2D Human Pose Estimation based on Object Detection using RGB-D information

Seohee Park, Myunggeun Ji, and Junchul Chun
Department of Computer Science, Kyonggi University
Suwon, South Korea
[e-mail: eehoeskrap@kgu.ac.kr, jmg2968@kgu.ac.kr, jcchun@kgu.ac.kr]
*Corresponding author: Junchul Chun

Received October 21, 2017; accepted December 13, 2017; published February 28, 2018

Abstract

In recent years, video surveillance research has been able to recognize various behaviors of pedestrians and analyze the overall situation of objects by combining image analysis technology and deep learning method. Human Activity Recognition (HAR), which is important issue in video surveillance research, is a field to detect abnormal behavior of pedestrians in CCTV environment. In order to recognize human behavior, it is necessary to detect the human in the image and to estimate the pose from the detected human. In this paper, we propose a novel approach for 2D Human Pose Estimation based on object detection using RGB-D information. By adding depth information to the RGB information that has some limitation in detecting object due to lack of topological information, we can improve the detecting accuracy. Subsequently, the rescaled region of the detected object is applied to ConVol.utional Pose Machines (CPM) which is a sequential prediction structure based on ConVol.utional Neural Network. We utilize CPM to generate belief maps to predict the positions of keypoint representing human body parts and to estimate human pose by detecting 14 key body points. From the experimental results, we can prove that the proposed method detects target objects robustly in occlusion. It is also possible to perform 2D human pose estimation by providing an accurately detected region as an input of the CPM. As for the future work, we will estimate the 3D human pose by mapping the 2D coordinate information on the body part onto the 3D space. Consequently, we can provide useful human behavior information in the research of HAR.

Keywords: RGB-D, Human Activity Recognition, Object Detection, Keypoint Localization, 2D Human Pose Estimation

A preliminary version of this paper was presented at APIC-IST 2017, and was selected as an outstanding paper. This work was supported by Kyonggi University Research Grant 2016. (Research Title: (2016-034) An automatic method for tracking and analyzing moving objects based on 3D CCTV.)

http://doi.org/10.3837/tiis.2018.02.015 **ISSN : 1976-7277**

1. Introduction

Recently, a variety of intelligent video surveillance analysis is possible due to the improvement of artificial intelligence technology and computing ability [1]. Therefore, deep learning technology is applied to computer vision-based Human Activity Recognition (HAR), which is one of the methods that need to be developed for intelligent video surveillance analysis [2, 3, 46]. In the intelligent video surveillance system, technologies that robustly detect various abnormal behaviors of human such as pedestrians are needed. These techniques are closely related to the field of HAR, and can be recognized through a top-down approach that detects humans in the image and estimates the poses of the detected humans to recognize and predict various behaviors [4]. In general, the CCTV images used in the field of video surveillance have occlusion problems due to loss of topological information caused by projecting 3D real world in 2D image. The occlusion that causes inaccurate object detection can be resolved by using depth information. In addition, we can estimate the human pose by obtaining the skeleton model representing the human pose within the detected region through the object detection process. Estimating the human pose is a process for expressing the appearance of a human, and is a necessary process to show the numerous poses the human body can take. Therefore, estimating human pose is important to recognize human activity [5, 6, 47].

In this paper, we propose a top-down approach to 2D Human Pose Estimation based on Object Detection using RGB-D information. First, we segment moving objects from the background in two horizontally connected CCTV environments. Depth information is combined with existing RGB information to solve occlusion and recognition problem due to lack of topological information of 2D image. The depth information can be obtained by calculating the disparity between the images generated from the left and right cameras and generating the depth map [25]. We perform object detection by segmenting an object using depth information in a region segmented by RGB information.

Then the region of the detected object is rescaled to generate input data for pose estimation. The generated input data is applied to ConVol.utional Pose Machines (CPM), which is a sequential prediction structure based on ConVol.utional Neural Network (CNN) [7, 8]. In this paper, the prediction result of body parts can be obtained by using a pre-trained model such as MPII Human Pose Dataset [9]. Additionally, CPM is used to generate belief maps to predict the localization of keypoint representing human body parts. This belief map predicts the body part sequentially through the loss function at the output of each step of the CPM and provides clues to estimate the human pose by performing localization on 14 key body points per person. Finally, the human pose can be estimated by obtaining a skeleton model for the detected human in the image using the ordered pairs of detected 2D keypoint coordinates.

The remainder of the paper is organized as follows. In Section 2, we introduce various researches for estimating the human pose based on deep learning in the field of HAR and explain the problem to be solved through the proposed method in this paper. In Section 3, we present an overview of the proposed 2D human pose estimation based on object detection using RGB-D information. The main process consists of two phases: object detection and human pose estimation. In Section 4, object detection method using RGB-D information is explained. In Section 5, a method for estimating 2D human pose is discussed. In Section 6 the experimental results of object detection and human pose estimation are included. The conclusion and further works to be studied more is discussed in section 7.

2. Related Work

In the field of video surveillance, development of artificial intelligence technology has enabled various intelligent image analysis by applying deep learning based learning method. The HAR field recognizing the behavior of objects such as human being can recognize various human behaviors by applying intelligent image analysis technology using deep learning. Therefore, in order to recognize such human behavior, the researches for object detection, body part detection, and human pose estimation have been introduced [10-13].

Finding people in images through object detection is an important issue in many applications [40, 41]. Especially, it is important to detect and analyze motion of objects such as pedestrians in the image. However, detecting a specific object among the many moving objects remains an unresolved problem due to the occlusion between them since 2D images lack topological information. In this paper, we propose the approach to solve this problem by adding depth information to the conventional image data set.

The main issues of human pose estimation research for recognizing human behavior are as follows. In computer vision, human beings can be regarded as objects of joints composed of moving parts that are connected to each other at specific joint points. Therefore, human pose estimation aims at extracting representative keypoints such as the joints of body parts from the feature of images. The human pose extracted through this method can be used to analyze human behavior in smart surveillance systems, control avatar motions in realistic animations, and interact with computers [14, 15, 16].

The human pose estimation algorithm can be divided into a bottom-up approach and a top-down approach [4, 8, 17, 18, 19]. In the algorithm, the image is regarded as the lowest level and the human pose configuration is considered as the higher level. In addition, recognizing human behavior is considered to be at a higher level. Firstly, a bottom-up pose estimation algorithm forms a feature by collecting pieces of evidence for estimating the pose from the image. This method fuses low-level image evidence to extract high-level features. Meanwhile, the top-down pose estimation algorithm estimates the pose by collecting low-level image evidence at a high level. A common approach in both methods is to use a human detector and perform a single person pose estimation for each detection. Also, in order to improve the estimation accuracy, an attempt to combine an object detection and a recognition method has been made [22].

