Received: 18 September 2018 | Revised: 21 February 2019 | Accepted: 28 April 2019

DOI: 10.4218/etrij.2018-0520

ORIGINAL ARTICLE

WILEY ETRI Journal

Vision-based garbage dumping action detection for real-world surveillance platform

Kimin Yun | **Yongjin Kwon** | **Sungchan Oh** | **Jinyoung Moon** | **Jongyoul Park**

SW·Contents Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea

Correspondence
Jongyoul Park, SW·Contents Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea.
Email: jongyoul@etri.re.kr

Funding information
This research was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis).

In this paper, we propose a new framework for detecting the unauthorized dumping of garbage in real-world surveillance camera. Although several action/behavior recognition methods have been investigated, these studies are hardly applicable to real-world scenarios because they are mainly focused on well-refined datasets. Because the dumping actions in the real-world take a variety of forms, building a new method to disclose the actions instead of exploiting previous approaches is a better strategy. We detected the dumping action by the change in relation between a person and the object being held by them. To find the person-held object of indefinite form, we used a background subtraction algorithm and human joint estimation. The person-held object was then tracked and the relation model between the joints and objects was built. Finally, the dumping action was detected through the voting-based decision module. In the experiments, we show the effectiveness of the proposed method by testing on real-world videos containing various dumping actions. In addition, the proposed framework is implemented in a real-time monitoring system through a fast online algorithm.

KEYWORDS
action recognition, garbage dumping action, human-object relation, machine vision application, visual surveillance

1 | INTRODUCTION

Several vision systems are being adopted in a variety of practical applications with the aid of dissemination of portable and the recent developments in vision technologies. In particular, the advent of deep learning techniques makes vision systems more useful in real-world environments because these techniques deal with diverse real images and videos. For example, object recognition and detection models have recently been implemented in several applications, including autonomous vehicles, image/video search engines, and automatic object tagging, owing to their breakthrough performance [1–3]. Action recognition, one of the most challenging problems in computer vision, has also been investigated through deep learning techniques [4–6]. Although there are some meaningful methods for action recognition, they are not as widely used in real-world applications. Unlike an object, human actions are shown differently according to scene and camera view. Thus, building a model to represent the general behavior is not easy. Therefore, in a practical application, limited and well-refined datasets are utilized to detect the target actions. However, most action datasets are taken by the director's control, thereby complicating their application in real-world videos.

For the practical behavior detection research, we newly set up a problem of detecting illegal dumping actions in surveillance cameras. Dumping actions are frequently found in real-world videos such as a situation where a piece of trash

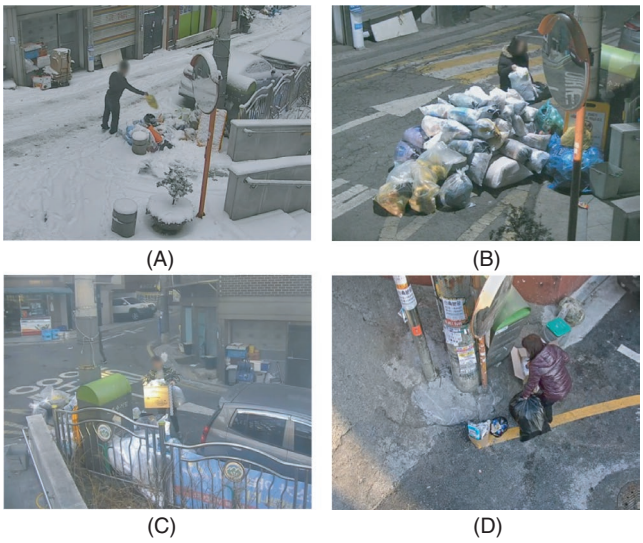


FIGURE 1 Appearance variations in real-world dumping action: (A and B) various dumping poses, (C) various types of garbage, and (D) various camera views

is thrown on the street. Thus, the detection of dumping actions can be a useful module in real-world vision systems. A similar study regarding dumping action detection is the detection of abandoned luggage through visual surveillance systems [7,8]. The abandoned luggage detection determines suspicious objects that are left unbothered for a while. There are two main differences between our problem and the abandoned luggage detection problem. First, the abandoned luggage detection problem focuses on detecting the abandoned object only after an event occurs. However, our problem detects the person who generated the event when it occurred. Second, we utilize real-world data instead of video recorded by a given scenario. In traditional data, people were found to intentionally leave objects in visible positions on the camera. In actual data, however, there is a large variation in the recorded data such as dumping spots and poses, as shown in Figure 1.

