Extended Temporal Convolutions for Human Action Recognition in Videos

Khwaja Monib Sediqi[1], Thai Leang Sung[1], Hyo Jong Lee[1,2]
[1]*Division of Computer Science and Engineering*
[1,2]*Center for Advanced Image and Information Technology*
[1,2]*Chonbuk National University*
Jeonju, South Korea
monib.korea@gmail.com, thaileang@jbnu.com, hlee@chonbuk.ac.kr

***Abstract*—In this paper we study the effect of extended temporal convolutions for human action recognition in videos. Our motivation emerges from empirical observation that a typical human action lasts several seconds constructing important spatial and temporal information. Preserving this information while training deep models help to improve performance significantly. Our empirical study leads to the design of a novel network architecture with extended temporal convolution for action recognition. Our novel method achieves comparable result to that of the state of the art. In this work, we also compare training a network for action recognition using RGB data and optical flow. Optical flow eases learning motion for action recognition and adds substantial improvement to the deep network model. Finally, we present an extended temporal convolution model fine-tuned on UCF-101 dataset.**

Keywords—Action recognition, video analysis, optical flow, convolutional neural networks, deep learning

I. INTRODUCTION

Recognition of human action in video is a challenging task that has recently received significant amount of attention among researchers of computer vision. Video based human action recognition has a vast and variant application area such as surveillance systems, robotics, health care and human computer interaction. Unlike classification in still images which is concerned with spatial information only, video data contains critical temporal information as well, which makes the classification task more challenging.

Recognition of human action from a video stream can be defined as classifying human actions automatically using a pattern recognition system with subtle human-computer interaction [1] [2]. Basically, an action recognition system analyzes certain video sequences or frames to learn the patterns of a particular human action in the training process and uses the learned knowledge to classify novel actions during the test phase.

In this research, we evaluate the extended temporal representation learning and the impact of high-quality optical flow extraction from the videos on action recognition in videos. Our experiment result confirms the advantages of extended temporal for learning efficient features for human action recognition.

II. RELATED WORK

As related works with action recognition in videos there were a few video oriented research were conducted. Video content analysis is one of the core problems in computer vision and has been studied for decades. Many research contributions in video processing have focused on developing spatiotemporal feature representation for video content analysis. A family of video content representation methods is based on shallow high-dimensional encodings of local spatiotemporal features. Some of these video representations include Spatiotemporal Interest Points (STIPS) [3], Histogram of Oriented Gradients (HOG) [4], Motion Boundary Histogram [5], Cuboids [6], and Action Bank [7]. These feature representations are hand-crafted and use different encoding technique such as those based on histogram or pyramids. Among these hand-designed representations, improved Dense Trajectories (iDT) [8], is widely considered the state of the art in handcrafted feature representation due to its bold results on video classifications.

With the breakthrough of deep learning in still-image recognition originated by AlexNet model [9], researchers devoted significant contribution to design similar model for video. 3D CNNs using temporal convolutions for recognizing human actions in video were arguably first proposed by M. Baccouche et al [10]. More recently 3D CNNs were shown to lead to strong action recognition results when trained on large amount of datasets. The 3D CNNs features also generalize well to other tasks, including action detection, video captioning [11] and gesture detection [12].

Current 3D CNN methods for action recognition mostly extend CNN architecture designed for static image classification. These networks are bound to learn short time feature representation. Basically, human actions in a video such as apply eye makeup, bench press or jogging lasts several seconds and spans sixty or hundreds of video frames. Breaking down this structure to a clip and congesting video information by simply average of the clips is likely to be optimal in the video level. Considering this, we design a new network architecture with extended temporal convolution to learn extended temporal information in video clips.

III. EXTENDED TEMPORAL CONVOLUTION

Based on our empirical observation most actions favor long extents, lasting 4-7 seconds on average. Computing an action from video consists of 120 frames on average. Unlike still images, video contains temporal information which is necessary to represent them during the learning. We believe that preserving long temporal resolution should enable the network to learn more complex features. Figure 1 depicts video frames extracted from two classes of action of UCF-101 dataset. Our novel extended temporal convolution network is able to learn this video representation over long period of time.

Figure 1: Video frames from two classes of action. (a) and (b) are video frames taken from UCF-101 dataset, indicates ApplyEyeMakeup and ApplyLipstick action, respectively. Action contains features: space-time patterns that lasts 4-7 seconds on average (roughly 120 frames). Extended temporal convolution is able to learn this video representation over extended periods of time.

