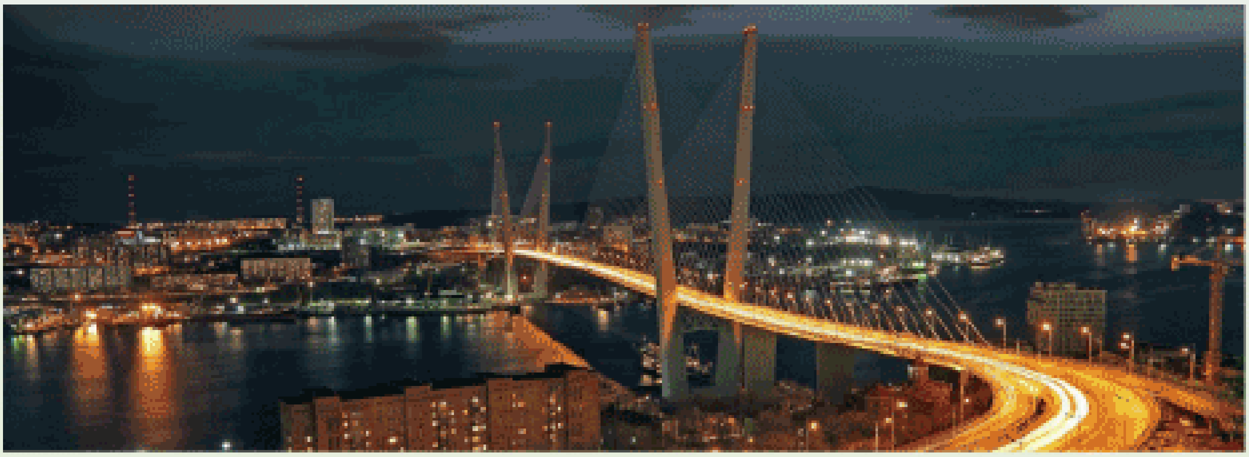
ISSN 2287-4348
Vol. 6 No. 2

ICSSC 2017

The 1st International Conference on Software & Smart Convergence

27-30 June 2017
Far Eastern Federal University, Vladivostok, Russia



Smart Media
KOREAN INSTITUTE OF SMART MEDIA

Oral Presentation I

Session C Smart Information

13:30 – 14:50, Campus E 321

Session Chair : Mucheol Kim(Wonkwang Univ.)

P.60	An Empirical Investigation of the Importance of IS Control Mechanisms Compatibility in Achievement of Superior IS Capabilities Elizaveta Srednik, Kyung Jin Cha*(Kangwon Nat'l Univ., Korea)*
P.65	Detection of Malicious Code using the FP-Growth Algorithm and SVM Yeongji Ju, Juhyun Shin*(Chosun Univ., Korea)*
P.69	Adapting Parallel Computer Simulation of Physical Fields to Multiple Problem Domains Andrey A. Chusov, Lubov G. Statsenko, Alexey P. Lysenko, Sergey N. Kuligin, Nelly A. Klescheva*(FEFU, Russia)*

Session D Computer Vision, Image Processing & Software Applications

13:30 – 14:50, Campus E 322

Session Chair : Soo-Hyung Kim(Chonnam National Univ.)

P.74	Deep RNN-CNN Based Activity Detection from Video Yali Nie*(Chonbuk Nat'l Univ., Korea)*, Yong Suk Cho*(Hansei Univ., Korea)*, Yongchae Jeong, Dong Sun Park*(Chonbuk Nat'l Univ., Korea)*
P.78	A Study of Character Input Interface based on Drag Gesture with a Smart Devices Kitae Bae*(SMIT, Korea)*, Libor Mesicek*(J.E. Purkinje Univ., Korea)*, Hoon Ko*(Sungkyunkwan Univ., Korea)*
P.82	WI-SUN based Cattle Shed Management System Sooho Jeong, Hyun Yeo*(Sunchon Nat'l Univ., Korea)*

