International Journal of Advanced Smart Convergence Vol.9 No.4 115-119 (2020)
http://dx.doi.org/10.7236/IJASC.2020.9.4.115

IJASC 20-4-14

A Study of Video-Based Abnormal Behavior Recognition Model Using Deep Learning

Jiyoo Lee[1 †], Seung-Jung Shin[2]

[1†] Ph.D. Candidate, Department of IT Convergence , Graduate School, Hansei University, Korea
iamjiyoo@gmail.com
[2]Professor, Department of ICT Convergence, Hansei University, Korea
expersin@hansei.ac.kr

Abstract

Recently, CCTV installations are rapidly increasing in the public and private sectors to prevent various crimes. In accordance with the increasing number of CCTVs, video-based abnormal behavior detection in control systems is one of the key technologies for safety. This is because it is difficult for the surveillance personnel who control multiple CCTVs to manually monitor all abnormal behaviors in the video. In order to solve this problem, research to recognize abnormal behavior using deep learning is being actively conducted. In this paper, we propose a model for detecting abnormal behavior based on the deep learning model that is currently widely used. Based on the abnormal behavior video data provided by AI Hub, we performed a comparative experiment to detect anomalous behavior through violence learning and fainting in videos using 2D CNN-LSTM, 3D CNN, and I3D models. We hope that the experimental results of this abnormal behavior learning model will be helpful in developing intelligent CCTV.

Keywords: Behavior Recognition, Deep Learning, CNN-LSTM, 3D CNN, I3D

1. INTRODUCTION

Video-based abnormal behavior detection is closely related to our safety. However, in the conventional video-based abnormal behavior detection, the surveillance personnel detected it with the naked eye in real time, but it was difficult for one person to monitor all the video data as the amount of video data to be monitored increased. As AI technology advances, it is now possible to overcome the limitations perceived by the naked eye. There is a need for a method to identify specific abnormal behaviors by viewing real-time images and analyzing abnormal behaviors or behavior patterns[1]. In addition, through the rapid development of deep learning, various models that can analyze video-based data have been proposed. However, it was difficult to realize video-based abnormal behavior detection because it was difficult to provide data on abnormal behavior due to personal information protection, and sufficient learning data could

Manuscript Received: October. 19, 2020 / Revised: October. 28, 2020 / Accepted: November. 4, 2020
Corresponding Author: expersin@hansei.ac.kr
Tel: +82-10-3664-0666, Fax: +82-31-450-5172
Professor, Department of ICT Convergence, Hansei University, Korea

not be guaranteed[2].

From December 2019, the National Information Society Agency started to provide AI data for CCTV abnormal behavior for public security as part of the "Infrastructure for Intelligent Information Industry", and it was possible to guarantee sufficient learning data through deep learning[3].

This paper intends to study the recognition method of a video-based abnormal behavior detection model using deep learning.

2. RELATED RESEARCH

The core technology in the era of the 4th industrial revolution can be said to be artificial intelligence (AI). The dictionary meaning of AI is a technology that realizes human learning, reasoning, perception, and comprehension skills through programs.

Video-based understanding of human behavior through application of deep learning Deep learning is a type of machine learning based on the way the human brain works. With the explosive increase in computing capacity and large datasets, deep learning is demonstrating its fast processing power in finding patterns in unstructured data such as images, tests, audio and video.

In order to recognize behavior, convolutional Neural Network(CNN) deep learning, a convolutional neural network, has been used a lot. CNN is a type of feed-forward neural network proposed for image recognition. It receives an input value of a predetermined size and outputs an output value of a predetermined size[6][8]. Among CNN algorithms, ResNet is a network with a total of 152 layers of 20 deep networks, which has good performance, but has a problem that learning is slow due to many parameters[7].

The CNN-LSTM model is a method that combines a CNN showing good results in still images with a Long Short Term Memory (LSTM) network that recognizes the current state as past information[4]. After applying a CNN to an individual image, it connects to the LSTM network to recognize the behavior so that not only the spatial pattern in each frame but also the temporal pattern in the frame sequence can be extracted. This method also has a problem in that it cannot consider frames of different viewpoints when summarizing one frame of an image.

3D CNN, which is a step up from the existing CNN, is a method of recognizing space-time. If the existing CNN performed only space, the 3D CNN operates including the time axis, so it is possible to recognize time division and behavior in successive images[5][10]. However, there is a problem that learning is difficult because there are many parameters to learn as the dimension is increased.

3. VIDEO ANALYSIS AND EXPERIMENT

3.1 Creating Learning Data

The data for the analysis of abnormal behavior was using AI Hub, an AI integrated platform built by the National Information Society Agency in 2017. AI Hub provides 700 hours (8400 cuts) video data set including 12 abnormal behavior behaviors (assault, fight, theft, vandalism, fainting, roaming, intrusion, speculation, robbery, data violence and harassment, kidnapping, drunk behavior) in CCTV for public security.

"Table 1" shows the types of abnormal motions, number of videos, and video duration provided by AI Hub. It can be seen that there is a lot of data in the order of actions taken and violence.

