ICIC Express Letters
Volume **12**, Number **7**, July **2018**

ICIC International ©2018 ISSN 1881-803X
pp. 715–721

ACTION RECOGNITION WITH ROBUST PCA SEGMENTATION DESCRIPTORS

Qingpei Xia, Junyong Ye, Tongqing Wang and Yiqiang Li

Key Laboratory of Optoelectronic Technology and Systems of the Ministry of Education
Chongqing University
No. 174, Shazheng Street, Shapingba District, Chongqing 400044, P. R. China
{ xiaqingpei; ygyocr; ocr; liyiqiang }@cqu.edu.cn

Received November 2017; accepted February 2018

Abstract. *Deep Convolutional Networks (ConvNets) have demonstrated remarkable progress of learning discriminative representation from raw visual data. However, deep ConvNets remain unable to achieve significant advantage for video-based action recognition. In human action recognition, visual feature descriptors are of vital importance. In this paper, we propose a novel thought of getting the low-rank features and sparse features from the video. This ideal is based on the Robust Principal Component Analysis (RPCA), and obtained through sparse component and low-rank on the raw image pixels of each video. Then we use the two-stream architecture for our robust PCA segmentation descriptors with fine-tuning. We evaluate our novel thought using two widely used action recognition benchmarks (HMDB51 and UCF101) which demonstrates both performance power and the classification capability of our approach.*

Keywords: Action recognition, Robust principal component analysis, Deep convolutional networks

1. **Introduction.** Action recognition based on video is a challenging task, and it has received a significant amount of attention in the research community [1, 2, 3], owing to its implantation in a number of real-world applications, such as security and behavior analysis. Inspired by the remarkable successes of ConvNets, a state-of-the-art still image representation in many image understanding tasks, recent works have explored ConvNets for action recognition.

Many popular methods have been proposed. For instance, Simonyan and Zisserman [3] calculate optical flow to explicitly capture motion information. Fernando et al. [4] obtain a new representation, which captures the video-wide temporal dynamics of a video for action recognition. Tran et al. [5] train a deep 3-dimensional convolutional network to get the state-of-the-art performance. The long short term memory [6] has also been used to learn an effective representation of videos in unsupervised settings by using them in an encoder-decoder frame work. Wang et al. [7] propose a multi-level video representation by stacking the activations of motion features, atoms, and phrases. However, a problem still exists when extending deep learning methods from image to video: unlike images, videos have variable temporal durations. How to represent a video optimally remains a difficult problem in action recognition.

The main obstacle to achieve an ideal case in practical recognition is compacting representation of videos. In this work, we are trying to use a new powerful representation of videos in the context of human action recognition. Our way to improve that is to capture the complementary information on appearance from still frames and motion between frames. To implement movement and background segmentation in this regard, a novel solution of obtaining the movement of people and complex backgrounds by using RPCA

DOI: 10.24507/icicel.12.07.715

has been created. This approach provides a stable tool for data analysis, the input dataset is decomposed into a sum of low-rank and sparse components. Sparse components frames capture the movement of people, while low-rank captures the complex background images of the video sequence, and hence, they help to reduce computational complexity.

The main contributions of this paper can be summarized as follows. 1) Our first contribution is to introduce the notion of movement and background images (Figure 1), which summarizes, in a compressed format, the gist of a representation of video. 2) To the best of our knowledge, there is rarely previous work using hand-craft feature, at the level of the image pixels, for action recognition. We compute this hand-craft feature directly at the level of the image pixels instead of using an intermediate feature representation. 3) With the proposed deep framework and novel representation of videos, we perform experiments on two challenging action recognition dataset, and finally we achieve excellent performance 66.7% on HMDB51 [8] and 89.4% on UCF101 [9]. Experimental results show that our new powerful representation of videos significantly improves action recognition performance.