In this paper, we propose a novel framework to detect garbage dumping actions in real time. The overall structure of our scheme is shown in Figure 2, First, a foreground region, joint heatmaps, and joint positions are obtained through background subtraction and joint estimation. Then, our scheme tracks pedestrians and detects the object being carried by them. After object detection, it is tracked with a correlation filter tracker that operates in real time. Simultaneously, the distance relationship is incrementally modeled between each person's joint and the tracked object. If a change is detected considering the distance relation, our method detects the dumping action. Experimental results show the effectiveness of the proposed method by comparing the state-of-the-art and ablation studies of the proposed modules.

2 | RELATED WORKS

The field of video surveillance has been actively studied as a useful application in the field of computer vision research. To assist surveillance observers who simultaneously monitor several cameras, various fields such as foreground detection, object detection, tracking, motion analysis, action recognition, and abnormal behavior detection have been developed together [9].

In the field of action recognition, a representative research such as two-stream network [4] has been proposed and large-scale video datasets are being built [10,11]. These datasets

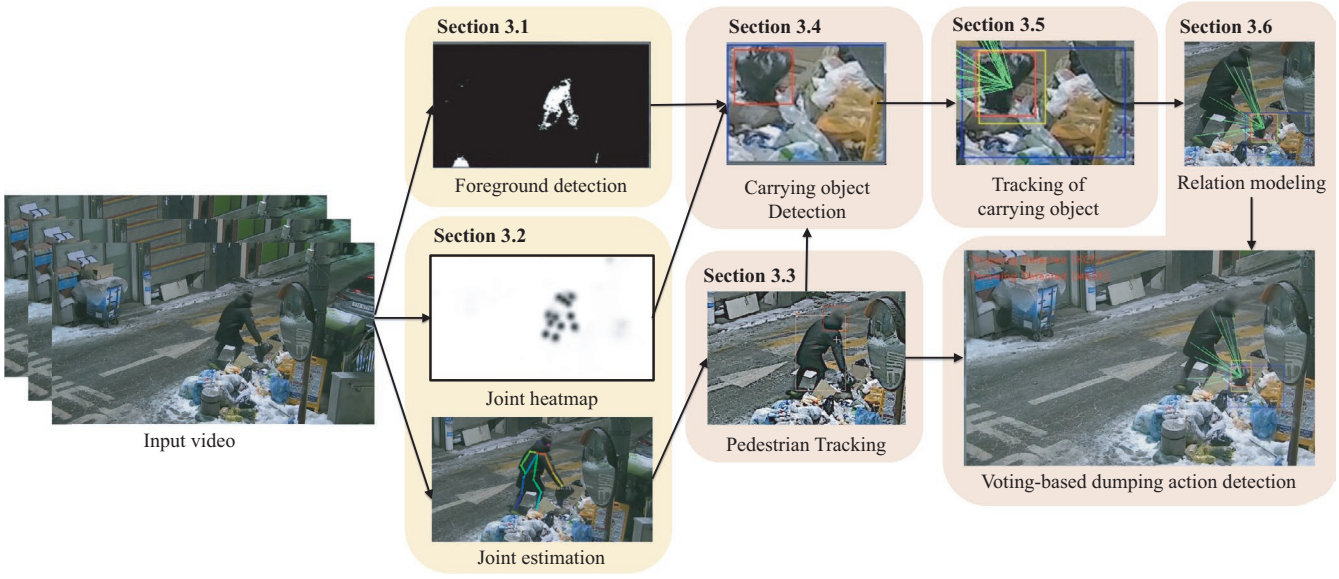


FIGURE 2 Overall framework of the proposed method

consist of movies or broadcast videos and the definition of each action is ambiguous. However, action recognition research in surveillance video focuses on well-defined actions according to the purpose of installation of the surveillance camera. Because surveillance target events such as violence and abnormal actions rarely occur, previous studies adapted the rule-based decision [12] or formulated this problem as the outlier detection problem [13,14]. This approach demonstrated the possibility of event detection when the event was difficult to model owing to a lack of data. However, if the normal events occur with the new pattern, these normal events are detected as abnormal.

