Video Surveillance System Based on 3D Action Recognition

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

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

Abstract**—** Human action recognition using depth-map images from 3D camera for surveillance system is a promising alternative to the conventional 2D video based surveillance. We propose a security-event detection method based on body part classification and human action recognition for more effective video surveillance system. Experimental results show that the body part classification accuracy of 65.0% and security event detection accuracy of 0.878 were achieved for 9 security events.

Keywords—Body part classification, 3D action recognition, Video Surveillance , depth-map image.

I. Introduction

Video surveillance system are widely used to monitor accidents and intrusion, or detect dangerous event for public security service. However conventional video surveillance system using 2D cameras has difficulties in detecting objects and various events precisely. We propose to use depth-map images from 3D cameras to supplement the 2D surveillance camera systems to detect objects and monitor abnormal events in detail. 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 proposed an action graph that graphically models human behavior [1]. Each node of the graph generated in this method expresses a major posture that is shared in common, and each edge is moved from one posture to the next posture to represent the motion probability. Second method used bi-gram with maximum likelihood decoding algorithm [2]. They proposed to use Bag-Of-Points (BOPs) as a set of major points forming contours to represent human outline information in 3D space. Third 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 [3]. The final algorithm of the sequential approach is to extract histograms of 3D joint location (HOJ3D) vectors using human joint information appearing in 3D depth [4]. 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 [5]. 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 [6]. Third method considered the depth images as a 4-dimensional space-time volume and proposed a Space-Time Occupancy Patterns (STOP) feature [7] that described the distribution of 3D depth information corresponding to a person. We propose a video surveillance system based on depth-map images from 3D camera in this paper. We evaluate body part classification, joint detection, action recognition, and finally predefined security event detection for more effective surveillance system. The evaluation results based on 9 predefined security events showed that our proposed method can be used in practical surveillance applications.

II. Propsed methods

In this section we provide detailed explanations about the proposed system based on based on body part classification and human action recognition. Fig 1.shows the block diagram of the proposed security event detection. The module is divided into two parts. First, we use a pre-processing module to refine the region of the depth image where the person is located. Then, each pixel of the depth image is classified into 43 parts such as head, neck, shoulder, arm, leg and torso through CNN classifier module. We extract 10 joints by the mean-shift method from the body parts classification results. Then, the Short Time Fourier Transform (STFT) pyramid is generated based on the action with the longest length L among the actions learned in the SVM training process based on the 1 joints in advance. From the generated STFT pyramid, we detect the joint-based human action located in the depth image of length L. The detected action classification results undergo a majority voting process from multiple frames. Second, the depth-based action detectors try to detect simpler actions such as gathering, group assault, appear, disappear, loitering, camouflage by using the motion information of the mean of the body in 3D images. These joint based feature vectors are then used to train with SVM for recognizing 9 predefined human actions after human action modeling.

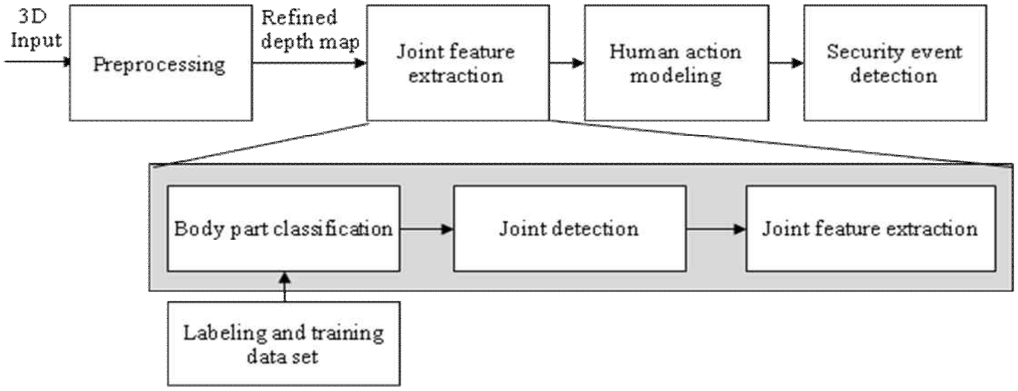

Fig. 1. Block diagram of proposed security event detection system

A. Preprocessing

The preprocessing module uses median filter and morphological operations. The median filter removes small noises and morphological operations such as closing to fill small holes appearing blob structures. Removing small noises and filling holes reduces unnecessary processing to make the foreground segmentation process more efficient.

B. Preprocessing

The preprocessing module uses median filter and morphological operations. The median filter removes small noises and morphological operations such as closing to fill small holes appearing blob structures. Removing small noises and filling holes reduces unnecessary processing to make the foreground segmentation process more efficient.

C. Body part classification and human action recognition

We use CNN Classifier for the body part classification (BPC). The BPC results are used for the joint detection. We synthetically generated depth and body part images using a free public software MakeHuman. After the BPC and joint prediction, the feature vectors for action classification is constructed. We use feature vectors constructed from pairwise relative positions of joints and their Fourier transformed results. [8] The joint based feature vectors are then used to train with SVM for recognizing 9 predefined human actions. Fig. 2 shows the information of body part classification .