IV. Network Architecture

We propose a novel network architecture design with extended temporal convolution departed from the work in [13]. Figure 2 illustrates our network architecture with extended temporal resolution. Our network has 5 convolutional layers, with 64, 128, 256, 256 and 256 kernel activation maps followed by two fully connected layers and a softmax output layer. We use 3D kernel of size of $3 \times 3 \times 3$ with stride 1 for all convolutional layers. Rectified linear unit is used in between each convolutional layer followed by a space-time max pooling layer. Max pooling kernels are of size $2 \times 2 \times 2$ with stride $2 \times 2 \times 2$ except for pool1 which has kernel size of 2x2x1. To help the network learn more complex spatial and temporal features we use 0.9 drop out in the first two fully connected layers. Fully connected layers are followed by ReLU layers. Softmax layer is used at the end of the network to output class score probability over classes of action.

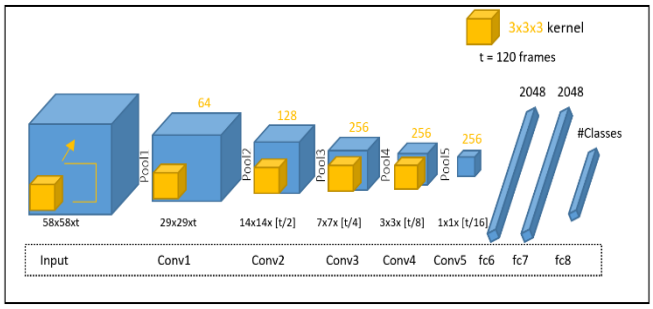

Figure 2: Our network has 5 Convolution, 5 max pooling, and 2 fully connected layers, followed by a softmax layer. All 3D convolution kernels are size $3 \times 3 \times 3$ with space-time stride of 1. The numbers above each box denotes number of filters. The max-pooling layers are denoted from pool1 to pool5. All pooling kernels are size $2 \times 2 \times 2$, except for pool1 which is $2 \times 2 \times 1$. Each fully connected layer has 2048 output units.

V. Implementation

We train our networks on the training split of both UCF-101 and HMDB51 datasets independently. We use SGD applied on mini-batches with loss function of negative log likelihood. Due to limitation in our GPU we use 10 clips of batch size as an input to the network. We set the weight decay to 3×10^{-5} and decrease it by a factor of 10^{-1} at every reduction of the learning rate. The momentum and dropout ratio are set to 0.9. We feed the network with random crop input patches of size $58 \times 58 \times 120$ from videos of size 171×128 pixels rescaled to 89×67 spatial resolution. We extend the temporal resolution to 120 frames and reduce spatial resolution to 58×58 have reduced network complexity. Some classes of action are relatively short and have 80-100 frames on average. We pad short clips and feed them to the network. Given the small size of UCF-101 and HMDB51, we fine tune network that has been trained on large datasets. In the first step, we start off by fine tuning 16f C3D network trained on Sports-1M dataset [1] to UCF-101 dataset. A randomly initialized fully connected layer of size 101 is added at the end of the network. We freeze the convolutional layers while fine tuning the fc layers. We start training our network with learning rate of 2×10^{-4} and decrease it to 2×10^{-5} after 25K iterations for 5K more iterations. In the second step, we feed 120frames to the network and fine tune all the layers. Convolutional layers are applied to 120frames. We re-trained fc layers of C3D, and down-sample temporal resolution of conv5 output and pass it to fc6. We run the same number of operations with a slight change in starting a learning rate of 2×10^{-5} and decrease it to 2×10^{-6}. We also examine the impact of optical flow on our network. In order to feed two channels of optical flow (opt-flow x, opt-flow y), we add a symmetric network model of our network to the existing network and use the idea in [14] to fuse the output. We benefit from Brox [15] optical flow for our experiment, owing to its high accuracy and good representation of motion in videos. Our experimental result shows that action recognition is easier to learn from motion representation rather than raw pixel values.

A. Datasets and Evaluation Metrics

The UCF-101 [16] is popular benchmark dataset used for action recognition. It consists of 101 action classes, which include 13,320 video clips. The videos are collected from YouTube, lasting 10-15 seconds on average with the total number 2.4M frames. The videos have 25fps frame rate with spatial resolution of 320 x 240 pixel.

The HMDB-51 [17] is another well-known benchmark dataset consist of 51 action classes, which include 6766 videos. The videos have 30 fps frame rate with spatial resolution 320 x 240 pixels. Although this dataset is considered large enough for action recognition in the past, the amount of data for training a deep learning model is limited.

