REALISTIC HUMAN ACTION RECOGNITION: WHEN CNNS MEET LDS

Lei Zhang[1], *Yangyang Feng*[1], *Xuezhi Xiang*[1], *Xiantong Zhen*[2]

[1] College of Information and Communication Engineering, Harbin Engineering University, Harbin, China

[2] Department of Medical Biophysics, University of Western Ontario, London, ON, Canada

ABSTRACT

In this paper, we proposed new framework for human action representation, which leverages the strengths of convolutional neural networks (CNNs) and the linear dynamical system (LDS) to represent both spatial and temporal structures of actions in videos. We make two principal contributions: first, we incorporate image-trained CNNs to detect action clip concepts, which takes advantage of different levels of information by combining the two layers in CNNs trained from images; Second, we further propose adopting a linear dynamical system (LDS) to model the relationships between these clip concepts, which captures temporal structures of actions. We have applied the proposed method on two challenging realistic benchmark datasets, and our method achieves high performance up to 86.16% on the YouTube and 82.76% UCF50 datasets, which largely outperforms most of the state-of-the-art algorithms with more sophisticated techniques.

Index Terms— Deep learning, Image-trained CNNs, Linear dynamical system, Concept confidence

1. INTRODUCTION

Human action recognition on realistic scenarios has recently drawn great interest for video analysis, which poses great challenges for accurate recognition. The main cue of an action contains spatial and temporal structures, both of which should be modeled for action representation. [1]. In this paper, we propose a new framework to leverage the strengths of deep learning and linear dynamic system for representations of realistic actions.

Deep learning [2] has shown great effectiveness in many applications including mainly image and signal processing. In fact, the idea of deep learning is to stack the basic unit into different hierarchical structure, which aims to extracts hierarchical information. Although for different applications, the deep learning ideas are similar, there is a common sense for architecture selection, such as recurrent neural network (RNN) [3] in natural language processing, deep belief nets (DBN) in speech signal processing [4] and CNNs in image understanding [2, 5].

Given the overwhelming success in the image domain for visual analysis and recognition tasks, CNNs have been widely used as the most powerful tool for feature learning from still image. Thanks to ImageNet large scale visual recognition challenge (ILSVRC), which brings the appealing growth of CNNs from 2012, and well-known open models as Alexnet in 2012, GoogLeNet [6] and VGGNet [7] in 2014 are derived from it. Besides the image classification tasks on CNNs, descriptors learned by CNNs can also boost the performances of a broad range of visual tasks, including object detection in [8], action recognition in [9] and scene labeling in [10].

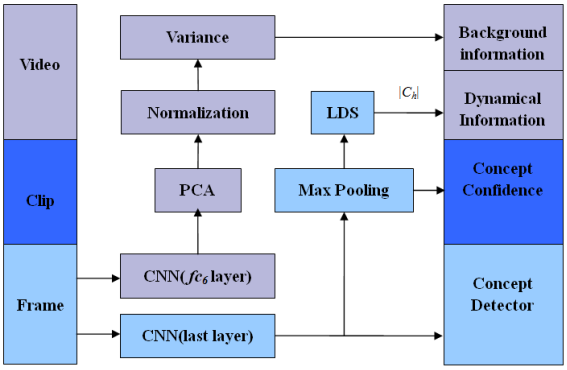

Fig. 1. The illustration of video representation on CNN (best viewed in color).

Recently, CNNs started to draw research interest in video analysis and event detection in the video domain. Generally speaking, there are two-way to apply CNNs in video processing. One is to extract frame level CNNs descriptors, as in [11, 12], accompanied by suitable pooling and encoding strategies, which can achieve state-of-the-art performance. The other way is to retrain CNNs by local cuboid representations, which is time-consuming and laborious task.

In this paper, we target at action recognition for realistic scenarios with actions in complex backgrounds and clutters. To capture the spatial appearance information of actions, we proposed adopt CNNs to extract features. To avoid the complicated re-training procedure, our work starts from CNNs learned by still images in ImageNet without fine-tuning on videos, and build a efficient framework to fuse the semantic meaning in video with dynamical temporal information as shown in Fig. 1. We adopt different colors to represent dif-

ferent levels in video representation. The lowest level is on frame, the middle level is on clips, and the highest level is to treat video as both clip sequences and frame sets, which is obtained from two different layers of CNNs.

