3D CONVOLUTIONAL NEURAL NETWORK WITH MULTI-MODEL FRAMEWORK FOR ACTION RECOGNITION

Longlong Jing[1], Yuancheng Ye[1], Xiaodong Yang[3], Yingli Tian[1,2]

[1]The Graduate Center, [2]The City College, City University of New York, NY, USA,
[3]NVIDIA Research, Santa Clara, CA, USA
{ljing,yye}@gradcenter.cuny.edu, xiaodongy@nvidia.com, ytian@ccny.cuny.edu

ABSTRACT

In this paper, we propose an efficient and effective action recognition framework by combining multiple feature models from dynamic image, optical flow and raw frame, with 3D convolutional neural network (CNN). Dynamic image preserves the long-term temporal information, while optical flow captures short-term temporal information, and raw frame represents the appearance information. Experiments demonstrate that dynamic image provides complementary information to raw frame feature and optical flow feature. Furthermore, with the approximate rank pooling, the computation of dynamic images is about 360 times faster than optical flow, and the dynamic image requires far less memory than optical flow and raw frame.

Index Terms— Action Recognition, 3D Convolutional Neural Network, Video Classification

1. INTRODUCTION

Since Krizhevsky *et al.* [1] won the first prize in ImageNet 2012 competition, CNNs have been widely applied in many computer vision fields such as image classification, object recognition, and segmentation, etc. CNN shows great advantages over traditional hand-designed features in these fields. Helped with CNN, powerful GPU, and the available large datasets, these fields are developing quickly.

With more videos available on the Internet, researchers begin to apply CNN in videos such as action recognition, video description generation, and action localization. Unlike the image-based applications which depend only on spatial information, the action recognition needs both the spatial and temporal information to identify actions in videos. Varol et al. [2] treated actions as 3D objects and identified them by capturing both the appearance information from every frame and the appearance evolution between frames.

Researchers have proposed different types of networks to extract appearance and temporal information for action recognition. Simonyan et al. [3] incorporated two CNNs to learn appearance information from raw RGB frames and learn temporal information from optical flow. Donahue et al. [4] employed VGG [5] to extract appearance information as a fixed length vector from continuous frames which was then used to train Recurrent Neural Network (RNN) to learn the temporal information between frames. Ji et al. [6] took 3D CNN to learn both the appearance and temporal information from multiple consecutive frames which proved to have a good performance. The networks of [2, 7] follow a similar structure as [6] but with different lengths of input clips.

The models developed for image classification such as [1, 5, 8, 9], which are good at extracting appearance information from images, can be used to extract appearance information of frames for action recognition. However, there is not a good way to capture the long-term temporal information of videos. The existing methods can only handle certain length of frames to extract short-term temporal information confined within a video clip. Varol et al. proved the importance of long-term temporal information for actions that lasting for longer time [2]. Following [2], short-term temporal information extracted from short video clip is not sufficient to identify some actions. Therefore more efficient methods are needed to extract long-term temporal information from videos.

One straightforward idea of extracting long-term temporal information is to feed networks with long frame sequences. However, with more video frames feed into the network, the network would be more complicated and easily overfitting. This usually requires massive data to train the network. Among all the networks for action recognition, Varol et al. claimed that their network can capture long-term temporal information [2]. However, the difference between the method in [2] and others is that they used a longer input sequence (100 frames). Even though, a 100-frame video clip is much longer than the input of other methods (usually under 16 frames), it is still too short for actions like TaiChi.

Another way is to compress the evolution of appearance in a video into shorter sequences such as 16 frames. With this way, the CNN models of papers [2, 6, 7] can be employed. Fernando et al. proposed to use ranking machine to capture video-wide temporal information for action recognition [10], which was expanded further by Bilen et al. to dynamic image [11]. Bilen et al. replaced the rank pooling with approxi-