Recently, the Robotics Institute at Carnegie Mellon University has released the OpenPose library, the first real-time system to detect 130 keypoints representing multiple bodies, hands, and faces in a single image [20]. This system detects keypoint by learning models based on CPM [8]. The OpenPose library uses a bottom-up approach to learn the association between detected keypoints based on CPM [8, 19, 21]. When a CPM is applied to an image of several persons to detect keypoints, only a keypoint for one person is detected, and detection is not performed for the rest of the persons. In this research, we have attempted to detect a target object by combining the CPM that detects only this single body keypoint with a method of detecting the accurate region of the object. This approach can be classified into a top-down approach by detecting the target object first and then detecting the location of the keypoints that represent each body part within the detected region.

In this paper, we propose a 2D pose estimation method based on object detection using RGB-D information. We adopt a top-down scheme and propose a 2D pose estimation method that combines an object detection process and a CPM structure. Usually, in the process of detecting an object, the occlusion between objects misleads the inaccurate object detection. In

order to resolve the problem we attempted to perform object detection by adding depth information.

3. The Proposed Approach for 2D Human Pose Estimation

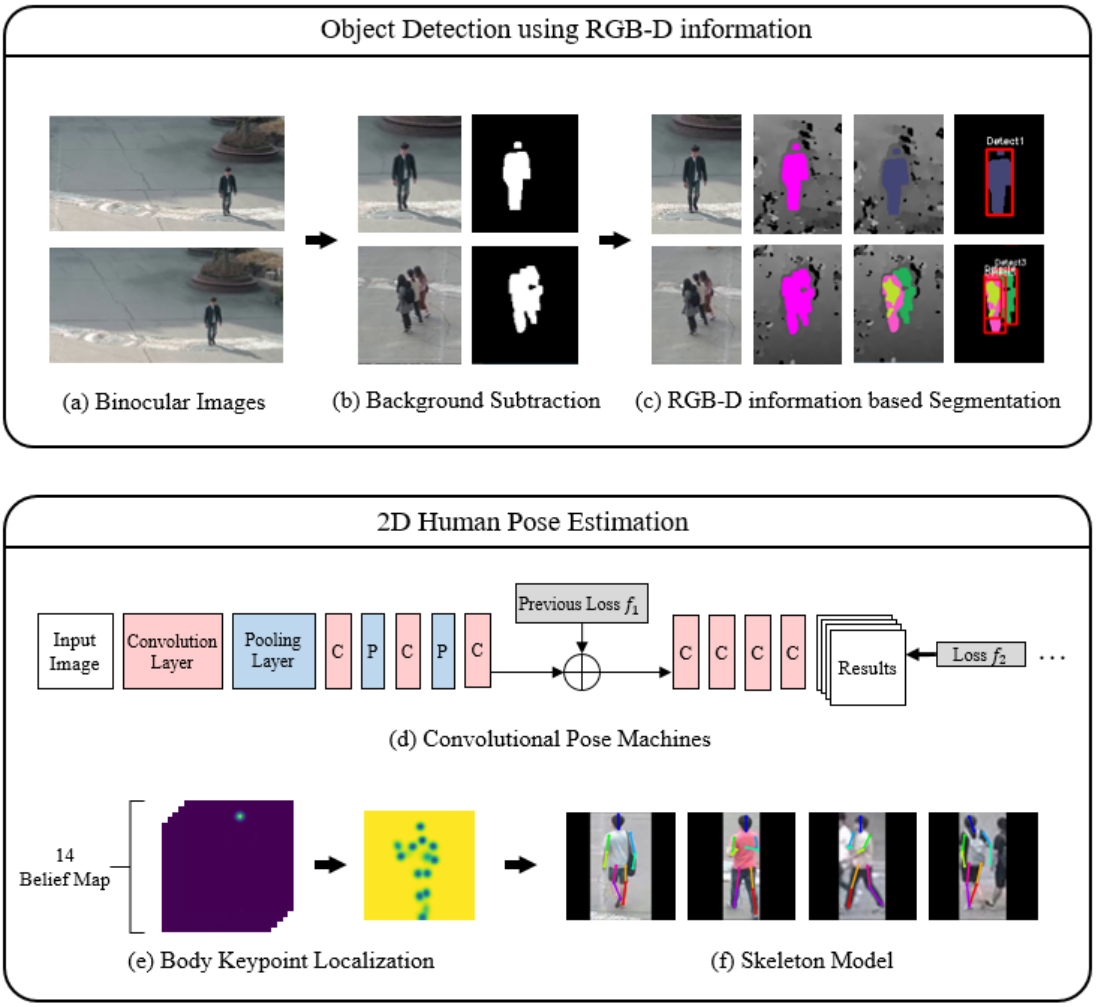


Fig. 1. Overview of the proposed approach

In order to recognize and analyze the human behavior in the video surveillance system, the process of estimating human pose should be preceded. The human pose can be estimated by constructing a skeleton model for expressing human appearance and behavior in the image. In this paper, we propose a 2D human pose estimation method based on object detection using RGB-D information. The overall procedure of the proposed method in this paper is shown in **Fig. 1**, which explains the method of object detection using RGB-D information and the method of 2D human pose estimation process.

In this research, two CCTV cameras installed horizontally are used to obtain depth information. First, the left and right images are obtained from two CCTVs as shown in **Fig. 1-(a)**. Then, in **Fig. 1-(b)**, the foreground is segmented from the background using the Mixture of Gaussians technique on the original image [23]. This method is used to perform background

modeling and to perform a difference operation on moving objects from the background model. Depth information-based segmentation is performed within the segmented region using RGB information. In this case, the depth information can be obtained by generating the depth map by calculating the disparity between the two images as shown in **Fig. 1-(c)** [25].

After the object detected based on the RGB-D information, the detected area is rescaled and cropped to generate input data. As illustrated in **Fig. 1-(d)**, the obtained single image is applied to the CPM and it generates 14 belief maps to detect each body part of a person while iterating through the output steps of the network [8, 16, 33]. The structure of the network is described in detail in Section 5. The belief map in **Fig. 1-(e)** shows the body part prediction result and subsequently the position coordinates of the keypoint representing the human joint can be obtained based on the belief map. Finally, as shown in **Fig. 1-(f)**, the human pose estimation is completed by making the skeleton model using the detected keypoint. In section 4, the object detection process using RGB-D information are expalined in detail. In section 5, the human pose estimation process of constructing a skeleton model using 2D keypoints is described.