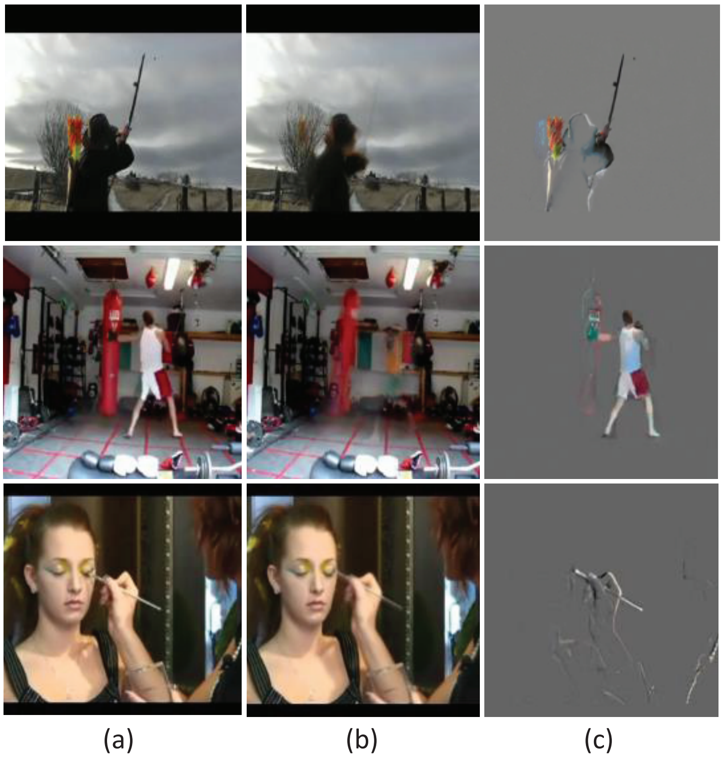

Figure 1. Movement and background segmentation images: (a) original video frames, (b) low-rank features (background segmentation), and (c) sparse features (movement segmentation)

2. **Video Representation.** In this section, the segmentation for movement of people and complex background from the videos at the level of the image pixels will be measured. Then, how to formulate these specific functions using RPCA will be recommended. Finally, we analyze the generalization capacity of the proposed sparse and low-rank images.

2.1. **RPCA for movement and background segmentation.** To mine human movement and background scene, we reshape the input video as a data matrix $D \in \mathcal{R}^{m \times n}$, and each video frame is a column vector of that matrix in a high dimensional ambient space. RPCA aims to recover a low-rank matrix L and its corresponding sparse component S

from the corrupted measurement D under a rather weak assumption, which can be solved by the following problem:

$$\min_{L,S} \|S\|_F, \text{ subject to } rank(L) \leq r, \quad D = L + S. \tag{1}$$

Here, $\|*\|_F$ is the Frobenius norm, and $r \ll \min(m, n)$ is the dimension of D. Mathematically, for an appropriate choice of parameter $\lambda > 0$, we have the following combinatorial optimization problem to solve:

$$\min_{L,S} rank(L) + \lambda\|S\|_0, \text{ subject to } D = L + S. \tag{2}$$

2.2. **Optimization method.** However, there is no efficient way to solve the optimization problem because this is a highly non-convex problem, while no efficient solution is known to track with it so far. Considering its convex relaxation instead, the solution using the matrix nuclear norm represented as $\|.\|^*$, employing the l_1-norm as a convex surrogate for l_0-norm is the best convex approximation [10, 11].

$$\min_{L,S} \|L_0\|^* + \lambda\|S\|_1, \text{ subject to } D = L + S. \tag{3}$$

When coming with constraint problem, we want to replace it via unconstrained objective function. The optimization problem can be solved via the method of augmented Lagrange multipliers. For the optimal Lagrange multiplier Y and the positive scalar μ, the Lagrangian function is given by

$$L(L, S, Y, \mu) = \|L_0\|^* + \lambda\|S\|_1 + < Y, D - L - S > + \frac{\mu}{2}|D - L - S|_F^2 \tag{4}$$

By using it, we can have the inexact approaches to minimizing for the matrix completion problem via iterative method [12]. The knowledge of salient regions and the distribution of outliers are important to drive the algorithm, when solving an approximated RPCA problem. In our method, we first extract the low-rank features and sparse components of each video by using RPCA decomposition.

2.3. **Generalization capacity.** A few examples of movement and background images are shown in Figure 1. Several observations can be made. Firstly, it is interesting to note that the sparse features tend to focus mainly on the acting objects, both humans and other objects such as Punching Bag in the "Boxing Punching Bag" action, or objects as a bow and arrows in the "Archery" action, since some actions are strongly associated with particular objects. Secondly, the obtained low-rank features will be transformed into corresponding low-rank images, which tends to be background pixels. The background scene can be evaluated the effect of the human action recognition without human, since some actions are strongly associated with particular objects. Finally, we observe that both sparse features and low-rank features are composed at the level of the image pixels.