For evaluation, we rely on standard evaluation metrics, i.e per-clip accuracy and per-video accuracy. For clip accuracy, we obtain per-clip accuracy by assigning each clip the class label with maximum softmax probability output and measure the number of correctly assigned labels over all clips. To measure video accuracy, we first obtain the video score by per-clip softmax score's average and take the maximum value of this average as video score. We then average overall videos to obtain video accuracy. We report our result according to the dataset's suggested evaluation protocol, which is the mean video accuracy across the given three test splits of the dataset.

B. Data Augmentation

Data augmentation is a crucial technique to improve the performance of deep learning models with having limited amount of the data. We applied data augmentation on both spatial and temporal resolution. During training stage, we used random clipping on temporal resolution and random cropping on spatial resolution. We also applied random left-right

flipping for each clip during the training. In order to evaluate the gain of data augmentation we first applied each data augmentation individually. We then combined them to evaluate the result. Our combined random clipping, random cropping and left-right flipping offer us 5.1% accuracy gain over clip accuracy and 4.1% accuracy gain over video accuracy.

VI. Result

In this section we present our experiment result. We train our network using UCF-101 and HMDB51 datasets. The datasets are provided in three splits of the training, validation and test sets. We train the network using the training split of the dataset, validate it on the validation split and provide result on the test split of the dataset. We first train the network using the RGB data. Our result shows that training network for action classification with RGB data gain 74.1% accuracy, worse than 2D CNN models. To evaluate network performance, we also train our network on flow data extracted from video clips using Brox optical flow algorithm. Table 1 reports that our novel extended temporal convolution network architecture with optical flow data outperforms many handcrafted algorithms and 2D CNN based networks for action recognition. Our model achieves good result comparable to that of the state of the art. Due to lack of resources and limited amount of data we are unable to train our network from scratch. We believe that training our network on large amount of data, our network will achieve higher accuracy.

Table 1: Comparison with other proposed deep learning models on action recognition. Our model achieves comparative performance on UCF-101 and HMDB51. Results are provided with mean accuracy across 3 splits of the datasets with @top-5 video classification.

Method	UCF-101	HMDB51
Single frame CNN Model (RGB) [14]	73.0	-
Single frame CNN Model (Optical flow) [14]	73.9	-
ImageNet + linear SVM (	68.8	-
Deep Networks [1]	65.4	-
Spatial stream (RGB) [14]	73.0	40.5
IDT +FV	85.9	57.2
Temporal stream (flow) [14]	83.7	54.6
C3D (3 nets) [13]	85.2	-
Two streams (RGB + Flow) [14]	**88.0**	59.4
Extended temporal convolution (ours) (RGB)	74.1	53.6
Extended temporal convolution (ours) (Optical flow)	86.2	**62.5**

VII. Conclustion

In this research we studied extended temporal convolution for recognizing human action in videos. Our study led to depart a new network architecture with extended temporal convolution for action recognition. We train our model on UCF-101 and HMDB51 benchmark datasets independently. We used extended temporal convolution on large number of video frames and obtained result comparable to the state-of-the-art on UCF-101 and HMDB51 datasets. We also presented that impact of optical flow over RGB data. Optical flow improves the result in video analysis significantly. We hope that our analysis will inspire new ideas and help to improve the efficacy and modeling for extended convolution network. Our future work will be devoted to search and improve further state-of-the-art architectures for human action recognition.

Acknowledgment

This research was supported by Basic Science Research Pr ogram through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (GR 2016R1D1A 3B03931911)

References

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

[2] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined heirarachical compound features," IEEE Transactions on Pattern Analysis and Machine Inteligence, vol. 33, no. 5, pp. 883 - 897, 2011.

[3] I. Laptev and T. Lindeberg, "Space-time interest points," in ICCV, 2003.

[4] N. Dalal and B Triggs, "Histogram of Oriented Gradients for Human Detection," Proc. CVPR, vol. 2, pp. 886 - 893, 2005.

[5] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in European Conference on Computer Vision, 2006.

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features.," in Proc. ICCV VS-PETS, 2005.

[7] S. Sadanand and J. Corso, "Action bank: A high level representation of activity in video," in CVPR, 2012.

[8] H. Wang and C. Schmid, "Action recognition with improved trajectories," in ICCV, 2013.

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

[10] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential Deep Learning for Human Action Recognition," in Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[11] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in CVPR, 2016.

[12] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in CVPR, 2016.

[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features With 3D Convolutional Networks," in The IEEE International Conference on Computer Vision (ICCV), 2015.

[14] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in NIPS, 2014.

[15] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in ECCV, 2004.

[16] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," in CRCV-TR-12-01, 2012.

[17] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in ICCV, 2011.

[18] Z. Qiu, T. Yao, and T.Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in CVPR, 2017.

[19] J. Carreira and A. Zisserman., "Quo vadis, action recognition? A new model and the kinetics dataset," in in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.