

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

- Obtaining the low-dimensional matrix

$$\mathbf{Y}_t^{(k)} = \mathbf{X}_t^{(k)} \mathbf{Q}_t^{(k)}.$$

- The PCA algorithm is applied on each matrix of CNN output 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 continue to apply PCA algorithm to further extract features on them.
- We employ the Fourier temporal pyramid (FTP) [13] on all elements of vectors to find the first q low frequency components.
- In the matching scheme, we perform one-vs-all strategy on the extracted feature vectors using linear support vector machines (SVM) [39].

Experiments

(1) Compared Methods:

(1) CCD [34]; (2) HON4D [5]; (3) SNV [9]; (4) CVP [36], (5) DVV [35]; (6) HOPC [14]; (7) HPM+TM [18].

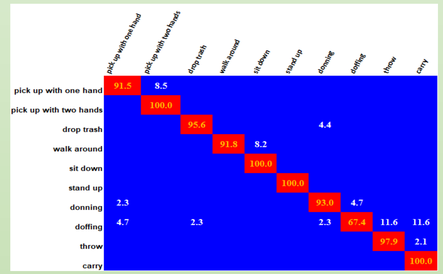
(2) Datasets: the UWA3D-II and the Northwestern-UCLA datasets

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**

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**



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

Conclusion

This work introduces 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.

Literature cited

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

[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

[32], Z. 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.

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

[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.

Acknowledgments

This research was financially supported by National Central University, Taiwan through the NCU International Student Scholarship.

For even more information, please contact Jia-Ching Wang at jiacwang@gmail.com or Manh-Quan Bui at bmanhquan2015@gmail.com