3. **Action Recognition with Movement and Background.** In this section, we give detailed descriptions of performing action recognition with our new video representation, called movement and background feature descriptor, which shares the benefits of hand-crafted features map for video understanding. Specifically, we first introduce the basic architectures of ConvNets we used. Then, we show how to adapt the movement and background images for ConvNets. Next we study the good practices training on large datasets to extract multi-scale convolutional feature maps in learning two-stream ConvNets. Finally, we describe the testing method for our ConvNets.

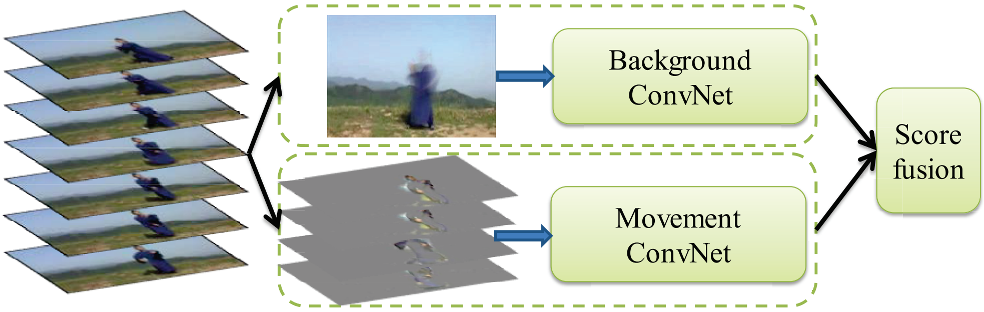

FIGURE 2. Two-stream ConvNets model

3.1. **Network architectures.** Figure 2 shows the outline of the two-stream ConvNets architecture. For our movement and background feature maps, two separate ConvNets are needed, namely movement net and background net. The movement part, in the form of motion across the frames, conveys the movement of the objects appearance in the video, which are trained on volumes of stacking frame images ($224 \times 224 \times 3N$, N is the number of stacking images). The background part, the low-rank of the video, carries information about scenes, whose input is a single image. In principle, any kind of ConvNets architecture can be adopted for our movement and background image features. In our implementation, we choose the ResNet-50 [13] instead of AlexNet, because this deep networks achieved 5.25% [13] top-5 error rates on the ImageNet ILSVRC-2012 validation set, respectively.

3.2. **Combining movement and background segmentation images.** Instinctively, we can use a single image for both movement stream and the background stream, which is simple but efficient. Inspired by [3], compared with single image as the input, they use image for the spatial stream and stacked optical flow fields for the temporal stream increased a lot. Therefore, we are also interested in exploring more modalities to enhance the discriminative power.

Movement Net. The sparse of the separation from each video date is the movement image, which represents the movement part that is content to the contextual information about the whole video. As shown in Figure 1, the moving people and target moving objects are the main composition of movement images. We would like to multiply movement images to capture the motion information. We can stack the consecutive movement frames to form one total input channels to represent the motion across a sequence of frames. In this way, the filters in the first layer of the movement stream are further modified by replicating the three RGB filter channels to a size of $3 \times L$ for operating over the movement image stacks, each of which has a stack of $L = 1, 5, 10$ frames.

Background Net. The background images are the low-rank of the separation from video, representing the same scene of the video-level. In a specific time point view, a single background image usually represents static appearance, and is the context-related for many previous and next frames even a whole video. The static appearance by itself is a useful clue, since some actions are strongly associated with scene. As shown in Figure 1, the background images describe the appearance of the scene, which may correspond to the place of motion occurrence. Thereby, we experiment with a single background image as input modality and investigate its performance in action recognition for background net.

4. **Experiments.** In this section, firstly, we present the details of datasets and their evaluation scheme. Then, we describe the details of our method. Finally, we give the experimental results.