Garbage dumping actions occur frequently and are considered an important event in surveillance but they have not been studied well in previous studies. There are several studies to detect the garbage dumping by establishing a specific problem, that is, throwing objects across a fence [15] or from a vehicle [16] and classifying the clean road and garbage area [17,18]. However, these methods do not use the general surveillance camera installed at a high position and detect the garbage instead of understanding why the person threw the trash.

The study on abandoned baggage detection is the most similar to our study but they focus on the detection of abandoned objects instead of detecting discarded behaviors. These methods mainly utilize two background model methods with different learning rates: a fast-adaptive background and a slow-adaptive background [7,8]. The fast-adaptive background includes the appearance of abandoned baggage but the slow-adaptive background does not include the baggage. By comparing these two backgrounds, the abandoned baggage or parked vehicle can be found. However, the determination of the object after the occurrence of an event takes some time. There are other studies [19,20] that detect surveillance activities such as loitering, falling, and fighting. However, most of them are not natural and have been concocted. Owing to this dataset bias, many false alarms occur, which interfere with the monitoring system. We design our algorithm using real data and minimize false alarms to enable practical applications.

3 | PROPOSED METHOD

The detection of garbage dumping behavior is an application of action recognition. However, the adoption of conventional action recognition methods is difficult. The data required to model the appearance characteristics of dumping behavior to utilize deep learning algorithms is insufficient. In addition, the trash discarding behavior has several appearance variations, which complicate the creation of the general appearance model. Figure 1A and 1B show the diversity of appearance in real-world dumping actions: a man throws the trash in Figure 1A and a woman puts garbage in the surrounding trash pile in Figure 1B. Figure 1C shows the shape diversity in the discarded object. It is difficult to define the general shape of the garbage because there are many types of discarded items such as paper boxes and plastic bags. In addition, depending on the installation height and angle of the camera, the appearance of the trash discarding behavior significantly changes, as shown in Figure 1D.

In addition to the difficulties resulting from the diversity of appearance, the proposed method is designed to meet the requirements of actual users. The conventional clip-based action recognition studies are not online detection methods and the abandoned luggage detection methods require a certain time for detection after the event occurs. However, in a real-time monitoring system, it is essential to notify the observer immediately after the occurrence of the event, implying that on-time detection is required instead of post-event detection. For this purpose, each module must be able to operate online and be efficient in the computation for real-time operation. Figure 2 depicts the overall framework of the proposed method. A detailed description of each module is given in the following subsection.

3.1 | Foreground detection

Because the video is composed of consecutive images, useful information can be obtained from the temporal property. Among these images, foreground detection provides information about the region where the change occurs in the video. To obtain the foreground region, we first modeled the background, which is a stable area, without changing the input frame and subtracting the input frame from the background. The latest methods using R-PCA [21,22] or CNN [23,24] exhibit good performance in foreground detection but they sacrifice speed and require scene-specific training and future frame information. Therefore, our framework adopts a scene conditional background modeling method that can handle the moving camera [25].

3.2 | Joint confidence map and joint estimation

The information about the posture and joint of a person is useful abstracted information for many applications such as human-computer interaction and behavior understanding. In particular, the extensive pose data [26,27] and state-of-the-art deep learning-based methods have enabled robust and fast joint detection [28,29]. Among them, we adopt the algorithm used by [28] called *Openpose*. This estimates a multiperson 2D pose using the multistage convolutional neural network that learns joint locations and associations. After an input image passes through this network, a joint confidence map

that has the same input size and joint coordinates with discrete positions are obtained.