4. Object Detection using RGB-D information

In the intelligent video surveillance system, robust detection of moving objects need to necessary to recognize human behavior. However, the loss of topological information occurs in the process of converting the 3D real world into 2D image information through the camera. As shown in **Fig. 2**, the loss of this information can mislead to the inaccurate object detection, thereby providing inaccurate information to the user.

Fig. 2. Occlusion Problem

Therefore, in this paper, object detection process is performed by adding depth information to existing RGB information to robustly detect an object. Detecting an object based on the RGB-D information is divided into two steps such as a process of segmenting an object from the background using RGB information and a process of performing a further segmentation using depth information on the preliminary segmented result based on RGB.

4.1 Background Subtraction

One of the simplest techniques for object detection is background subtraction, which identifies regions of interest by segmenting moving objects from the background. A moving object is segmented by extracting a region that changes over time from a background image. The process of detecting an object from the background is processed by subtracting the foreground model from the background model. First, the CCTV left image is used for background modeling. An initial background model is calculated using a Mixture of Gaussians technique which indicates the color distribution of each pixel point [23, 43]. In the Adaptive Mixture of Gaussian model, the following equation (1) is the probability of a pixel with intensity I_t at

time t .

$$P(I_t) = \sum_{i=1}^{K} w_{i,t} \cdot \eta(I_t, \mu_{i,t}, \sum_{i,t}) \tag{1}$$

Each gaussian distribution is sorted in descending ordey by the weight $w_{i,t}$, and if B gaussian distributions in K gaussian distributions are called background models, the following equation (2) is satisfied.

$$B = \arg\min_b (\sum_{K=1}^{b} w_k > T) \tag{2}$$

This background model is continuously updated over time to perform background modeling. The moving object is segmented by subtracting the moving object from the modeled background. In addition, noise is removed by performing a morphological operation to detect only objects of interest [24]. **Fig. 3** shows the result of segmentation using only RGB color information. The left color images are input images and the right binary images are the results of background subtraction.

Fig. 3. Background Subtraction

4.2 RGB-D information based Segmentation

In order to solve the problem shown in **Fig. 2**, in this paper, object detection is performed by utilizing depth information. The depth information can be obtained by calculating the disparity between the left and right images taken from two CCTVs installed horizontally [25]. The diparity of the images can be calculated by matching the feature points representing the same point on the two images in block units. The depth value calculated through the stereo-based block matching is represented by a value from 0 to 255, and the depth map can be generated based on the depth value.

The depth value of the generated depth map is used to perform a further division in the region segmented by RGB. Therefore, the region segmented in the previous step is set as the domain from which the depth value is to be extracted. The depth value is sequentially searched in the domain to compare the previous pixel value with the current pixel value. Since the same object can be segmented into different objects or other objects can be segmented into the same object, a certain depth value range is set to perform the segmentation. If it is included in a certain range, it is regarded as the same object, otherwise it is regarded as another object. The

object is segmented by adding depth information to the RGB information, and the detection result is represented by performing labeling based on the center point of each segmented region.

The result of object detection using RGB-D information is shown in **Fig. 4**. In **Fig. 4-(a)** shows the original image, **Fig. 4-(b)** is a depth map, **Fig. 4-(c)** shows the RGB-based segmentation result on the depth map, **Fig. 4-(d)** is the result of segmenting using the depth information in the set domain, and **Fig. 4-(e)** is the labeled object detection result.

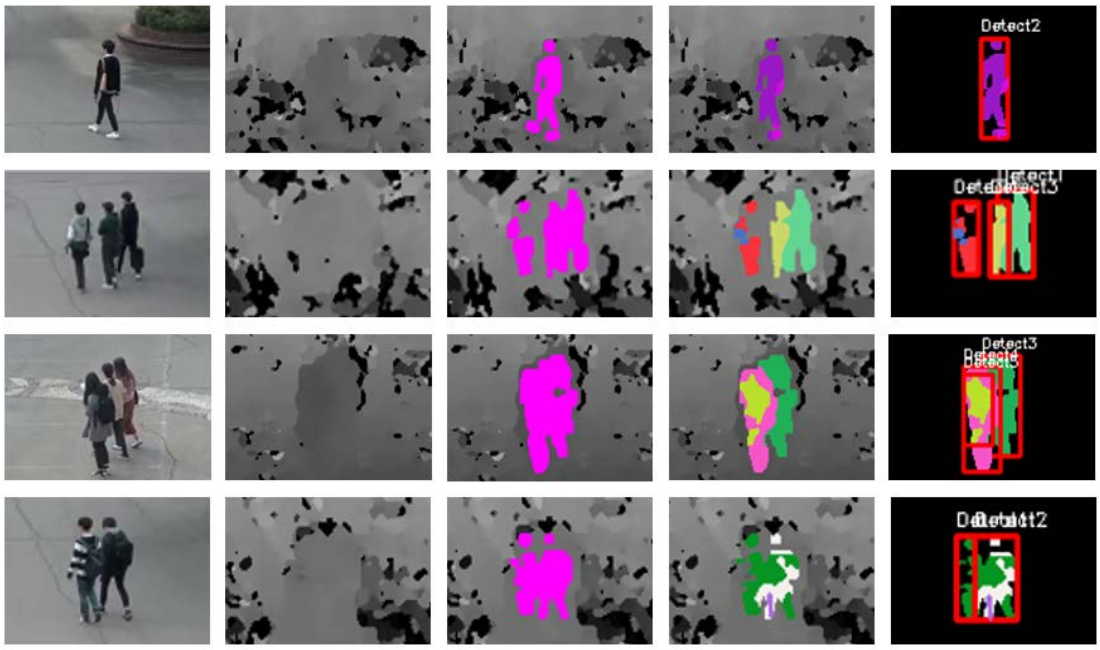


Fig. 4. Object Detection Result using RGB-D information

5. 2D Human Pose Estimation

In this research, we tried to estimate the human pose by first obtaining the region containing human and secondly executing the top-down approach of estimating the pose inside the region. The problem of estimating the human pose can be defined as the problem of detecting the position of the human joint [26, 46, 47]. Therefore, the human pose estimation process can be performed by detecting the keypoint representing the position of the joint.