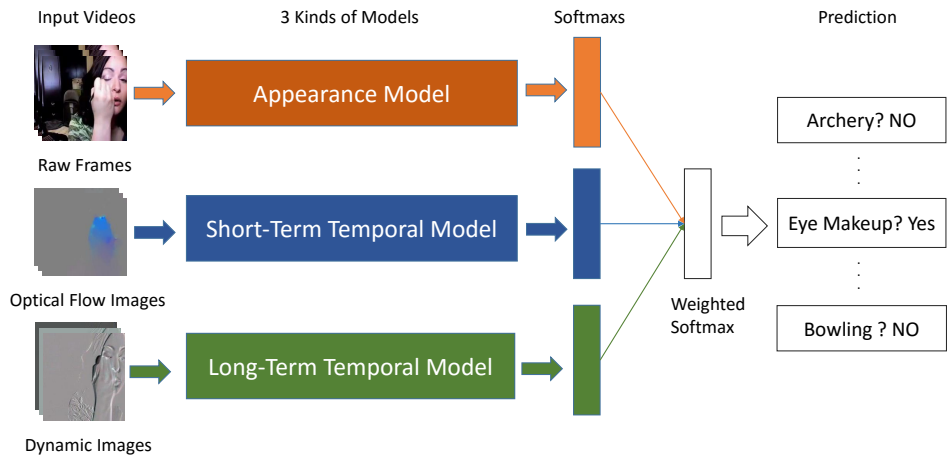

Fig. 1. The pipeline of the proposed framework. First, three models are trained to extract different kinds of information from videos. RGB Network is for extracting appearance information from raw RGB frames, Optical Flow Network is for capturing short-term temporal information from optical flow, and Dynamic Image Network is for representing long-term temporal information from the dynamic image. Finally, the outputs of the softmax of the three models are weighted to a final softmax score.

mate rank pooling to obtain dynamic image which reserves the long-term temporal information [11]. With this model, we can compress the temporal information span in a video into a short sequence, and the network can learn it from these dynamic images.

In this work, approximate rank pooling is applied in every video to generate 16 dynamic images. In addition to dynamic images, short-term temporal information is extracted from optical flow, and appearance information is extracted from raw RGB frames. We explore different fusions of the three models. Our multi-model framework achieves 88.6% in UCF101 and 57.9% in HMDB51 datasets.

2. PROPOSED FRAMEWORK

The pipeline of our framework is illustrated in Fig. 1. The three networks in Fig. 1 are corresponding to RGB Network, Optical Flow Network, and Dynamic Image Network. RGB Network is for extracting appearance information from raw RGB frames, Optical Flow Network is for extracting short-term temporal information from optical flow, and Dynamic Image Network is for capturing long-term temporal information from the dynamic images generated from the video. For every video, the softmax outputs of the three models are weighted to form the final softmax score.

During training the RGB Network, raw RGB frames of each video are divided into 16-frame clips, and each of them is fed to a 3D CNN which then serves as appearance feature extractor. For Optical Flow Network, the optical flow for each video is computed, then the optical flow for every video is divided into 16-frame clips without overlap between consecutive clips. These optical flow clips are fed to another 3D CNN to train Optical Flow Network which serves as short-term temporal information extractor. For Dynamic Image Network, each video is divided into 16 clips with equal length, and each of them is compressed into one dynamic image by approximate rank pooling. Then these dynamic images are fed into a 3D CNN to train the Dynamic Image Network.

2.1. DYNAMIC IMAGE GENERATION

The dynamic image generation is based on the idea of [10]. This paper proposed to employ rank pooling to capture video-wide temporal information for action recognition. They take ranking machine to learn functions that can order the frame temporally, and use the parameters learned by ranking machine to represent the evolution of the appearance within the video.

Inspired by Fernando et al. [10], Bilen et al. optimized the computation process and replaced the ranking machine with approximate rank pooling [11]. With the approximate rank pooling, the dynamic images are obtained by directly applying rank pooling on the raw image pixels of a video. The parameters of the frames only depend on the relative position of the frame in the video sequence, and can be pre-computed, which speeds it orders of magnitude.

The process of generating dynamic images by approximate rank pooing is indicate by Eq. (1).

$$\hat{\rho}(I_1, \ldots, I_T; \psi) = \sum_{t=1}^{T} \alpha_t \psi(I_t). \quad (1)$$

$$\alpha_t = 2(T - t + 1) - (T + 1)(H_T - H_{t-1}), \quad (2)$$

where $\psi(I_t)$ is pixels of the t-th frame, T is the length of the original video, $H_t = \sum_{i=1}^{t} 1/t$ is the t-th Harmonic number and $H_0 = 0$. The dynamic image is actually linear combinations of every frame and the parameters α_t for each frame can be obtained from Eq. (2).

In experiments, every video is divided into 16 segmentations with equal length and without overlap between consecutive segmentations, then we directly apply approximate rank pooling in each clip to generate dynamic images. From Eq. (1), the computation of dynamic image is linear combinations of the raw pixels in the frame, and it is very fast compared to optical flow computation.

