Enhanced 3D Action Recognition Based on Deep Neural Network

Sungjoo Park
Information Media Research Center
Korea Electronics Technology Institute(KETI)
Seoul, Korea
bpark@keti.re.kr

Dongchil Kim
Information Media Research Center
Korea Electronics Technology Institute(KETI)
Seoul, Korea
dekim@keti.re.kr

Abstract**—**In the video surveillance system operating in the real environment, the recognition of detailed behavior of objects has important meaning in terms of understanding whether security events have occurred. In particular, the perception of behavior in a poor environment such as low light and overlapped objects is recognized as an important technical factor that must be overcome in the existing 2D image-based surveillance system. Action recognition of objects using 3D depth map provides a way to solve these issues. In this paper, we propose the enhanced 3D action recognition method based on convolution neural network (CNN) for the video surveillance system. And we evaluated the action recognition performance using the real environment DB, and the recognition result for 6 detailed behaviors was confirmed to be an average of 68.56%.

Keywords—3D action recognition, video surveillance, depth-map image analysis, deep neural network

I. Introduction

Intelligent video surveillance systems are widely used to provide security services such as intrusion detection and recognition of dangerous and abnormal behaviors of specific objects. These intelligent video security systems perform object detection and action recognition using CCTV camera video input. However, it is true that the performance of the intelligent video security system used in real security services is limited by external environmental factors such as low light, fog, and rain. In order to overcome these environmental factors, a method of using 2D image data and 3D depth map information together has been proposed. For this purpose, recently, CCTV camera products that provide distance information through an infrared sensor and temperature information using a thermal image sensor are attracting attention. In particular, the infrared sensor has been mainly used for gesture recognition in entertainment services, but lately it is also used in CCTV cameras for services such as people counting using depth map information. Action recognition using 3D depth map can provide a meaningful way to overcome environmental factors such as illumination, clustered background and camera motion that can occur in common CCTV cameras [1].

There are two major approaches in human action recognition using 3D cameras: sequential approach and space-time approach. There have been many studies in the sequential approach. First method used R transformation including Radon transformation on the 3D depth map to construct feature vectors and applied PCA and LDA to the feature vectors [2]. The second algorithm of the sequential approach is to extract histograms of 3D joint location (HOJ3D) vectors using human joint information appearing in 3D depth [3]. The temporal and spatial approach uses the entire depth map for learning and does not detect the body part separately, thus reducing the error rate from the detection of the body part. However, only rough behavior analysis is possible and there may be limitations in detailed analysis. The algorithms of the spatiotemporal approach are as follows. First method extracted feature vectors using Depth Motion Map (DMM) and Histogram of Oriented Gradient (HOG), and then proposed an action recognition method using SVM [4]. Second method extracted features using Random Occupancy Patterns (ROP) that randomly samples 4-dimensional sub-volumes of different positions and sizes within the 4-dimensional space-time volume [5]. Third method considered the depth images as a 4-dimensional space-time volume and proposed a Space-Time Occupancy Patterns (STOP) feature [6] that described the distribution of 3D depth information corresponding to a person.

In this paper, we propose an enhanced action recognition for video surveillance system using 3D depth map based on deep convolution neural network. To do this, we constructed training datasets with most existing action datasets and our own datasets (KETI-RGB+D). And the evaluation results show that our proposed method can recognize detailed actions of real objects and can be used in surveillance applications.

II. Datasets for 3D Action Recognition

We analyzed the existing studies, and derived requirements for an action dataset suitable for 3D action analysis in the video security system. The construction of such a dataset is essential to support quantitative evaluation of various approaches as well as training of deep learning algorithms for action recognition. In order to improve and optimize the performance of action recognition in the intelligent video surveillance system, we built our own KETI-RGB+D dataset including 2D images and 3D depth map information about various poses of objects.

In previous studies, the MSR Action3D dataset is one of the earliest datasets for 3D skeleton-based activity analysis. This dataset consists of selected instances in the context of

978-1-6654-8550-0/22/$31.00 ©2022 IEEE

interacting with game consoles such as high arm waves, horizontal arm waves, hammers and hand grabs [7].

NTU RGB+D is a state-of-the-art large-scale benchmark for action recognition. It represents a series of standards and experience for building large-scale data. Recently reported results for this data set achieved satisfactory accuracy in this benchmark [8]. SYSU-3D-HOI is a dataset containing human-object interactions. It consists of a total of 480 clips and 11 human actions are defined.