In this work, we represent the position of the joint for the pose estimation through 2D coordinates, and the body estimated from the predicted joint position is represented by the skeleton model. The models that follow this skeletal structure are called kinematic chain models [27]. The kinematic model provides enough information to estimate the human pose. The kinematic model allows us to incorporate prior beliefs about joint angles. To perform human pose estimation based on this model, CPM [8], which is an pose estimator, was used in this paper. In Section 5, we describe the 2D human pose estimation process in five phases: Image Rescale and Crop, ConVol.utional Pose Machines, Generating Belief Map, Body Keypoint Localization, and Skeleton Model.

5.1 Images Rescale and Crop

According to the top-down approach, the human pose can be estimated in the region of the detected object using the RGB-D information. Information on the body part can be obtained by providing the pose estimator with an accurate region for estimating the human pose. The region of the detected object contains only one person and estimates the pose for only one person respectively. Therefore, the input data is generated by using the region of the detected object by the RGB-D information. Also, since the detected regions have different sizes, an input image is generated through the process of rescaling and cropping the region of the obejct [17, 28]. The process is as follows. Based on the RGB-D-based object's center, the image is cropped based on the labeled region. While the aspect ratio of the detected object is maintained, padding is added to the margin portion of the image to generate input data with a resolution of 100x100 pixels. The input data of the generated single image becomes an input to CPM. The result of rescaling and cropping the detected area is shown in **Fig. 5**.

Fig. 5. Rescaled and cropped images from the object regions

5.2 ConVol.utional Pose Machines

CPM is a CNN based sequential prediction framework which returns precisely N belief maps (or heatmaps) for individual body joints [7]. The CPM was pre-trained using the MPII Human Pose Dataset, which contains over 25,000 images annotated body parts and contains over 400 human activities [8, 9]. CPM is a state-of-the-art pose estimation system, achieving 88.5% PCKh in the MPII data set. The standard Percentage of Correct Keypoints (PCK) metric represents the percentage of detections that fall within a normalized distance of the ground truth. For MPII dataset, distance is normalized by a fraction of the head size (referred to as PCKh) [32, 42].

CPM is a pose estimator that implements the functions of existing Pose Machine architecture using CNN [7, 8, 32]. Pose Machine consists of two steps to predict the body part, $g_t(\cdot)$, that are trained to predict the location of each part in each level of the hierarchy. In the first step, the estimators produce an estimate of the confidence of each body part location based on the features computed in the image. In the next step, the estimators refine these confidences using additional information from the results of the previous step via the feature function ψ [7].

The $p-th$ is a pixel location of anatomical landmark, $Y_p \in Z \in \Re^2$, where Z is the set of all (u,v) location $Y=(Y_1,...,Y_P)$ for all P part. In each stage $t \in \{1...T\}$, the classifiers g_t predict beliefs for assigning a location to each part, based on features extracted from the image at the location z [7, 8]. X_z is the feature vector in the image patch for the t level of the

hierarchy centered at location Z in the image. In the first stage $t=1$, therefore produces the following belief values:

$$g_1(\mathrm{X}_z) \rightarrow \{b_1^p(Y_p = z)\}_{p\in\{0\ldots P\}} \tag{3}$$

In the equation (3), $b_1^p(Y_p = z)$ is the score predicted by the classifier g_1. All the beliefs of part p are evaluated at every location $z=(u,v)^T$ in the image.

$$b_t^p[u,v] = b_t^p(Y_p = z) \tag{4}$$

In subsequent stages, the classifier predicts a belief for assigning a location to each part $Y_p = z$, based on features of X_z and feature computed on the confidences via the context feature function Ψ for each part in the previous stage. The context feature function Ψ at a location z takes the confidence maps for the location of each part as input and generates features extracted at location z of the confidence map b_t^p. The context feature is a concatenation of scores at location z extracted from the confidence maps of all parts [7]. At each stage, the belief provides an refined estimate of the location of each part p.

$$g_t(\mathrm{X}_z^{'}, \Psi_t(z, b_{t-1})) \rightarrow \{b_t^p(Y_p = z)\}_{p\in\{0\ldots P+1\}} \tag{5}$$

CPM is the replacement of these Pose Machine functions with a conVol.utional architecture. The CPM architecture used in this work is shown in **Fig. 6**.

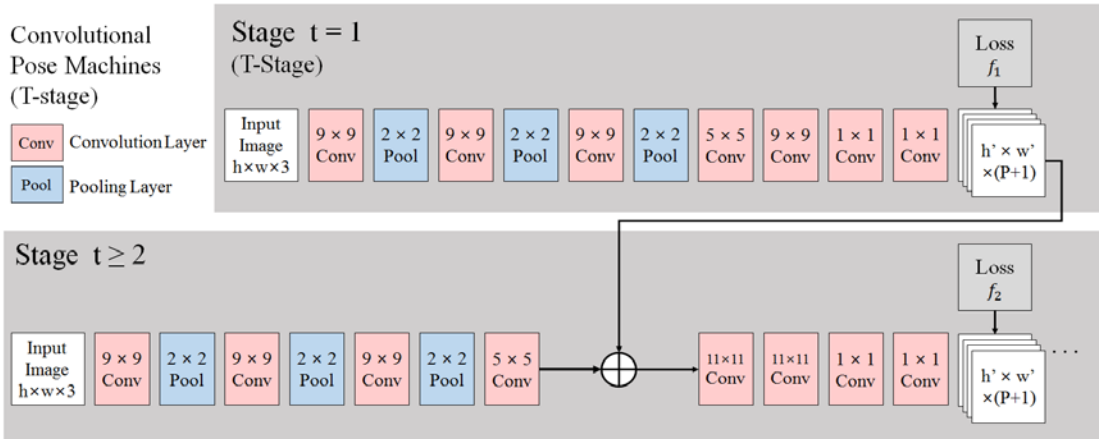


Fig. 6. Architecture of ConVol.utional Pose Machines

The CPM architecture consists of several steps $t\in\{1,\ldots,T\}$ to generate a belief map. The predicted results of the previous step are used in the next step to predict the body part p sequentially. Thus, in step t, it is trained to repeatedly generate a belief map for body part p to predict position $Y=(Y_1,\ldots,Y_P)$ on the image of all body parts P. An ideal belief map $b_*^p(Y_p = z)$ for body part p at all positions $z=(u,v)^T$ (Z is the set of all positions (u,v) of the

image, $\forall_z \in Z$) in an image is generated by placing a Gaussian peak in the ground truth of each body part p. Also, by defining the loss function as follows in the output of each step t that minimizes the distance between the predicted result for each body part and the ideal belief map, the network is repeatedly approached to the correct body part position. Therefore, the loss function to be minimized in the result of each step is shown in equation (6) (P parts plus one for background).

$$f_t = \sum_{p=1}^{P+1} \sum_{z \in Z} \| b_t^p(z) - b_*^p(z) \|_2^2$$

Using this loss function, the positional coordinates of the correct body part are obtained by adding the losses at each step. This is shown in equation (7).

$$F = \sum_{t=1}^{T} f_t \tag{7}$$

In this work, the pose was estimated using a model pre-trained by CPM. The CPM network structure consists of a conVol.utional layer that extracts only meaningful features from the input image and a pooling layer that performs subsampling to reduce the extracted features [29]. The activation function for extracting meaningful output values uses a ReLU (Rectified Linear Unit) function that outputs 0 if the input value is smaller than 0 and outputs the input value when it is larger than 0 [30, 44].