2.2. 3D CONVOLUTIONAL NEURAL NETWORK

Ji et al. are the first to apply 3D CNN model for action recognition and achieve promising performance [6]. Compared to 2D CNN, the extra dimension in 3D CNN is temporal dimension which makes it possible to extract both spatial and temporal information from multiple frames. Among the 3D networks for action recognition, the network C3D proposed by Tran et al. [7] is a deeper network and achieves better performance, therefore we choose C3D in our framework. C3D includes 5 convolution layers, 2 fully convolution layers and one softmax layer. The size of the kernel in C3D is 3*3*3 which makes it more suitable for extracting spatial-temporal information than 2D CNN. In order to keep the temporal information in the first convolution layer, the size of the first max pooling is 2*2*1. Except the max pooling in the first layer, all the max pooling in later layers have a size of 2*2*2.

In [7], the input of C3D is consecutive 16 raw RGB frames, and C3D can capture the temporal information and appearance information from these input images. However, for actions that last for a long time, 16 consecutive images are not sufficient to represent the action. After compressing the temporal information into multiple dynamic images, C3D can learn the long-term temporal information from these dynamic images. When this kind of long-term temporal information is combined with short-term temporal information and appearance information, the multi-model framework can obtain a high performance.

2.3. Multiple Feature Models

Three networks are trained to model different types of features: RGB Network, Optical Flow Network, and Dynamic Image Network. After training finish, the three networks serve as feature extractor for every video. RGB Network which trained with raw RGB frames focuses on appearance information, Optical Flow Network which trained with optical flow captures short-term temporal information, and Dynamic Image-Network which trained with dynamic image learns long-term temporal information. The details of the network are as follows.

RGB Network. Raw RGB frames of every video are divided into 16-frame long clips without overlap between consecutive clips. Then these clips are fed into C3D to train RGB Network. After the training is finished, the activations of the softmax layer are extracted for every video clip. All the activations of softmax of the same video are averaged to form a final softmax score, which represents the appearance information of the short video clips.

Dynamic Image Network. Each video is divided into 16 segmentations with equal length and without overlap between consecutive segmentations, then approximate rank pooling is applied in each segmentation to generate dynamic images. After approximate rank pooing, UCF101 dataset is compressed into 13320 16-frame clips, and HMDB51 dataset into 6766 16-frame clips. The dynamic image is used to train Dynamic Image Network. After training finished, these dynamic images are passed to Dynamic Image Network, and the activations of the softmax layer are extracted. The feature extracted by Dynamic Image Network captures long-term temporal information span in a whole video.

Optical Flow Network. Varol et al. [2] showed that optical flow computed by [12] has the best performance among several methods [12, 13, 14]. We employ the method developed in [12] to compute optical flow for UCF101 and HMDB51 datasets. The optical flow is used to train Optical Flow Network. After the training is finished, the activations of the softmax layer are extracted for every video clips. All the activations of softmax of the same video are averaged to form a final softmax score. The features extracted by Optical Flow Network from optical flow model short-term temporal information within a clip.

2.4. Fusion of Multiple Feature Models

After training all the models, three kinds of features are obtained for every video, RGB feature, Dynamic Image feature, and Optical Flow feature. When fusing these models, the softmax of all the models are weighted to form the final softmax score similar to [3, 15, 16, 17]. For example, when fusing the RGB Network and Optical Flow Network, the softmax of these two models are averaged to form the final classification score. All the fusions among different feature models follow the same procedure.

3. EXPERIMENTS

3.1. DATASETS

We evaluate the performance of our proposed framework on UCF101 [18] and HMDB51 [19] datasets. UCF101 consists of 101 action categories, 13320 video clips. HMDB51 includes 51 action classes, 6766 video clips extracted from a variety of sources ranging from digitized movies to YouTube. All the experiments are conducted on the first split of these two datasets.

Table 1. The accuracies of different modalities. 'DI' indicates dynamic image, and 'OF' is optical flow. The accuracy of the single model such as Dynamic Image is the mean precision of classification. For the fusion of models, the accuracy comes from the weighted softmax score of different models.

Input	UCF101	HMDB51
RGB with 3D CNN[7]	82.5	50
OF with 3D CNN	78.2	48.9
Multilayer RGB with 3D CNN[20]	85.4	53.1
Multilayer OF with 3D CNN	82.5	53.0
DI with 2D CNN[11]	70.9	35.8
DI with 3D CNN	78.4	46.8
RGB + DI with 2D CNN[11]	76.9	42.8
RGB + DI with 3D CNN	85.8	53.6
RGB + OF with 3D CNN	87.6	56
RGB + OF + DI with 3D CNN	88.6	57.9

3.2. LEARNING PROCESS

Three kinds of networks, RGB Network, Dynamic Image Network, and Optical Flow Network are trained with the same structure as C3D. However, these two datasets are too small to train such scale of networks. In order to avoid overfitting, we train the network based on the model pre-trained in Sport-1M which has 1 million videos in 487 categories. The initial learning rate is 10^{-4} and decreased 90% every 20000 iterations. The optimization is done after 60000 iterations.