KETI RGB+D is a dataset optimized for video surveillance systems. This dataset consists of 13 classes, 1,080 samples, 17 human subjects, 40 long video samples (1,520 action samples). It is generated with different height (2m, 1m) and 3 different angle for each action [9].

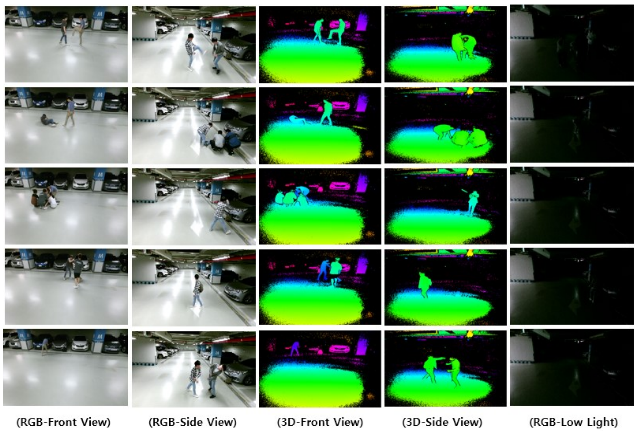


Fig. 1. Example of KETI RGB+D dataset

III. Architecture of the Proposed Action Recognition Method

We think that CNN is suitable for spatiotemporal feature learning that recognize specific actions rather than a simple scene recognition using 3D depth map information.

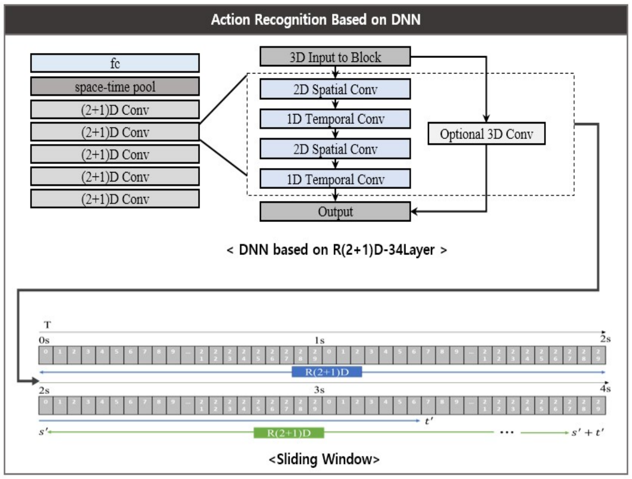


Fig. 2. Architecture of proposed action recognition based on DNN

Fig 1 illustrates the architecture of proposed DNN based on R(2+1)D-34Layer which performs convolution and pooling in spatiotemporal domain. And a sliding window method is adopted to secure the efficiency of 3D action recognition for video surveillance system in real environment.

IV. Experimental Results

The proposed 3D action recognition method was trained using the NTU RGB+D dataset and the KETI RGB+D dataset. The KETI RGB+D dataset used for training has 20 video clips, and each video chip consists of 4 predefined front view motion images. In the entire training dataset, 70% of the data was used to train the proposed method and the remaining 30% was used for validation.

To evaluate the performance of human action recognition, the KETI RGB+D dataset consisting of 20 left/right view video clips was used. The performance evaluation of the recognition accuracy of the proposed method is shown in Table I. The recognition result for 6 detailed behaviors was confirmed to be an average of 68.56%.

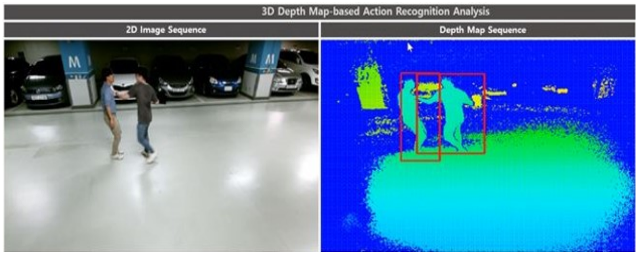
TABLE I. Accuracy of Action Recognition

Action Type	Mean Accuracy(%)[a]
Punch	62.96
Prowl	77.36
Falling	83.33
Kick	50.00
Vandalism	87.69
Grouping	50.00
Average	**68.56**