3.3 | Pedestrian tracking

In the previous section, we obtained the joint confidence map and discrete joint coordinates by the pose estimation method. However, because it is a detection-based method that operates on single frame without temporal information, we cannot accurately determine the information about every person owing to false positives and false negatives (missing). Apart from the problems of missing and false alarms, pose estimation generates an output regardless of the order of existence of multiple people. To ensure that each person has the same ID over time, a multitarget tracking scheme is required. We employed the tracking-by-detection framework [30] based on the Hungarian method [31], which operates online in real time. While the original method used a full-body bounding box from the deformable part model [32], the whole-body boxes often are overlapped and occluded when people walk together. When each person's bounding box overlaps, a complex cost function for matching is required to correctly link it with the tracking process. Additionally, the size of full-body bounding box significantly changes depending on the movement of the hand and foot but the change in the size of head bounding box is relatively smaller. The position of the head bounding box has a tendency to move linearly in the direction that the person is heading to, which assists in linking the detections to the tracking process. Therefore, we used the head bounding boxes as inputs to pedestrian tracking.

Figure 3 shows the tracking performance difference according to the input bounding box type under the same matching cost function. As shown in the top row of Figure 3, when the full-body bounding box is used, the ID of the person is newly given or ID switch occurs. However, when the head bounding box is used, the ID remains unchanged after overlap occurs, as shown in the bottom row of Figure 3.

FIGURE 3 Pedestrian tracking performances: top row and bottom row show the results of using the full-body bounding box and head bounding box, respectively. When the full-body bounding box is used, the ID of the person passing in front of the camera is newly given as (A and B) and both IDs change as (C and D)

FIGURE 4 Examples of detection results of Faster R-CNN [3] using COCO dataset [26]

3.4 | Carrying object detection

The dumping action can be defined as a situation in which a person is separated from a human carrying an object. If a bounding box of a person and an object is given as the ground truth, the situation where two objects are moving away can be easily detected [12]. However, objects that people carry are difficult to detect even with state-of-the-art detection algorithms. Figure 4 shows the detection results of Faster R-CNN [3] trained on the COCO dataset [26]. Pedestrians are relatively well-detected but human-carried objects are rarely detectable because it is difficult to define the shape characteristics of humans carrying objects. In other words, because the type of garbage is diverse, the performance of garbage detection is not satisfactory owing to the intra-variation problem.

Therefore, in this study, we find the human carrying object using joint information and foreground region instead of object detection. The overall scheme is presented in Figure 5. Using the foreground image, joint position and heatmap (color reverse map of confidence map), and pedestrian tracking, we establish the initial candidate region for carrying object near each hand. Let $\mathbf{b}_{\text{init}} = (x_{\text{init}}, y_{\text{init}}, w_{\text{init}}, h_{\text{init}})$ denote the position (center coordinate, width, and height) of an initial box. The width w_{init} and height h_{init} are set to half of those of the pedestrian box as

$$w_{\text{init}} = w_{\text{person}}, \quad h_{\text{init}} = h_{\text{person}}/2. \tag{1}$$

Then, for the left hand, we set the right-top point of the initial box to be the left hand's coordinates and shifted it slightly to the right as

$$x^{L}_{\text{init}} = x^{L}_{\text{hand}} - \frac{w_{\text{init}}}{2} + \frac{w_{\text{init}}}{5}, \tag{2}$$