The image input to the network structure of the CPM is normalized with a resolution of 368×368 pixels, and the filter size and stride (2×2) are downsampled by 8 multiples after passing through the Max-Pooling Layer three times. The input image ($h = 368, w = 368$) is downsampled to a resolution of 184×184, 92×92, 46×46 pixels. When the feature is extracted through downsampling, the body part can finally be predicted.

5.3 Generating Belief Map

The CPM returns N belief maps for individual body joints, which can predict the position of keypoints representing the body joints through the belief map and estimate the human joint. Each stage of CPM is trained to generate a belief map repeatedly for each part location. The belief map is generated sequentially through the loss function at the output of each output stage t of the CPM [8, 16, 33]. A belief map for each generated body part p represents the body part prediction result.

In this process, the points displayed on each belief map represent the maximum height values of the confidence generated by the CPM, and the belief maps representing the maximum values of each confidence become the predicted results for each body region. In the distribution of confidence values for the belief map, the x and y axes represent the resolution of 368×368, the input image size of the CPM, and the z axis represents the confidence value. The distribution of confidence values for the 14 keypoints representing human joints to generate a belief map is shown in **Fig. 7**. In the figure, yellow indicates a high confidence value, and purple indicates a low confidence value.

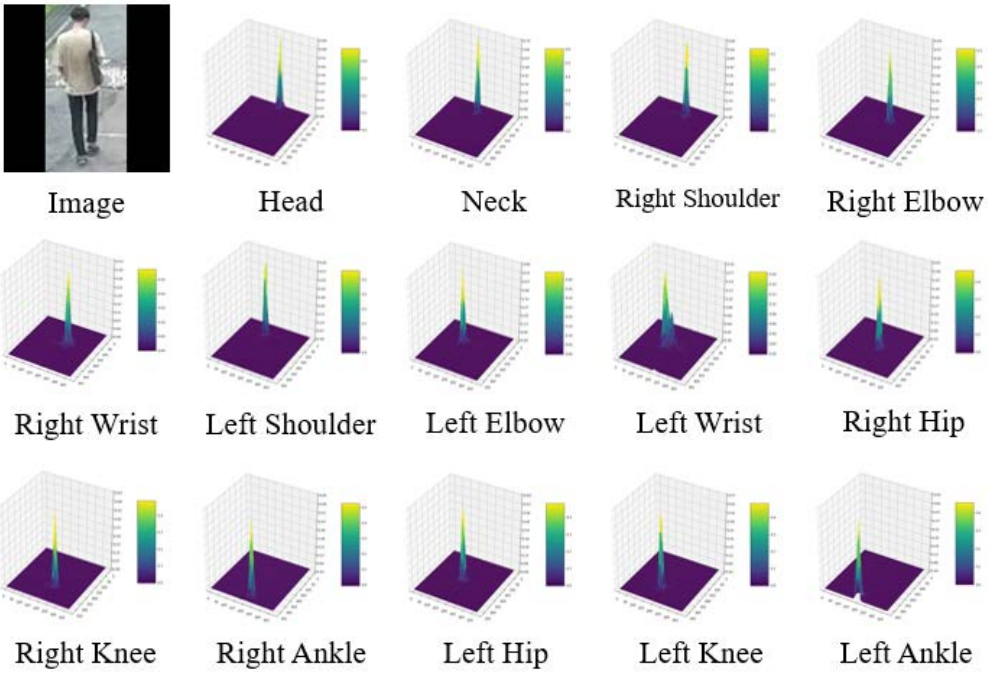


Fig. 7. Distribution of confidence values

In this paper, the confidence value is predicted by using a model pre-trained by CPM on the image of the detected region rescaled using RGB-D information. In addition, the coordinate point of the keypoint can be calculated using the maximum value of the confidence value predicted for each joint region of the body [8, 31, 32]. The belief map showing 14 keypoints per person is illustrated in **Fig. 8**. In this figure, 2D belief map was generated using the distribution of confidence values in **Fig. 7**. 2D Human Pose Estimation can be performed through the result of the keypoint localization using 2D belief map.