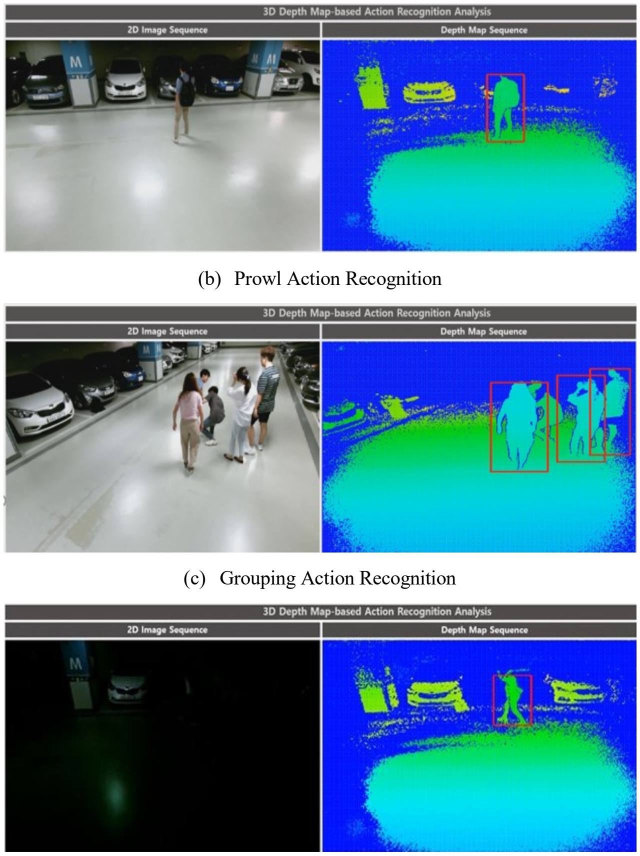
[a] Mean accuracy = sum of successful action recognition / total number of evaluation datasets X 100

Fig. 3 shows the implementation results of video surveillance system using proposed human action recognition method.

The Kinect video camera was used as an input device. User interface has been implemented to provide users with 2D image information, 3D depth map information, and recognized action information. The system can process 20 frames per second video input(800x600)



(a) Punch Action Recognition



(b) Prowl Action Recognition

(c) Grouping Action Recognition

(d) Vandalism Action Recognition

Fig. 3. Implementation Results of 3D action recognition for video surveillance system

V. Conclusions and Future Work

We try to improve object and action recognition performance by applying deep learning technology to the video surveillance systems. In addition, we are also try to find a technical way to overcome environmental factors such as lighting, background clusters, and camera motion that can occur in general CCTV cameras. In this respect, the action recognition technology using 3D depth map can be considered as a meaningful approach to improve the performance of the intelligent video surveillance system.

In this paper, we proposed an 3D motion recognition and security event detection method applicable to video surveillance systems. The proposed method performs DNN based on R(2+1)D-34Layer and adopts a sliding window for high-speed operation. Experimental results showed that the proposed method can recognize human action and efficiently detect security events in video surveillance systems.

We will further try to build more diverse object action datasets and improve the action recognition performance by learning more various information.

Acknowledgment

This work was supported by the ICT R&D program of MSIT/IITP. [2021-0-00348, Development of A Cloud-based Video Surveillance System for Unmanned Store Environments using Integrated 2D/3D Video Analysis]

References

[1] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu, “PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding,” arXiv preprint arXiv:1703.07475, 2017. descriptor defined on the Radon transform”. Computer Vision and Image Understanding, 2006, vol. 102, pp. 42-51.

[2] L. Xia, C.-C. Chen, and J. K. Aggarwal, “View invariant human action recognition using histograms of 3d joints,” IEEE Conf. Comput. Vision Pattern Recognition Workshops (CVPRW), 2012, pp. 20-27.

[3] X. Yang, C. Zhang, and Y. Tian, “Recognizing actions using depth motion maps-based histograms of oriented gradients,” Proc. 20th ACM international Conf. Multimedia, 2012, pp. 1057-1060.

[4] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, “Robust 3d action recognition with random occupancy patterns,” ECCV. Springer, 2012, pp. 872-885.

[5] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Zicheng Liu, and M. M. Campos, “Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences,” Progress Pattern Recognition, Image Anal. Comput. Vision, Applications. Springer, 2012, pp. 252–259.

[6] W. Li, Z. Zhang, and Z. Liu. “Action recognition based on a bag of 3D points,” In CVPR, 2010.

[7] A. Shahroudy, J. Liu, T.-T. Ng, and G.Wang, “NTU RGB+D: A large scale dataset for 3D human activity analysis,” In CVPR, 2016.

[8] Jiang Wang, Zicheng Lu et al., “Mining actionlet ensemble for action recognition with depth cameras,” IEEE Conference on Computer Vision and Pattern Recognition pp1290-1297, 2012

[9] Sungjoo Park, Dongchil Kim, "Study on 3D Action Recognition Based on Deep Neural Network," In ICEIC, 2019.