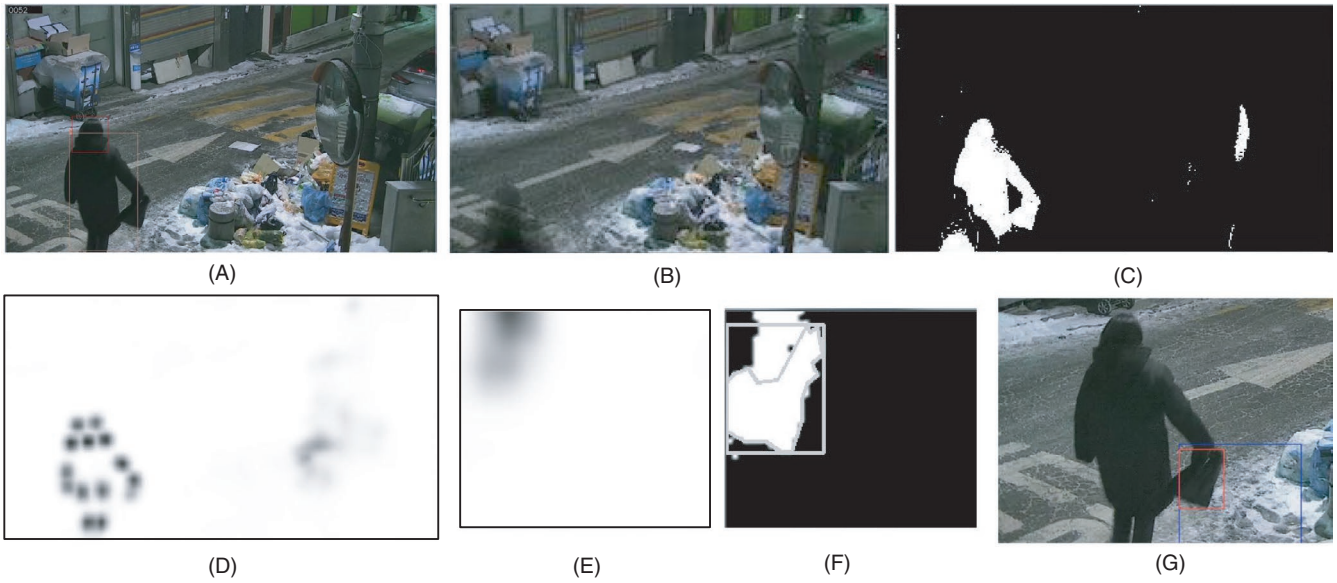


FIGURE 5 Overall procedure for carrying object detection. The joint heatmap is obtained by applying color reversal to the confidence map of the joint to facilitate the calculation: (A) Input video, (B) Background model, (C) Foreground image, (D) Joint position & heatmap info, (E) Initial candidate region near hand, (F) Refined candidate region of carrying object, and (G) Final carrying object proposal

$$y^L_{\text{init}} = y^L_{\text{hand}} + \frac{h_{\text{init}}}{2}. \tag{3}$$

In the case of right hand, we set the left-top point of the initial box to be the right hand's coordinates and slightly shifted it to the left as

$$x^R_{\text{init}} = x^R_{\text{hand}} + \frac{w_{\text{init}}}{2} - \frac{w_{\text{init}}}{5}, \tag{4}$$

$$y^R_{\text{init}} = y^R_{\text{hand}} + \frac{h_{\text{init}}}{2}, \tag{5}$$

where the superscript L and R indicate left and right, respectively. As shown in the blue rectangle in Figure 5G, this initial rectangle covers the candidate area containing a human-held object around the hand. By establishing the area for the initial bounding box to be proportional to the size of the person, the object candidate region changes according to the size of the person, which varies depending on the camera installation angle or scene depth. Here, these parameters in (1)–(5) act as a guide for the initial candidate. The establishment of the initial candidate region is heuristic but in most dumping events, people do not throw away objects bigger than their bodies. In addition, because the initial candidate region will be refined to fit the carrying object in the next step, the initial region is set to a large region to include the carried object.

Then, the scheme determines whether the initial box contains an object using the heatmap and foreground map near hand. As shown in Figure 5, heatmap and foreground patches Figure 5E and 5F are extracted at the initial box from an image-size heatmap and foreground map as Figure 5C and 5D. If a person is holding an object, an object region will appear in the foreground because objects and people move together in the video. However, because the appearances of objects are different than those of the joints, the heatmap does not contain the object region. Therefore, by comparing the foreground and heatmap, the human-held objects can be detected.

Let $H_{\mathbf{b}_{\text{init}}}$ and $F_{\mathbf{b}_{\text{init}}}$ be the heatmap patch as shown in Figure 5E and the foreground patch as shown in Figure 5F on the position of $\mathbf{b}_{\text{init}}$, respectively. Then, a patch M, which represents the foreground pixel that is not included in the human region, is obtained as

$$M(i) = \begin{cases} 1 & \text{if } H_{\mathbf{b}_{\text{init}}}(i) \cdot F_{\mathbf{b}_{\text{init}}}(i) < th_{\text{h}}, \\ 0 & \text{otherwise}, \end{cases} \tag{6}$$

where th_{h} is the threshold parameter for heatmap. In this obtained binary image M, each isolated chunk is represented by the contour and bounding rectangle as Figure 5F by finding the convex hulls [33].