To capture the temporal structure information of actions, we propose using a linear dynamical system (LDS) [13] to model the relationships between concept clips obtained from the CNNs. Linear dynamical systems are dynamical systems whose evaluation functions are linear and can be used to understand the qualitative behavior of general dynamical systems. We introduce the LDS to capture the temporal dynamical structure in a video sequence.

We make two principal contributions: first, we incorporate image-trained CNNs to detect action clip concepts, which takes advantage of different levels of information by combining the two layers in CNNs trained from images; Second, we further propose adopting a linear dynamical system (LDS) to model the relationships between these clip concepts, which captures complex dynamical temporal structures of actions in a video sequence.

We apply the proposed method to challenging realistic human action datasets. The experimental results have demonstrated that the proposed framework is effective to capture the spatial and temporal features of actions to achieve efficient and accurate human action recognition.

2. SPATIAL REPRESENTATION WITH CNNS

In order to leverage the great strength of the deep learning techniques, we propose using the convolutional neural networks (CNNs) to capture the spatial appearance of actions in video frames. In stead of training new CNNs on video sequences, which would be extremely expensive to compute and even feasible for large scale human action datasets, we propose adopting the image-trained CNNs, a set of filters, to extract spatial appearance features from video frames.

The VGG network is currently the most preferred choice in the community when extracting CNNs features from images. In particular, we adopt the VGGNet with 21 layers composed of convolution, full-connect and max pooling layers.

Instead of directly applying VLAD or BoW on image representation of CNNs, we fully utilize the CNNs both for concept detector and for the background variation reflection. We adopt the last layer in CNNs in Fig. 1 to reflect the semantic concept of each frame in video, and the 6^{th} full connection (fc_6) layer output to represent the global background information.

2.1. Descriptor on whole videos

Inspired by [11], which treats the filters as latent concept classifiers, we believe that the filter output in convolution layer at least can reflect the global content information of each frame. This is the main reason for selecting the fc_6 layer output to extract background information.

The dimension of fc_6 layer output is 4096, and each dimension corresponds to the response of filter with $6 \times 6 \times 256$ cuboid size. This vector with 4096 dimension can fully reflect the content of frame from different aspects, such as color, shape and so on. By PCA dimensional reduction and normalization in Fig. 1, the variance on the output of fc_6 layer is computed as background information, which gives the divergence of the whole video and is helpful to distinguish the actions with complex backgrounds.

Combined with a linear dynamical system, we can capture the full information of actions in a video, which includes the spatial appearances, i.e., the layout of human body in clips, and temporal relations among clip sequences.

2.2. Concept confidence on clips

We start with extracting the frame level CNNs descriptors on the model shared by VGG group. For the last layer in CNNs, we can treat it as a classifier or a concept detector which can give the likelihood details of each object among 1000 categories in ImageNet, which is shown in Fig. 1. By max-pooling, we can represent concept confidence on clip level in algorithm 1.

Algorithm 1 Concept confidence $L_{V_n^c}$

Input: Input videos

Output: Video representation by concept confidence

1. Dividing the n^{th} video into several clips, and a video is treated as the clip sequence with latent temporal structure.
$V_n = \{V_n^1, ...V_n^c, ..., V_n^C\}$

2. For each clip, it is treated as a frame set without temporal information.
$V_n^c = \{f_n^c(1), ...f_n^c(t), ..., f_n^c(T)\}$

3. For one frame $f_n^c(t)$, compute the likelihood $l_{f_n^c(t)}$ by the concept detector, which is the score vector, output of the last layer in a CNN.

4. For all $l_{f_n^c(t)}$ in V_n^c, compute the concept confidence as $L_{V_n^c}(i) = \max_{t \in \{1...T\}} l_{f_n^c(t)}(i)$, where i corresponds the dimension index in $l_{f_n^c(t)}$.