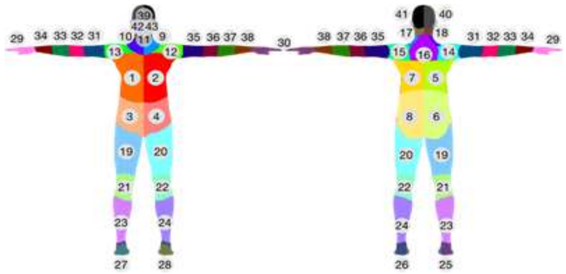

Fig. 2. Information of 43 body parts (front, back)

III. Experimental Results

A. Preprocessing

To separate the object from the background, we calculate the difference of depth between the object and the background and extract the blob of object with CCA. Fig. 3 shows the process of object extraction from background.

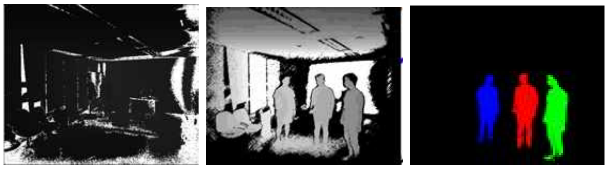

Fig. 3. Preprocess for object extraction from background

B. Body part classification

The training of body part detection process requires a depth image and its corresponding color-labeled image pair. To obtain these color-labeled images, synthetic images are created with 3D human body models. The color-labeled image represents each part of the body in a different color and we generated 100,000 depth and body part images based on Mocap data[9]. Fig. 4 shows the synthetic depth and body part labeled images.

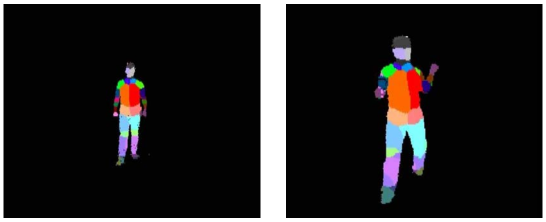

Fig. 4. Synthetic depth and body part labeled images.

These images are used to generate a classifier by off-line learning using CNN. After the body part classification, the position of the joint is found by calculating the mean of each body parts. [10] The body part classification process is greatly affected by the type of features, type of classifier, and the way classifier is trained. The features are typically chosen as depth gradient in unary or binary fashion. The size of window and allowed displacement from the window center are also important factors. We used real world unit (cm) rather than pixels in calculating the displacement to address the scale issue. We followed typical body part classification method and adopted CNN classifier. After training the CNN classifier, we report the results as mean accuracies for the entire test sample. As a result of learning 500 images consisting of poses up to ± 45 ° from the frontal view, we obtained an average classification accuracy of 80% and 65% for the synthesized and real depth images, respectively. We used depth images containing kick and punch motion for the body part classification evaluation.

C. Human action recognition

In order to evaluate the performance of the security event detection from human action, we collected videos for the 9 security events, 10 times for each events. The 9 types of security events are defined as appearing, disappearing, loitering, camouflage, line cross, punch, kick, falling, and abandonment. Five events among them such as appearing, disappearing, loitering, camouflage, line cross use depth image based features and other four events such as punch, kick, abandon, and falling use joint-based features for the detection process. Table I shows the comparison of performance with Kinect based system. The performance evaluation for detection accuracies of the proposed ssysem is shown in Table II.

TABLE I. ACCURACY OF HUMAN ACTION RECOGNITION

Action	**Mean accuracy**	
	Kinect based system[a]	***Proposed system***
Kick	0.5	1.0
Abandon	1.0	0.9
Fall	0,.7	0.8
Punch	1.0	0.8
Average	**0.83**	**0.875**

TABLE II. ACCURACY OF SECURITY EVENT DETECTION

Event	**Mean accuracy**	**Event**	**Mean accuracy**
Kick	1.0	Line cross	1.0
Abandon	0.9	Appear	1.0
Falling	0.8	Disappear	1.0
Punch	0.8	Camouflage	0.7
Loitering	0.7		
Average	**0.878**		

[a] Kinect based system used Kinect's object segmentation and joint detection function.

IV. CONCLUSIONS AND FUTURE WORK

Conventional 2D camera based video surveillance systems have been widely used in these days for security system, but their roles are still limited as recording events and serving in post-event analysis. There have been much efforts in developing 2D camera based surveillance systems that can be utilized as pre- or in-event alert media. In these days, the advancements of computing softwares and hardwares and 3D cameras have increased the feasibilities of developing more intelligent surveillance systems using 3D information. We proposed security event detection method in this paper. The proposed method performs depth gradient based feature construction, CNN based body part classification, joint prediction, joint pair based feature construction, and action recognition. The action recognition module is applied on an input video for each fixed length and decides whether a security event is included or not. Experimental results show that the proposed method can be used as an effective 3D camera based surveillance system. We are trying to further improve the action recognition performance by employing more robust features and classification methods.

ACKNOWLEDGMENT

This work was supported by the ICT R&D program of MSIT/IITP. [2014-3-00077, Development of global multi-target tracking and event prediction techniques based on real-time large-scale video analysis]

REFERENCES

[1] W. Li, Z. Zhang, and Z. Liu. Expandable data-driven graphical modeling of human actions based on salient postures. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(11): pp. 1499–1510.

[2] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points", CVPRW, 2010

[3] Tabbone, S. Wendling, L. Salmon, and J.-P.: "A new shape descriptor defined on the Radon transform". Computer Vision and Image Understanding, 2006, vol. 102, pp. 42—51.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[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] CMU Mocap Database (2016, November 21). Received from http://mocap.cs.cmu.edu/

[10] SHOTTON, Jamie, et al. Real-time human pose recognition in parts from single depth images. Communications of the ACM, vol. 56 no. 1, pp. 116-124, 2013.