Then, we selected the largest chunk, greater than the minimum size th_{area}, and the bounding box of the largest chunk was set as the refined object box $\mathbf{b}_{\text{refine}}$. By conducting the tests to determine the human joint region contained in the refined box $\mathbf{b}_{\text{refine}}$, we decided whether to use this refined box as the object position through the following equation:

$$\frac{1}{N}\sum_{i=1}^{N} H_{\mathbf{b}_{\text{refine}}}(i) > th_{\text{h}}. \tag{7}$$

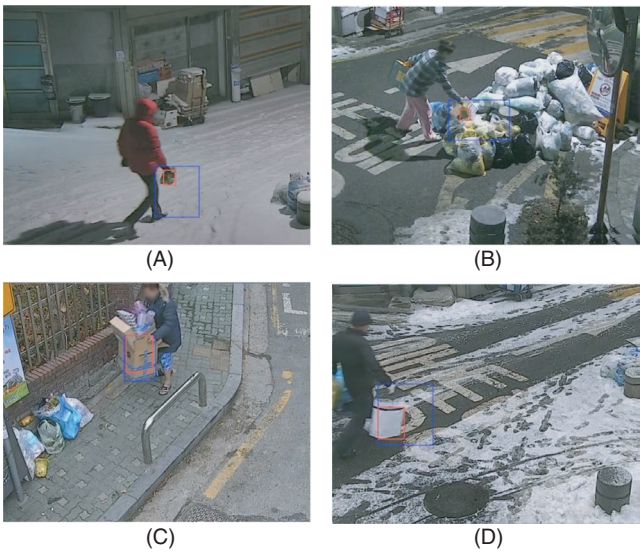


FIGURE 6 Example results of human-held object detection. The blue rectangle is the initial box $\mathbf{b}_{\text{init}}$ and the red rectangle is the refined box $\mathbf{b}_{\text{refine}}$

If the tracking of carrying object fails, the reinitialize scheme works and finds the object near the hand again. The details of the tracking of carrying object and re-initialize scheme are described in the next section. The red box in Figure 5G shows the example of refined object boxes. Through the proposed method, it is possible to detect small objects and various types of garbage that are difficult to detect with conventional object detection methods, as shown in Figure 6.

3.5 | Tracking of carrying object

Although $\mathbf{b}_{\text{refine}}$ as a human-held object rectangle is obtained, it is not efficiently detected when the confidence of hand joint is low or the foreground has some noise. To compensate for this instability, we combined single-target tracking to maintain temporal consistency in the object region. Among the various single-target tracking algorithms [34–36], we adopted the kernelized correlation filter method (KCF) [37]. This correlation filter-based tracking is very efficient as well as provides a comparable performance in comparison with deep learning-based tracking. However, single-target tracking methods are generally assumed when the position of an object is accurately given in the first frame. Instead, in our problem, the position $\mathbf{b}_{\text{refine}}$ in Section 3.4 is used for the object's initial position. We also implemented a reinitialization scheme when the tracker confidence was smaller than the given threshold, implying that our method discards the error accumulated in the tracker and finds the object using the scheme in Section 3.4, which enables robust tracking like a tracking-by-detection scheme.

3.6 | Voting-based dumping action detection

Finally, for the detection of dumping action, we want to detect the situation when the object is far from the pedestrian. The nave method is to measure the distance between a person's hand coordinates and hand-held object and if this distance is more than a certain length, it can be detected as an act of discarding. However, there are several drawbacks such as the tracking is not perfect, the position of the estimated hand is incomplete, and the length relation is not consistent depending on the camera view and carrying type.