3. TEMPORAL MODELING WITH LDS

The temporal structure contains informative and discriminative information, which can be explored to achieve effective representations of human actions. Although concept confidence can provide reasonable estimation of the probability of the content in each clip, how best to integrate this information with temporal structure remains a challenge. We introduce a linear dynamical system (LDS) to model the temporal relations among clips.

We represent a video V as a sequence of n clips, with no overlap. As described above, each clip descriptor by concept confidence can express the content from semantic aspect, and the clip descriptor sequence could be treated independently as an individual time series. In order to exploit the interactions across clips, a linear dynamical system (LDS) is adopted here with an additional advantage of dimensionality reduction.

LDS can be defined as:

$$z_{n+1} = Az_n + w_{n+1} \tag{1}$$

$$x_n = Cz_n + \varepsilon_n \tag{2}$$

where z_n is the state at time n and x_n is the output of system at the same time.

This model assumes the observed output sequences are generated from a series of hidden variables with a linear projection matrix C, and the hidden variables evolve over time with a linear transition matrix A.

As observed in [13], if we decompose matrix A as follows, the complex eigen values can represent some properties of the signal such as the frequency and phase.

$$A = U\Lambda U^* \tag{3}$$

where $UU^* = I$ contains the eigenvectors of A and Λ is a diagonal matrix of eigenvalues of A. Furthermore, in order to obtain the same observation sequences from Λ as the transition matrix, we need to compensate the output matrix C as:

$$C_h = CU \tag{4}$$

It has already been shown that C_h can discriminatively represent original videos by exploiting the temporal structure between clips. By dropping off the phase information, which is only the delay in time domain, only the magnitude of C_h is considered in our method, corresponding to the dynamical information in Fig. 1.

4. EXPERIMENTS AND RESULTS

We apply the proposed method to challenging realistic human action recognition datasets. The proposed method achieves high performance and largely outperforms the state-of-the-art algorithms which use sophisticated techniques. The significantly improved performance has demonstrated the effectiveness of the proposed method for human action recognition.

4.1. Datasets

YouTube action [14]: This dataset contains 1168 sequences of 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horse back riding,soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. This dataset is challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background and illumination conditions.

UCF50 [15]: This dataset contains 6685 sequences of 50 action categories, consisting of realistic videos taken from YouTube. Comparing to the Youtube action dataset above, object scale, viewpoint, cluttered background, illumination conditions are also changed among different categories.

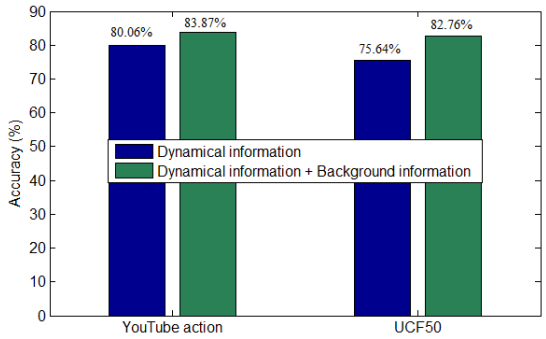

Fig. 2. Performance of video level descriptor (dynamical information + background information in Fig. 1).

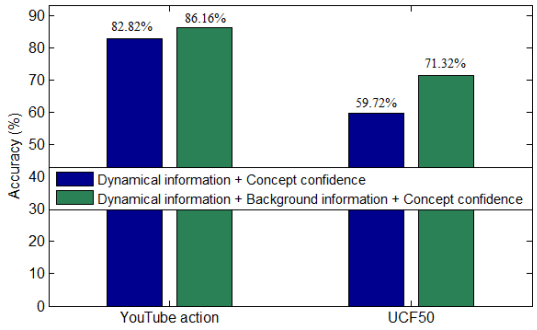

Fig. 3. Effect with additional clip level descriptor (concept confidence in Fig. 1).