Table 1. Abnormal behavior dataset

Abnormal behavior name	*Number of images*	*Video time*
Assault	913	78:05:41
Fight	1174	99:54:45
Burglary	839	69:33:24
Vandalism	490	41:28:46
Swoon	912	84:26:16
Wander	645	55:24:50
Trespass	259	22:03:15
Dump	259	22:03:15
Robbery	259	22:03:15
Datefight	693	58:21:45
Kidnap	262	22:24:08
Drunken	1262	104:20:37
Total	8436	717:03:33

In this paper, data of two abnormal behaviors (assault, fainting) out of 12 abnormal behaviors were used, and only outdoor data was used to provide three types of data: indoor, outdoor, and chroma key for each data.

3.2 Pre-processing of Learning Data

The abnormal behavior data provided by AI Hub has a problem that human behavior is very small compared to the image size, and the abnormal behavior time is very short compared to the image length, which makes learning difficult.

Therefore, as shown in Figure 1(a), for efficient learning of each image, preprocessing was performed to cut out only the scene of the image where assault or fainting occurred. As shown in Figure 1(b), pre-processing of crop was performed to increase the size of a person.

Figure 1 Actual data (a) and cropped data (b)

3.3 Experiment Method

In this paper, for a video-based abnormal behavior learning experiment, two existing research methods, ResNET-LSTM and 3D ResNet, and I3D model were used as behavior recognition models.

Inflated 3D Convolution Network (I3D) is a method of adding RGB frames and optical flow frames as inputs,

unlike existing models to increase the performance of behavior recognition. Compared to the conventional 3D CNN, this method can improve performance by constructing a model with fewer parameters and performing training[6][9].

For behavior recognition, learning was conducted with a total of 573 training data and 37 test data.

When predicting every frame, a case in which the predicted value is changed in real time occurs, so a method of using the largest predicted value by summing the predicted values for the last 10 frames was applied.

3.4 Experiment Environment

The environment for the classification experiment was run on Cuda 9.1 on Ubuntu 16.04 LTS. The detailed experiment environment is described in Table 1 below.

Table 2. Experiments environments

Name	Spec
OS	Ubuntu 16.04 LTS
RAM	32G
GPU	NVIDIA GPU 1080 Ti
CUDA	9.1

3.5 Experiment Result

Tables 3 and 4 show the video behavior recognition results through the three execution methods described in 3.3.

To analyze the results of Tables 3 and 4, when looking at the video behavior recognition inference time, the reasoning time was shorter and the accuracy was the best for I3D compared to the other two models. However, the inference time with 3D RestNet50 showed little difference. However, you can see that there is a difference in accuracy.

Table 3. Training data Inference and Accuracy result

Name	Inference	Accuracy
3D ResNet50	0.035	0.75
ResNet50 +LSTM	0.06	0.43
I3D	0.029	0.93

Table 4. Test data Inference and Accuracy result

Name	Inference	Accuracy
3D ResNet50	0.035	0.75
ResNet50 +LSTM	0.06	0.59
I3D	0.029	0.98

4. CONCLUSION

It is important to identify crime prevention, disaster, and emergency situations through the analysis of abnormal behavior or behavior patterns. Since it is impossible to identify and check images with limited personnel, researches that can detect abnormal behavior using artificial intelligence technology are being actively conducted.

We presented a video anomaly detection model based on deep learning. Through the experimental results of 2D CNN+LSTM, 3D CNN, and I3D, abnormal behavior was learned and the performance of the I3D model was confirmed to be good in accuracy. However, in this paper, since learning was performed using only 610 data, pre-processing was necessary because there were fewer actions to learn compared to images and the time for abnormal behaviors was shorter compared to the image length. If there is a large amount of data that can be continuously learned, research is needed to detect various abnormal behaviors, not just violence and fainting, by learning patterns by themselves, out of manual preprocessing.

REFERENCES

[1] Baek, S.-I., Lim, G.-G., & Yu, D.-S, "Exploring Social Impact of AI," *23*(4), 3–23, 2016.
DOI: https://doi.org/10.22693/NIAIP.2016.23.4.003

[2] P. Y. Simard, D. Steinkraus and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis." *Seventh International Conference on Document Analysis and Recognition*, 2003. Proceedings., Edinburgh, UK, 958-963, 2003.
DOI: 10.1109/ICDAR.2003.1227801..

[3] AI HUB, http://www.aihub.or.kr, 2019.

[4] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory." *Neural computation*, vol. 9, 1735-80.
DOI: https://doi.org/10.1162/neco.1997.9.8.1735,1997.

[5] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition", in ICML,2010.
DOI: https://doi.org/10.1109/TPAMI.2012.59

[6] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset." *2017 IE EE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4724-4733, 2017.
DOI:10.1109/CVPR.2016.90

[7] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 770-778, 2016.
DOI: https://doi.org/10.1109/cvpr.2017.502

[8] Sooyeon Lim, "Estimation of gender and age using CNN-based face recognition algorithm," *International Journal of Advanced Smart Convergence* Vol.9 No.2 203-211, 2020.
http://dx.doi.org/10.7236/IJASC.2020.9.2.203

[9] T. Kwon, J.-Y. Lee, and K.-D. Jung, "Design of Falling Recognition Application System using Deep Learning," *International Journal of Internet, Broadcasting and Communication*, vol. 12, no. 2, pp. 120–126, May 2020.
DOI:10.7236/IJIBC.2020.12.2.120

[10] Ji-Sub Kim, Chang-Jun Nan, & Byoung-Tak Zhang. "Deep Learning-based Video Analysis Techniques," *Com munications of the Korean Institute of Information Scientists and Engineers* ,33(9), 21-31,2015.