To resolve these problems, we estimated all joints and voting-based decision based on 1D Gaussian modeling. In our method, the distance of the held object and each joint was modeled by the 1D Gaussian model (mean m_i and variance s_i^2) using the moving average as

$$m_i^{(t)} = (1-\eta)m_i^{(t-1)} + \eta d_i^{(t)}, \tag{8}$$

$$s_i^{2(t)} = (1-\eta)s_i^{2(t-1)} + \eta v_i^{(t)}, \tag{9}$$

where i indicates the joint index and $d_i^{(t)}$ is the Euclidean distance between the center of target $(x_{\text{obj}}^{(t)}, y_{\text{obj}}^{(t)})$ and i-th joint coordinates (x_i, y_i) at time t.

$$d_i^{(t)} = \sqrt{\left(x_{\text{obj}}^{(t)} - x_i^{(t)}\right)^2 + \left(y_{\text{obj}}^{(t)} - y_i^{(t)}\right)^2} \tag{10}$$

and $v_i^{(t)}$ are defined as

$$v_i^{(t)} = \left(d_i^{(t)} - m_i^{(t)}\right)^2. \tag{11}$$

This updated process proceeds only for the joints with confidence larger than th_c. Likewise, we tested the joints with the confidence of at least th_c and determined whether each joint relation is normal through the following equation:

$$l_i^{(t)} = \begin{cases} 1 & \text{if } \left(d_i^{(t)} - m_i^{(t-1)}\right) > \sqrt{s_i^{2(t-1)}} * th_j, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

where th_j is the threshold parameter. Unlike a typical test, we only considered $d_i^{(t)}$ when it was greater than $m_i^{(t-1)}$. In other words, we only considered the situations where the distance between the object and human joints increases according to the dumping action situation. As the last step, we used the voting method that measured the percentage of joints that passed the test of (12) from the total number of joints N as

$$I_{\text{final}}^{(t)} = \begin{cases} 1 & \sum_i I_i^{(t)}/N > th_{\text{v}} \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where $I_{\text{final}}^{(t)} = 1$ indicates that t-th frame contains a dumping action and th_{v} is the threshold parameter.

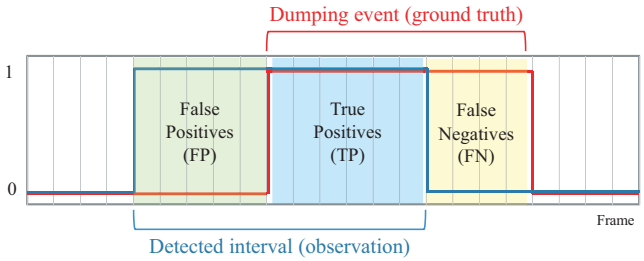


FIGURE 7 Illustration of false positives, true positives, and false negatives when true dumping event and detected event intervals are given

4 | EXPERIMENTS

The proposed method was implemented using C++ and OpenCV library. The joint information was estimated using the *OpenPose* library based on Caffe using C++. We experimented with fixed parameters to determine the parameters with the best experimental performance without tuning for each scene. For the threshold parameters, th_{h} equaled 0.5, th_{c} and th_{v} were 0.3, and th_{j} was 2 in the conventional background model [38]. Because the proposed method captures the dumping action by modeling the relationship between a person and object, we abbreviated the relation-based detection as RBD to represent our proposed method.

4.1 | Dataset

As an experimental dataset, we received videos from the closed-circuit television (CCTV) camera of local governments that were already installed. The video resolution was 1,280 × 720 pixel (HD). We collected videos from eight different spots, which were originally recorded for general surveillance purposes as well as illegal garbage dumping. Original videos were collected for several hours at each location; however, we created clips of approximately 10 minutes except for the section with no movements in scene. Although the target behavior was the dumping action of a person, in actual surveillance scenes, most events were normal, such as people or cars passing by. Thus, each compressed clip was re-edited to have a similar duration for both normal and dumping action events. As evident from the various experimental pictures, some of the images were taken at night but the lighting situation was adequate owing to the streetlamp at the installation location. We targeted those events that could be judged to determine dumping behavior in humans. Then, the ground truth of each clip was tagged for the starting and ending time of the dumping action. This ground truth was used to measure the false positives, true positives, and false negatives by overlapping the detected event intervals, as shown in Figure 7.