4.2. Settings

For the hidden state number h of LDS in following experiments, we set it to be 8 in Eq. (5), where s_* is the singular value of matrix $|C_h|$ and in practice. We keep 90% energy to determine h. In addition, each video is divided into 10 clips, such that the clip level representation for each video with different durations can have the same size. For action recognition, a linear SVM [16] is applied for human action recognition.

$$h \leftarrow \arg_h \frac{\sum_{j=1}^{h} s_j^2}{\sum_{i=1}^{m} s_i^2} \tag{5}$$

We conduct extensive experimental comparison with state-of-the-art algorithms on these two datasets to comprehensively

Algorithm	Accuracy
Dynamical information + background information + concept confidence	**86.16%**
Hierarchical feature on ISA + BoF + Chi-square kernel [17]	75.80%
Dense trajectory + HOG +HOF + MBH + BoF [18]	84.10%
SIFT trajectory + HOG +HOF + MBH + BoF [18]	73.20%
KLT trajectory + HOG +HOF + MBH + BoF [18]	79.50%
Dense cuboids + HOG + HOF + MBH +BoF [18]	81.40%
Static + motion feature [14]	71.20%
Relative motion descriptor (RMD) + Modes [19]	81.70%
Dense trajectory + BoF [20]	84.20%

Table 1. Performance comparison of accuracy with state-of-the-art approaches on the YouTube dataset

Algorithm	Accuracy
Dynamical information + background information	**82.76%**
Relative motion descriptor (RMD) + Modes [19]	81.80%
Lagrangian particle trajectories [21]	81.03%
Dense trajectory + HOG +HOF + MBH + BoF [18]	84.50%
SIFT trajectory + HOG +HOF + MBH + BoF [18]	71.80%
KLT trajectory + HOG +HOF + MBH + BoF [18]	78.10%
Dense cuboids + HOG + HOF + MBH +BoF [18]	80.20%
Motion feature [15]	76.90%
GIST3D + STIP [22]	73.70%
Orientation-based descriptor+ Gabor STIP [23]	72.90%
Motion interchange patterns [24]	72.68%

Table 2. Performance comparison of accuracy with state-of-the-art approaches on the UCF50 dataset

investigate the proposed method.

4.3. Performance on the video level descriptor

Fig. 2 gives the performance of video level descriptors, including individual dynamical information with and without background information.

It can be seen that for individual dynamical information, the performance achieves 80.06% on YouTube action dataset while it is 75.64% for UCF50 dataset, proving the effectiveness of dynamical information. The main reason lies in that dynamical information aims to dig the temporal structure, which is important in action representation.

With additional background information, the performance can be further enhanced by the proposed method. Compared with the algorithms that neglects background information, the proposed method improves the performance by a large margin of 3.81% on the YouTube dataset and of 7.12% on the UCF50 dataset. This finding demonstrates the effectiveness of background information, which can capture the divergence for the whole video.

4.4. Fusion with the clip level descriptor

Fig. 3 shows the behavior of additional clip level descriptor. on the YouTube dataset, additional concept confidence outperforms the system without concept confidence by 2.78% and 2.29% on the first two bars in Fig. 2 and Fig. 3 respectively. However, on the UCF50 dataset, additional concept confidence does not further improve the performance.

Note that dynamical information is extracted on concept confidence by LDS, then concept confidence can be viewed as static descriptor while dynamical information reflects the temporal structure information. The behavior in Fig. 3 certifies that static descriptor is helpful when itself with great discriminative ability, such as YouTube dataset with a smaller category condition. However, with the increase of categories, the static descriptors tend to be less discriminative due to the sharing background information among different actions. In this circumstance, the merit of dynamical information is manifested.

4.5. Comparison to state-of-the-art performance

To show the advantages of proposed approach for human action recognition, we have also conducted extensive comparison with state-of-the-art results in Table 1 and Table 2 on both YouTube and UCF50 datasets respectively. It is firstly observed that on the YouTube dataset, the proposed method beats the most of the previous methods with more sophisticated techniques and computationally more expensive stratifies, even including the dense trajectory with the BoF framework. Secondly, on the UCF50 dataset, the proposed method outperforms most of methods and is even competitive with the dense trajectories plus the HOG, HOF and MBH descriptors, which however is computationally more expensive than the proposed method.