International Conference on Software & Smart Convergence (ICSSC) 2017, June 2017

Deep RNN-CNN Based Activity Detection from Video

Yali Nie[1], Yong Suk Cho[2], Yongchae Jeong[3] and Dong Sun Park[4]
[1] Department of Electronics Engineering, Chonbuk National University
Jeonju, South Korea
[e-mail: nieyali@jbnu.ac.kr]
[2] Department of Media and Advertising, Hansei University
Gunpo, South Korea
[e-mail: adcho@hansei.ac.kr]
[3] Division of Electronics and Information Engineering, Chonbuk National University
Jeonju, South Korea
[e-mail: ycjeong@jbnu.ac.kr]
[4] IT Convergence Research Center, Chonbuk National University
Jeonju, South Korea
[e-mail: dspark@jbnu.ac.kr]
*Corresponding author: Dong Sun Park

Abstract

Deep learning has achieved great success for visual recognition or classification in still images. However, it has not yielded significant gains for classification and detection tasks from video. Our work proposes a new method to classify multiple activity in a video sequence. We fix the frame as input and combine Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to extract features and predict the different activities respectively. Pose and action are very similar concepts. A sequence of pose will comprise of an action in some certain environment, so with a stack of frames, it is very much possible to get rich source for activity detection from a video. We demonstrate the effectiveness of the proposed method on two datasets HMDB51 and UCF101.

***Keywords*:** Activity, Convolutional Neural Network, Detection, Recurrent Neural Network

1. Introduction

Recognizing activities from video has become a popular topic along with deep learning performing a dramatic change in the field of computer vision. It is a challenging problem due to factors such as large pose and scale variations, fast motions, body part occlusion and varying number of persons per video. In activity classification, the performance of a detection system depends on whether it can extract and simultaneously classify the activity rightly. However, there are two crucial obstacles: appearances and dynamics in action video. We propose to address the multiple-activity classification problem by means of a deep learning model comprised of the convolutional neural network and the long short-term memory (LSTM): recurrent neural network. Recently, ConvNet have witnessed big success in image classification, object detection [1], pose estimation [2,3,13,14], and other complex events. Similarly, LSTM can deal well with sequential research, such as image caption [16], speech recognition and tracking [11]. Some work[4,5,6] also use deep learning to learn actions from video and get good results, Inspired

This work was supported by the Brain Korea 21 PLUS Project, National Research Foundation of Korea. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No.NRF-2016R1D1A1B03936269).

by[15] and using CNN and RNN compound model, we are able to track multiple pose and select the one that best explain the activity class from a video.

2. Related Work

Research a sequence of pose expresses a human action. From video, we can get more information than a single image. Multiple-activity classification from video falls into two categories: (1) single image pose estimation, (2) video activity recognition.

Single Image Pose Estimation. Several works [2,13] have been in done towards designing effective deep neural network to detect key points location. In this work, some challenges are: foreground occlusions, background clutter, large scale, pose changes and multi-instance objects. Pictorial structures model [14] usually combine of unary and graph. Convolution Pose Machine [13] uses a sequential prediction framework to get the spatial correlations among body parts. State-of-the-art performance is achieved by the multi-context attention network [2], which use repeated pooling down and up sampling process with visual attention mechanism to learn the spatial distribution.

Video Activity Recognition. Human motion capturing has been long studied in machine learning. [3] tries to focus on articulated pose estimation in unconstrained video. Single person pose estimation in videos [12] aim to increase performance by utilizing temporal smoothing constraints. [11] multi-person Pose Track provides a method to track pose in the video, but it doesn't make a classification of the different activities.

In this paper, we combine the ConvNet and LSTM into one network. We can get sequence of video features from the CNN, which is an input to a RNN.

3. Proposed Architecture

3.1 Architecture

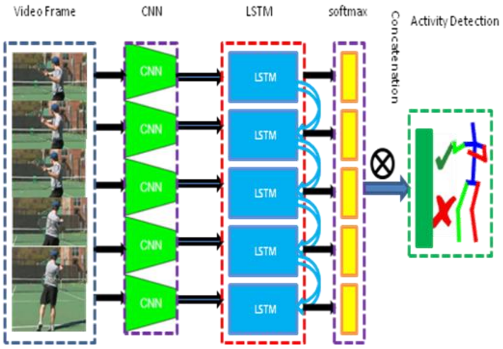


Fig. 1. Activity detection via a model combined with CNN and LSTM.

We design a deep network that deals with a sequence of frame from a video which shows as **Fig. 1**. With the fixed number of 16 frames, each frame is followed by a ConvNet architecture to generate a feature vector. The LSTM layer with input gate, forget gate and output gate at one time, and the hidden state will remember the action performance of a person. The output will predict next pose based on the past memory content. Figure 1 shows the proposed architecture. A sequence prediction of pose estimation from LSTM processing will fuse to produce the final classification of activity.