4.1. **Datasets.** In order to test our features of movement and background, we conduct experiments on two public large datasets, namely HMDB51 [8] and UCF101 [9]. The HMDB51 dataset is composed of 6,766 video clips from 51 action categories. The UCF101 dataset contains 101 action classes and there are at least 100 video clips for each class. As UCF101 is larger than HMDB51, we use the UCF101 dataset to train two-stream ConvNets initially, and then transfer this learned model on the HMDB51 dataset.

4.2. **Implementation details.** It is challenged to train a deep ConvNets for action recognition as the available dataset is extremely small compared with the ImageNet dataset, and action from a video is more complex than object from a single image. We choose the training dataset of UCF101 split 1 for learning two-stream ConvNets as it is probably the largest public available dataset. We use Caffe to construct the deep neural network and use SGD to optimize it. The network weights are learnt using the mini-batch (set to 128) with momentum (set to 0.9).

For movement net, we first resize the frame to make the side as 256 × 256, and then randomly crop a 224 × 224 region from the frame. Specifically, we choose different stack frames of movement fields. We pre-train the network with the public available ResNet [13] and set initial learning rate to 0.01, decreasing to 10^{-3} after 20K iterations. It is then reduced to 10^{-4} after 40K iterations and training is stopped at 60K iterations.

For background net the learning rate is initially set to 10^{-3} and decreases to 10^{-4} after 10K iterations. It is then reduced to 10^{-5} after 20K iterations and training is stopped at 30K iterations.

4.3. **Evaluation of movement and background nets.** In this section, we evaluate performances of the proposed model in different input modalities on UCF101 dataset, and the experimental results are summarized in Table 1.

Table 1. Exploration of different input modalities for two-stream ConvNets on the UCF101 dataset (split 1)

Modality	From scratch (%)	Fine-tuning (%)
Movement Image ($L = 1$)	76.8%	86.5%
Movement Image ($L = 5$)	75.3%	85.3%
Movement Image ($L = 10$)	75.4%	85.7%
Background Image	72.8%	81.3%

We observe that the best recognition performance for movement images is 86.5%, and for background images is 81.3%. This result indicates that movement images and background images encode action information, which shares the benefits of hand-crafted features map. Although background maps do not count the action information, the performance appearance surprised us, which indicate that action recognition can be done just by analyzing motion of the background. This clearly shows the effect of a background sequence for understanding an action label. One result we should note that though stacking multiple ($L > 1$) displacement fields provides the network with long-term motion information in other papers , which is not more discriminative than the one frame (we set $L = 1$) in our result.

4.4. **Different consensus functions.** Each net may not get perfect accuracy, while combination of the two nets further improves the results. Here we evaluate several candidates: (1) max pooling, (2) average pooling, (3) variance reciprocal weight. We can get the action recognition probability from each net; the error distance $(x - \mu)$ can be seen as one minus probability. And variance is sample $\sigma^2 = \sum_{i=1}^{N}(x_i - \mu)^2/N$, and the weight is $1/\sigma^2$, (4) static weight (2/5 for background net, 3/5 for movement net).

Table 2. Exploration of different functions for movement and background nets on UCF101 dataset (split 1)

Modality	Two-Stream
Max	87.6%
Variance Reciprocal Weight	88.2%
Average	88.8%
Static Weight	89.1%

As shown in Table 2, the best result is obtained by static weight. As discussed before, we believe that both average and max are not enough to guarantee good convergence, while variance reciprocal weight prone to reduce error, the weight average has strong response as well as the more important information for action recognition to consensus function.

4.5. **Results.** After exploring the good practices for our two-stream, we are ready to build up our final action recognition method. Specifically the movement and background nets were trained on UCF101 [9], and HMDB51 [8]. As can be seen from Table 2, our movement and background nets outperform the deep architectures by the weighted average. The combination of the two nets further improves the results (in line with the single-split experiments above), and is comparable to the very recent state-of-the-art models.

The results are summarized in Table 3, where we compare our method with other deep learning methods. We anticipate that combining the proposed movement and background images will benefit the accuracies further. Some shallow representation is still essential ingredients of the state-of-the-art methods. The most prominent one is fusing the two-stream network at the convolutional layer instead of fusing before softmax layer. Finally, we conclude that with further representation, our movement segmentation allows for state-of-the-art accuracy in action recognition.