The largely improved performance of the proposed method over the state-of-art algorithms has clear shown the effectiveness of the proposed methods for human action recognition on realistic datasets, which also indicates its great potential to be used in practical applications.

5. CONCLUSION

In this paper, we have presented a new framework for human action recognition on realistic datasets. To achieve informative and effective representations of actions, we proposed using convolutional neural networks (CNNs) to capture the spatial appearances and adopting a linear dynamical system (LDS) to model the temporal structures. We apply the proposed method to realistic human action recognition on the YouTube and UCF50 datasets, which are realistic and extremely challenging. The proposed method achieves high recognition accuracy and outperforms most of the state-of-the-art algorithms, which shows the effectiveness of the proposed method for human action recognition in realistic scenarios.

References

[1] Xiantong Zhen, Ling Shao, Dacheng Tao, and Xuelong Li, "Embedding motion and structure features for action recognition," *IEEE TCSVT*.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[3] Lipton Zachary C., berkowitz John, and Elkan Charles, "A critical review of recurrent neural networks for sequence learning," in *arXiv preprint arXiv: 1506.00019*, 2015.

[4] Nicolas Boulanger-Lewandowski, Yoshua bengio, and Pascal Vincent, "Modeling tempora dependencies in high-dimensional sequences: application to polyphonic music generation and transcription," in *ICML*, 2012.

[5] Zifeng Wu, Yongzhen Huang, and Liang Wang, "Learning representative deep features for image set analysis," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1960–1968, 2015.

[6] Zhirong Wu, Yinda Zhang, Fisher Yu, and Jianxiong Xiao, "A gpu implementation of googlenet," Tech. Rep., Technical report, Princeton University, 2014. 6.

[7] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, 2013.

[9] Karen Simonyan and Andrew Zisserman, "Two-stream covolutional networks for action recognition in videos," 2014.

[10] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Learning hierarchical features for scene labeling," *TPAMI*, vol. 35, no. 8, pp. 1915–1929, 2013.

[11] Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann, "A dsicriminative cnn video representation for event detection," in *CVPR*, 2015.

[12] S. Zha, F.Luisier, W. Andrew, N. Srivastava, and R. Salakhutdinov, "Exploiting image-trained cnn architectures for unconstrained video classification," in *arXiv:1503.04144v2*, 2015.

[13] Lei Li, B Aditya Prakash, and Christos Faloutsos, "Parsimonious linear fingerprinting for time series," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 385–396, 2010.

[14] Jingen Liu, Jiebo Luo, and Mubarak Shah, "Recogning realistic actions from videos "in the wild"," in *CVPR*, 2009.

[15] Kishore K. Reddy and Mubarak Shah, "Recognizing 50 human action categories of web videos," in *Machine Vision and Applications Journal (MVAP)*, 2012.

[16] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," *TIST*, vol. 2, no. 3, pp. 27, 2011.

[17] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng, "Learning hierachical invariant spatio-temporal feature for action recognition with independent subspace analysis," in *CVPR*, 2011.

[18] Heng Wang, Alexander Klaser, and Cordelia Schmid, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, vol. 103, no. 1, pp. 60–79, 2013.

[19] Olusegun Oshin, Andrew Gilbert, and Richard Bowden, "Capturing relative motion and finding modes for action recognition in the wild," *Int. Journal Computer Vision and Image Understanding, CVIU*, vol. 125, pp. 155–171, 2014.

[20] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng lin Liu, "Action recognition by dense trajectories," in *CVPR*, 2011.

[21] Sinisa Todorovic, "Human actions as stochastic kronecker graphs," in *ECCV*, 2012.

[22] Solmaz B, Assari SM, and Shah M, "Classifying web videos using a global video descriptor," *Machine Vision and Application*, vol. 24, no. 7, pp. 1473–1485, 2012.

[23] Everts I, van Gemert J, and Gevers T, "Evaluation of color stips for human action recognition," in *CVPR*, 2013.

[24] Kliper-Gross O, Gurovich Y, Hassner T, and Wolf L, "Motion interchange patterns for action recognition in unconstrained videos," in *ECCV*, 2012.