3.3. RECOGNITION ACCURACY

The results are reported in Table. 1. C3D with dynamic images achieves 78.4% in UCF101 which is 7.5% higher than dynamic image with 2D CNN [11]. When combining the feature of dynamic images and raw RGB frames, our multi-model achieves 85.8% which is 8.9% higher than dynamic images and raw RGB frames with 2D CNN. These two facts demonstrate the advantages of 3D CNN over 2D CNN for action recognition. The fusion of dynamic image feature and RGB feature achieves 85.8% which demonstrates that dynamic image feature is complementary to RGB feature. When we fuse the three models, the multi-model reaches 88.6% in UCF101 and 57.9% HMDB51, which demonstrates that dynamic image feature is complementary to optical flow feature and appearance feature.

3.4. COMPUTATION SPEED

We compare the computation speeds for dynamic image and optical flow. We randomly select 10 videos containing 996 frames from HMDB51 dataset. The computing speed for dynamic image and optical flow are reported in Table. 2. It shows that dynamic image is about 360 times faster than optical flow. From Eq. (1), the computation cost for dynamic image depends on the length of the clip and the number of generated dynamic images. No matter how many dynamic images generated for a video, the increase of the time cost only depends on the number of sum operation. However, the number of the sum operation is far less than that of multiplications. So the computation cost of dynamic image for a video is nearly constant.

Table 2. The memory and speed comparison of dynamic image and optical flow. The Dynamic Image 16 in the first row means generating 16 dynamic images for every video, while Dynamic Image 32 in the second row means generating 32 dynamic images for every video. The computation of dynamic images is hundred-time faster than optical flow and needs far less memory than that of raw frame and optical flow .

Generated Data	Time	Memory
Dynamic Image 16	0.348s/Video	16 frame/Video
Dynamic Image 32	0.382s/Video	32 frame/Video
Optical Flow	140s/Video	99 frame/Video

3.5. MEMORY

With dynamic image, we can compress the appearance evolution span in a video into multiple images. The number of dynamic images generated for every video can be adjusted according to the length of the video. Compared to RGB frames and optical flow, dynamic image needs less memory. In our experiments, every video is compressed into 16 dynamic image, which requires less memory than optical flow. Accomplished with the speed advantage, this is a very promising to be used for real-time applications.

4. CONCLUSION

We have proposed a framework to fuse the short-term temporal information, long-term temporal information and appearance information for action recognition. The ranking machine can preserve the evolution of appearance into multiple dynamic images while the optical flow can capture the local temporal information. Experimental results demonstrate that dynamic image feature is complementary to RGB feature and optical flow feature, and 3D CNN is more suitable for action recognition than 2D CNN. The performance of action recognition is significantly improved by fusing all the three feature models. Furthermore, the dynamic image is hundred-time faster and requires less memory than optical flow. For real-time applications, the dynamic image is a promising alternative to optical flow without decreasing much of the performance.

5. ACKNOWLEDGMENT

This work was supported in part by NSF grants EFRI-1137172, IIP-1343402, and IIS-1400802.

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.

[2] Gül Varol, Ivan Laptev, and Cordelia Schmid, "Long-term temporal convolutions for action recognition," *arXiv:1604.04494*, 2016.

[3] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 568–576. Curran Associates, Inc., 2014.

[4] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015.

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

[6] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[7] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[8] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars, "Modeling video evolution for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5378–5387.

[11] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[12] Thomas Brox, Andr"ës Bruhn, Nils Papenberg, and Joachim Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European conference on computer vision*. Springer, 2004, pp. 25–36.

[13] Vadim Kantorov and Ivan Laptev, "Efficient feature extraction, encoding and classification for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2593–2600.

[14] Gunnar Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.

[15] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: towards good practices for deep action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 20–36.

[16] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1961–1970.

[17] Yemin Shi, YongHong Tian, Yaowei Wang, and Tiejun Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *CoRR*, vol. abs/1609.03056, 2016.

[18] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[19] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre, "Hmdb51: A large video database for human motion recognition," in *High Performance Computing in Science and Engineering 12*, pp. 571–582. Springer, 2013.

[20] X. Yang, P. Molchanov, and J. Kautz, "Multilayer and multimodal fusion of deep neural networks for video classification," in *ACM Multimedia*, 2016, pp. 978–987.