Table 3. Comparison of movement and background networks to the state-of-the-art methods

Modality	HMDB51	UCF101
HOG [14]	40.2%	72.4%
HOF [14]	48.9%	76.0%
MBH [14]	52.1%	80.8%
HOF+MBH [14]	54.7%	82.2%
C3D (3 nets) [5]	–	82.3%
LTC [6]	–	84.3%
Two-Stream ConvNet [3]	59.4%	88.0%
MoFAP [7]	61.7%	88.3%
Ours	**66.7%**	**89.4%**

5. **Conclusions.** We present robust PCA segmentation descriptors, a novel and powerful compact representation for videos, which can directly assign videos at the level of the single images. We extract the sparse component, represent movement features and low-rank feature, represent background images of each video by using RPCA decomposition. A visual inspection outlines the richness of movement and background descriptors in describing complex motion patterns as simple 2D images. As such, movement and background descriptors are directly competitive to existing CNN architectures allowing for action recognition learning. We conduct a systematic study and also introduce a consensus function to find the best for our two-stream. We show that movement and

background segmentation can share appearance and motion information simultaneously. In the future, we will investigate slow fusion strategies to combine background net and movement net at some convolutional layers instead of fusing before softmax layer.

Acknowledgment. This work was supported by the Fundamental Research Funds for the Central Universities of China under Grant 106112017CDJXY120001.

REFERENCES

[1] H. Jhuang, T. Serre, L. Wolf and T. Poggio, A biologically inspired system for action recognition, *Proc. of 2007 IEEE Int. Conf. on Comput. Vis.*, Rio de Janeiro, Brazil, pp.1-8, 2007.

[2] H. Wang and C. Schmid, Action recognition with improved trajectories, *Proc. of 2013 IEEE Int. Conf. on Comput. Vis. and Pattern Recognit.*, Portland, Oregon, pp.3551-3558, 2013.

[3] K. Simonyan and A. Zisserman, Two-stream convolutional networks for action recognition in videos, *Advan. in Neu. Infor. Pro. Sys.*, Montréal, Canada, pp.568-576, 2014.

[4] B. Fernando, E. Gavves and J. Oramas, Rank pooling for action recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.39, no.4, pp.773-787, 2017.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, Learning spatiotemporal features with 3D convolutional networks, *Proc. of 2015 IEEE Int. Conf. on Comput. Vis*, Santiago, Chile, pp.4489-4497, 2015.

[6] N. Srivastava, E. Mansimov and R. Salakhudinov, Unsupervised learning of video representations using LSTMs, *International Conference on Machine Learning*, Lille, France, pp.843-852, 2015.

[7] L. M. Wang, Y. Qiao and X. O. Tang, MoFAP: A multi-level representation for action recognition, *International Journal of Computer Vision*, vol.119, no.3, pp.254-271, 2016.

[8] H. Kuehne, H. Hildegard and G. Hueihan, HMDB: A large video database for human motion recognition, *Proc. of 2011 IEEE Int. Conf. on Comput. Vis.*, Barcelona, Spain, pp.2556-2563, 2011.

[9] K. Soomro, A. Zamir and M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, *arXiv*, 2012.

[10] J. Wright, A. Ganesh, S. Rao, Y. Peng and Y. Ma, Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization, *Advances in Neural Information Processing Systems*, British Columbia, Canada, pp.2080-2088, 2009.

[11] S. J. Huang, J. Y. Ye, T. Q. Wang, L. Jiang and Y. Li, Extracting refined low-rank features of robust PCA for human action recognition, *Arabian Journal for Science and Engineering*, vol.40, no.5, pp.1427-1441, 2015.

[12] Z. C. Lin, M. M. Chen and Y. Ma, The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices, *arXiv*, 2009.

[13] K. M. He, X. Y. Zhang, S. Q. Ren and J. Sun, Deep residual learning for image recognition, *Proc. of 2016 IEEE Int. Conf. on Comput. Vis. and Pattern Recognit.*, Las Vegas, USA, pp.770-778, 2016.

[14] H. Wang and C. Schmid, LEAR-INRIA submission for the THUMOS workshop, *PICCV Workshop on Action Recognition with a Large Number of Classes*, vol.2, no.7, pp.8-15, 2013.