3.2 Processing

The output of the CNN is represented by describing the spatial information of the image, which is taken as the input of a LSTM cell at the time. The cell gates are as follows:

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + b_f) \quad (2)$$

$$O_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + b_o) \quad (3)$$

$$g_t = \delta(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$

$$m_t = o_t \odot \delta(c_t) \quad (6)$$

where $\odot$ stands for the element-wise multiplication. σ and δ represents the nonlinear activity function : sigmoid function and tanh function. m is the hidden state. LSTM followed by a softmax ($K+1$), here K is the number of activity classes in the dataset.

4. Experimental Classification Results and Analysis

For the task of activity classification, we use two large action datasets, which are HMDB51 and UCF101. HMDB51 is composed of 6,766 video clips and 51 actions classes, collected from various realistic videos, including 3,570 trained and 1,530 test videos. The UCF101 consists of 101 action categories and 13,320 video clips with 9,537 trained and 3,783 test videos.

We train the network with a random 224 × 224 crop from the sequence frame in the video as input. During training, the batch size of 256 is used, we train the whole model with RMSprop for 20k iterations and the learning rate as 10-4. 10 timesteps and 4000 hidden nodes are consisted of LSTM layer, following a softmax layer to recognize the activity.

Our method compared with other methods is as shown in **Table 1**. The performance of our method gets an accuracy of 70.6% on the UCF101 dataset and 41.0% on the HNDB51 dataset respectively. We can see that the activity detection in video still has a large scope for improvement, especially on the HMDB51 dataset. Even though deep learning is very powerful, to use it for the real world, is still a long way to go.

Table 1. Comparison of our method performance with other methods on UCF101 and HMDB51

Method	**Dataset**	
	UCF101	**HMDB51**
Model[7]	55.4%	23.6%
Wang[8]	41.5%	16.9%
Two-stream [9]	73.0%	40.5%
HOG[10]	72.4%	40.2%
Ours	70.6%	41.0%

5. Conclusions

In this paper, we have demonstrated that combining CNN and LSTM together to make an activity detection from a video can be efficiently utilized. Our deep structured architecture model evaluates on the activity dataset which are UCF101 and HMDB51. The sequence to sequence method can learn better for more challenging task in video processing.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *in CVPR*, 2014.

[2] X. Chu, W. Yang, W. Ouyang, C. Ma, A. Yuille and X. Wang, "Multi-Context Attention for Human Pose Estimation," *in CVPR*, 2017.

[3] U. Iqbal, M. Garbade, and J. Gall, "Pose for Action – Action for Pose," *arXiv preprint arXiv:1603.04037*, 2016.

[4] B. XiaohanNie, C. Xiong and S. Chunzhu, " Joint Action Recognition and Pose Estimation From Video," *in CVPR*, 2015.

[5] A. Montes, A. Salvador, S. Pascual and X. Giro-i-Nieto, "Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks," *arXiv preprint arXiv:1608.08128*, 2016.

[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang and L. Van, "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition," *arXiv preprint arXiv :1608.00859*, 2016.

[7] S. Purushwalkam and A. Gupta, "Pose from Action: Unsupervised Learning of Pose Features based on Motion," *arXiv preprint arXiv:1609.05420*, 2016.

[8] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," *arXiv preprint arXiv:1505.00687*, 2015.

[9] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *arXiv preprint arXiv:1406.2199*, 2014.

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

[11] U. Iqbal, A. Milan and J. Gall, "PoseTrack: Joint Multi-Person Pose Estimation and Tracking," *in CVPR*, 2017.

[12] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman, "Personalizing human video pose estimation," *in CVPR*, 2016.

[13] S. Wei, V. Ramakrisma, T. Kanade and Y. Sheikh, "Convolutional Pose Machines," *in CVPR*, 2016.

[14] X. Chen and A. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," *in NIPS*,

76 ICSSC 2017 ICSSC 2017

2014.

[15] M. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A Hierarchical Deep Temporal Model for Group Activity Recognition," *arXiv preprint arXiv:1511.06040* , 2016.

[16] A. Karpathy and F.-F. Li, "Deep visual-semantic alignments for generating image descriptions," *in CVPR*, 2015.