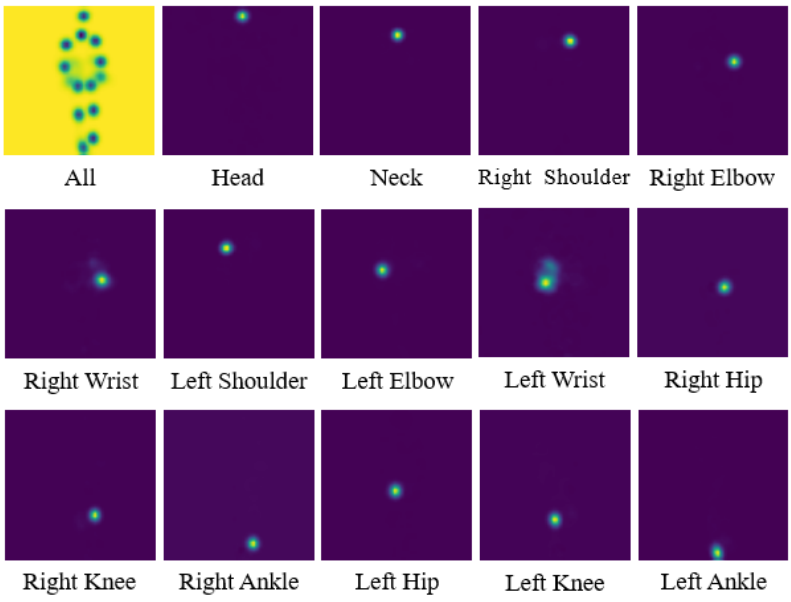


Fig. 8. Belief Map

5.4 Body Keypoint Localization

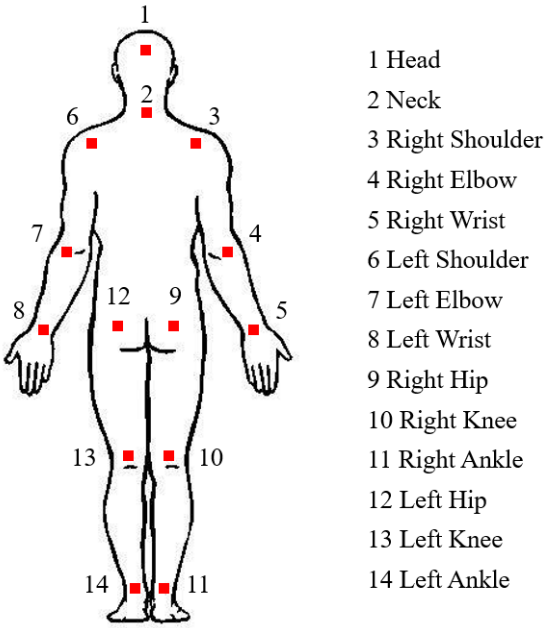


Fig. 9. Keypoints

The generated belief map represents the location of the joint, which is a body part useful for estimating the human pose. In this research, 14 keypoints were set by dividing the position of the joint into Head, Neck, Shoulder, Elbow, Wrist, Hip, Knee, and Ankle as shown in **Fig. 9**. In order to detect the position of the keypoint using the belief map generated by the CPM, the position of the body keypoint can be predicted by calculating the centeral location of the point represented in the belief map [8, 31]. There are 14 keypoints per person. Keypoints: Head, Neck, Right Shoulder, Right Elbow, Right Wrist, Left Shoulder, Left Elbow, Left Wrist, Right Hip, Right Knee, Right Ankle, Left Hip, Left Knee, Left Ankle. The results of Body Keypoint Localization are shown in **Fig. 10**.

Fig. 10. Body Keypoint Localization

5.5 Skeleton Model

In estimating the pose for HAR, we use CPM to generate a belief map through each output stage. The generated belief map represents the predicted position of the keypoint, and the coordinate order pair of the keypoint can be calculated using the prediction result. Therefore, a skeleton model for representing the human pose can be generated by using the detected 2D keypoint ordered pair (x, y). Such a skeleton model is capable of 2D human pose estimation. The result of 2D human pose estimation based on RGB-D information is finally shown in **Fig. 11**.

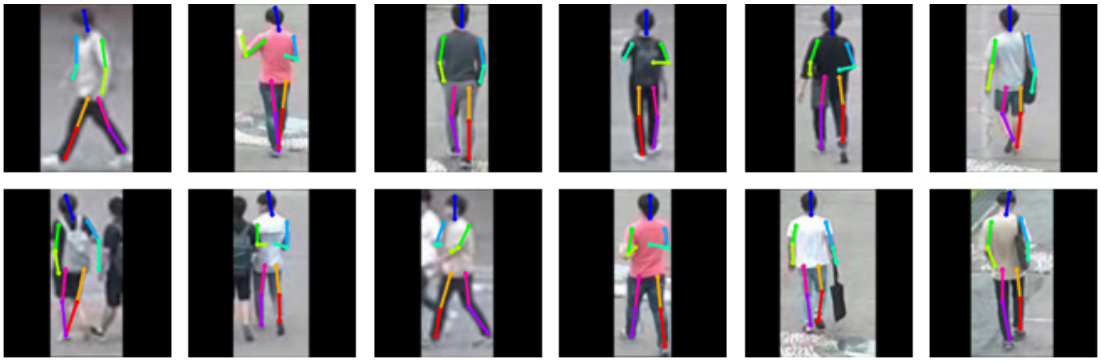
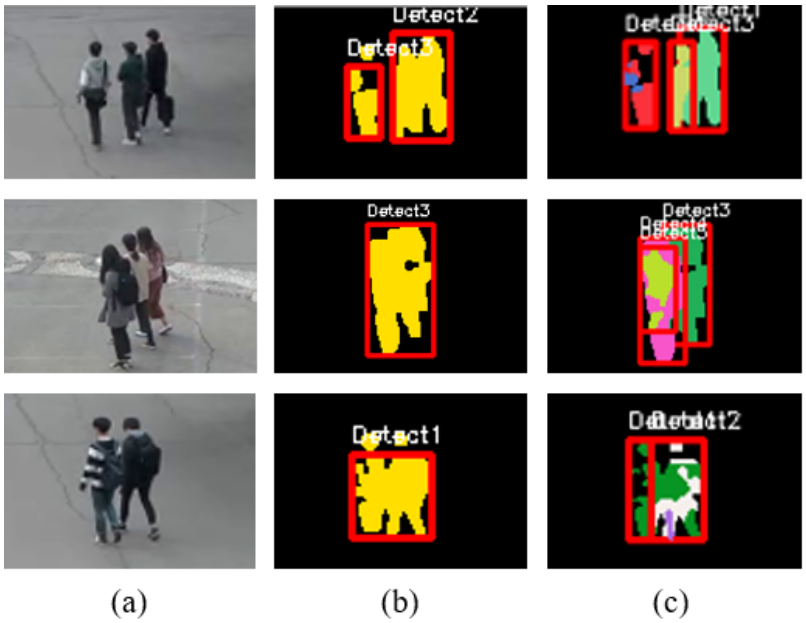


Fig. 11. Skeleton Model

6. Experimental Results

In the experiments, 800×450 pixel images from 3D CCTV are used. Additionally, experiments were conducted using TensorFlow which is an open source library of deep running [45]. As for the experimental environments, AMD Ryzen 7 1700 3GHz CPU, 16GB RAM, NVIDIA GeForce GTX 1080 Ti GPU, and Windows 10 are utilized.



(a) (b) (c)

Fig. 12. The result of comparing object detection method

Fig. 12 shows the comparative results between object detection method using only RGB information and object detection method using RGB-D information. **Fig. 12-(a)** is input image. **Fig. 12-(b)** is object detection using only RGB information. **Fig. 12-(c)** is object detection based on RGB-D information. From the experiment, we can support the efficiency of the proposed approach that is possible to detect objects robustly by resolving occlusion problem due to lack of topological information in RGB model.

In this work, object detection was performed using the top-down approach, and then the human pose was estimated within the detected region. As a result, the accurate region of the target object is provided to the pose estimator using the RGB-D information, and the localization can be estimated by detecting the keypoint position. The results of the 2D human pose estimation are shown in **Fig. 13**. (Top) Input image of CPM. (Middle) Keypoint predictions using Belief Maps. (Bottom) 2D Human Pose Estimation.

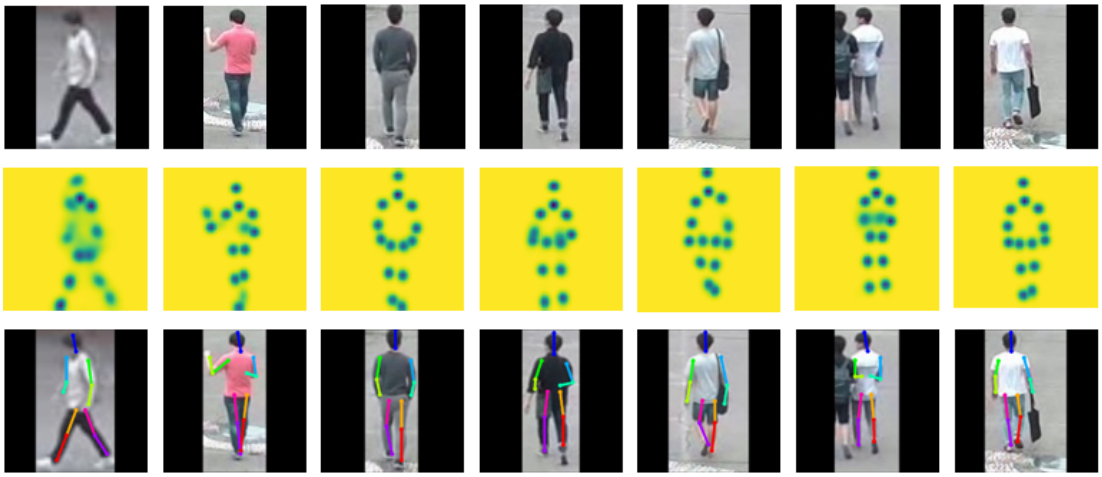


Fig. 13. The results of Human Pose Estimation based on Object Detection using RGB-D information

7. Conclusion and Future Works

In this paper, we proposed a 2D human pose estimation method based on object detection using RGB-D information. Based on the proposed approach, robust object detection was acquired using depth information, and pose estimation was performed within the region of the detected object. In the pose estimation process, keypoint prediction were obtained by sequentially generating belief maps using the pre-trained pose estimator CPM. Consequently, the human pose can be estimated by producing a human body model represented by a skeleton model using 14 keypoints of human joint positions.

The proposed method solves the occlusion problem by performing object detection by adding depth information, and based on this object detection method, 2D pose estimation is performed using CPM, which is an pose estimator. Through a combination of RGB-D based object detection and CPM, we attempted a top-down approach to 2D human pose estimation. The human pose estimated with the proposed approach can be used in analyzing human behavior in smart surveillance systems, controlling the motion of avatars in animation. Therefore, as for the future work, the proposed method can be extended to the research of 3D pose estimation after mapping to a 3D human model using predicted 2D key point coordinates [34-39].

References

[1] Grant, Jason M., and Patrick J. Flynn, "Crowd Scene Understanding from Video: A Survey," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, Vol.. 13, No. 2, pp. 19, 2017. Article (CrossRef Link)

[2] Paul, Manoranjan, Shah ME Haque, and Subrata Chakraborty, "Human detection in surveillance videos and its applications-a review," EURASIP Journal on Advances in Signal Processing, Vol. 176, No. 1, pp.1-16, 2013. Article (CrossRef Link)

[3] San, Phyo P., et al, "DEEP LEARNING FOR HUMAN ACTIVITY RECOGNITION," 2017. Article (CrossRef Link)

[4] Gong, Wenjuan, et al, "Human Pose Estimation from Monocular Images: A Comprehensive Survey," *Sensors*, Vol. 16, No. 12, pp. 1-39, 2016. Article (CrossRef Link)

[5] Zhang, Shugang, et al, "Vision-Based Human Activity Recognition: A Review," *Journal of Healthcare Engineering*, Vol. 2017, pp. 1-31, 2017. Article (CrossRef Link)

[6] Vrigkas, Michalis, Christophoros Nikou, and Ioannis A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, Vol. 2, article 28, 2015. Article (CrossRef Link)

[7] Ramakrishna, Varun, et al., "Pose machines: Articulated pose estimation via inference machines," in *Proc. of European Conference on Computer Vision*, pp. 33-47, 2014. Article (CrossRef Link)

[8] Wei, Shih-En, et al, "ConVol.utional pose machines," in *Proc. of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724-4732, 2016. Article (CrossRef Link)

[9] Andriluka, Mykhaylo, et al, "2d human pose estimation: New benchmark and state of the art analysis," in *Proc. of Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686-3693, 2014. Article (CrossRef Link)

[10] Poppe, Ronald, "A survey on vision-based human action recognition," *Image and vision computing*, Vol. 28, No. 6, pp.976-990, 2010. Article (CrossRef Link)

[11] Weinland, Daniel, Remi Ronfard, and Edmond Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer vision and image understanding*, Vol. 115, No. 2, pp. 224-241, 2011. Article (CrossRef Link)

[12] Aggarwal, Jake K., and Michael S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, Vol. 43, No. 3, pp. 16, 2011. Article (CrossRef Link)

[13] Chen, Lulu, Hong Wei, and James Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, Vol. 34, No. 15 pp. 1995-2006, 2013. Article (CrossRef Link)

[14] Presti, Liliana Lo, and Marco La Cascia, "3D skeleton-based human action classification: A survey," *Pattern Recognition*, Vol. 53, pp. 130-147, 2016. Article (CrossRef Link)

[15] Chen, Wenzheng, et al, "Synthesizing training images for boosting human 3d pose estimation," in *Proc. of 3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, pp. 479-488, 2016. Article (CrossRef Link)

[16] Tome, Denis, Chris Russell, and Lourdes Agapito, "Lifting from the deep: ConVol.utional 3d pose estimation from a single image," *arXiv preprint arXiv:1701.00295*, 2017. Article (CrossRef Link)

[17] Papandreou, George, et al, "Towards Accurate Multi-Person Pose Estimation in the Wild," *arXiv preprint arXiv:1701.01779*, 2017. Article (CrossRef Link)

[18] Insafutdinov, Eldar, et al, "Articulated multi-person tracking in the wild," *arXiv preprint arXiv:1612.01465*, 2016. Article (CrossRef Link)

[19] Cao, Zhe, et al., "Realtime multi-person 2d pose estimation using part affinity fields," arXiv preprint arXiv:1611.08050, 2016. Article (CrossRef Link)

[20] OpenPose: A Real-Time Multi-Person Keypoint Detection and Multi-Threading C++ Library, 2017. Article (CrossRef Link)

[21] Simon, Tomas, et al, "Hand Keypoint Detection in Single Images using Multiview Bootstrapping," *arXiv preprint arXiv:1704.07809*, 2017. Article (CrossRef Link)

[22] Dimitrijevic, Miodrag, Vincent Lepetit, and Pascal Fua, "Human body pose detection using bayesian spatio-temporal templates," *Computer vision and image understanding*, Vol. 104, No. 2, pp. 127-139, 2006. Article (CrossRef Link)
[23] Zivkovic, Zoran, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. of* Proceedings of the 17th International Conference on Pattern Recognition, pp. 28-31, 2004. Article (CrossRef Link)
[24] Niblack, Wayne, "An introduction to digital image processing," *Strandberg Publishing Company*, 1985. Article (CrossRef Link)
[25] Hirschmuller, Heiko, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 30, No. 2, pp. 328-341, 2008. Article (CrossRef Link)
[26] Toshev, Alexander, and Christian Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proc. of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653-1660, 2014. Article (CrossRef Link)
[27] Lehrmann, Andreas M., Peter V. Gehler, and Sebastian Nowozin, "A non-parametric bayesian network prior of human pose," in *Proc. of Proceedings of the IEEE International Conference on Computer Vision*, pp. 1281-1288, 2013. Article (CrossRef Link)
[28] Linna, Marko, Juho Kannala, and Esa Rahtu., "Real-time human pose estimation from video with conVol.utional neural networks," *arXiv preprint arXiv:1609.07420*, 2016. Article (CrossRef Link)
[29] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep conVol.utional neural networks," *Advances in neural information processing systems*, pp. 1097-1105, 2012. Article (CrossRef Link)
[30] Dahl, George E., Tara N. Sainath, and Geoffrey E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, pp. 8609-8613, 2013. Article (CrossRef Link)
[31] Belagiannis, Vasileios, and Andrew Zisserman, "Recurrent human pose estimation," in *Proc. of Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on. IEEE*, pp. 468-475, 2017. Article (CrossRef Link)
[32] Chen, Ching-Hang, and Deva Ramanan, "3D Human Pose Estimation = 2D Pose Estimation + Matching," *arXiv preprint arXiv:1612.06524*, 2016. Article (CrossRef Link)
[33] Rafi, Umer, et al, "An Efficient ConVol.utional Network for Human Pose Estimation," *BMVC*, Vol. 1, 2016. Article (CrossRef Link)
[34] Bogo, Federica, et al, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. of European Conference on Computer Vision*, Springer International Publishing, pp. 561-578, 2016. Article (CrossRef Link)
[35] Lassner, Christoph, et al, "Unite the People: Closing the Loop Between 3D and 2D Human Representations," *arXiv preprint arXiv:1701.02468*, 2017. Article (CrossRef Link)
[36] Martinez, Julieta, et al, "A simple yet effective baseline for 3d human pose estimation," *arXiv preprint arXiv:1705.03098*, 2017. Article (CrossRef Link)
[37] Ramakrishna, Varun, Takeo Kanade, and Yaser Sheikh, "Reconstructing 3d human pose from 2d image landmarks," *Computer Vision–ECCV 2012*, pp.573-586, 2012. Article (CrossRef Link)
[38] Loper, Matthew, et al, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics (TOG)*, Vol. 34, No. 6, pp.248:1-248:16, 2015. Article (CrossRef Link)
[39] Mehta, Dushyant, et al, "VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera," *arXiv preprint arXiv:1705.01583*, 2017. Article (CrossRef Link)
[40] Enzweiler, Markus, and Dariu M. Gavrila, "Monocular pedestrian detection: Survey and experiments," in *Proc. of IEEE transactions on pattern analysis and machine intelligence*, Vol. 31, No. 12, pp. 2179-2195, 2009. Article (CrossRef Link)
[41] Hu, Weiming, et al, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 34, No. 3, pp. 334-352, 2004. Article (CrossRef Link)

[42] Newell, Alejandro, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *Proc. of European Conference on Computer Vision*, Springer International Publishing, pp. 483-499, 2016. Article (CrossRef Link)
[43] Zivkovic, Zoran, and Ferdinand Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern recognition letters*, Vol. 27, No. 7, pp. 773-780, 2006. Article (CrossRef Link)
[44] Nair, Vinod, and Geoffrey E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. of Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807-814, 2010. Article (CrossRef Link)
[45] TensorFlow: An open-source software library for Machine Intelligence. Article (CrossRef Link)
[46] Uddin, Md, and Jaehyoun Kim, "A Robust Approach for Human Activity Recognition Using 3-D Body Joint Motion Features with Deep Belief Network," *KSII Transactions on Internet & Information Systems*, Vol. 11, No.2, 2017. Article (CrossRef Link)
[47] Uddin, Md, and Jaehyoun Kim, "Human Activity Recognition Using Spatiotemporal 3-D Body Joint Features with Hidden Markov Models," *KSII Transactions on Internet & Information Systems*, Vol. 10, No.6, 2016. Article (CrossRef Link)

Seohee Park received the B.S. degree in computer science from Kyonggi University, South Korea, in 2017. She is currently working toward the M.S. degree in computer science from Kyonggi University, South Korea. Her major research interests are vision-based human pose estimation, and human activity recognition.

Myunggeun Ji received the B.S. degree in computer science from Kyonggi University, South Korea, in 2017. He is currently working toward the M.S. degree in computer science from Kyonggi University, South Korea. His major research interests are computer vision, Augmented Reality.

Junchul Chun is currently a professor in the department of computer science and a principal investigator in the Graphics and Image Processing Research Laboratory at Kyonggi University, Suwon South Korea. He received a B.S. degree from Chung-Ang University majoring in computer science. He also received M.S. and Ph.D. degrees from the department of computer science and engineering at the University of Connecticut, CT., U.S.A, in 1992 and 1995, respectively. He was a visiting research scholar in PRIP (Pattern Recognition and Image Processing) at Michigan State University and University of Colorado at Boulder, 2001 and 2009 respectively. He is currently a chief director of CCSRC (Contents Convergence Software Research Center) at Kyonggi University. He served as the present of the Korean Society of Internet Information during 2015. His major research fields are augmented reality, vision-based animation, and human computer interaction.