4.2 | Qualitative Comparisons

Figure 8 shows the result of detecting the dumping activity using the proposed algorithm. The red messages at the top left indicate that the proposed method detected a dumping

FIGURE 8 Qualitative results of dumping action. The red messages at the top left indicate that the proposed method detected a dumping action

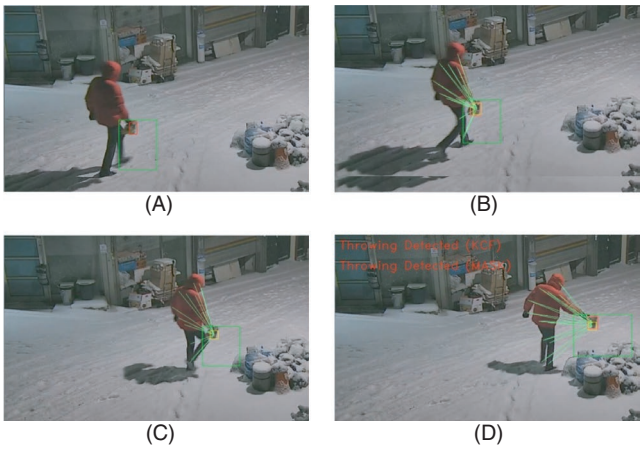


FIGURE 9 Qualitative results of detecting the dumping action in frame sequences

action. The message has two lines because our scheme can simultaneously operate with and without single target tracking in the Section 3.5. The detailed explanation is given in the quantitative comparisons below.

The proposed method can detect objects in various forms of garbage. In addition, the proposed method can detect various dumping actions irrespective of the dumping styles, such as bending the body and throwing it away.

Figure 9 shows the qualitative results of the proposed method in frame sequences. As shown in Figure 9A, the object carried by the person was detected. Then, through the online update scheme, the relation between the person and object was constructed, as shown in Figure 9B and 9C. When the dumping action occurred, our voting-based decision method detected the relation change and initiated the system alarm, as shown in Figure 9D.

4.3 | Quantitative Comparisons

To quantitatively evaluate the performance, we used the frame-level *precision*, *recall*, and *F*-measure. Because we focused on the online algorithm without future frame information, we adopted these frame-level measures instead of the conventional clip-based measures. Expressed in mathematical form, based on the ground truth label and detected label $l^{(t)}_{\text{final}}$ in (13), we measured the frame-level *precision* and *recall* as

$$precision = \frac{TP}{TP+FP}, \quad recall = \frac{TP}{TP+FN} \tag{14}$$

where *TP* is the number of true positives, *FP* is the number of false positives (false alarms), and *FN* is the number of false negatives (missing) over all frames. Considering the overall performance measure, we used the *F*-measure that represents a harmonic mean of *precision* and *recall*.

The problem addressed in this study is a new issue that has not been covered in previous studies, and it is impossible to apply conventional clip-based methods that require the start and end timing of a behavior. Therefore, we implemented three methods that generate detection results for each frame: Pose SVM [39], ST-GCN [40], and Multi-CNN [41]. Both Pose SVM and ST-GCN utilize joint coordinates to distinguish the dumping action. To train Pose SVM and ST-GCN, the ground truth of dumping action was assigned as a target value and each person's pose coordinates were given as input. Multi-CNN [41] uses three convolutional neural networks in a surveillance environment similar to our problem. To learn garbage dumping behavior in this model, labelling of dumping behavior was added in CNN and the modified model was retrained using the dumping action data.

Table 1 lists the quantitative results of each method. Pose SVM or ST-GCN using only joint information had the lowest overall performance owing to inaccurate joint information. Multi-CNN, which contains information about RGB image and motion change, is relatively better than Pose-based methods but still unsatisfactory. The proposed method is also based on inaccurate pose information but it exhibits better performance with tracking and voting-based decisions, which compensates for the incorrect input. To facilitate an overall understanding of the results, a *precision-recall* plot is shown in Figure 10. In the case of Pose SVM and ST-GCN, both *precision* and *recall* were low. As shown in Figure 8, the postures of the person dumping the garbage varies such as bending pose, throwing pose, and putting garbage on top of something. Therefore, pose coordinate features have limitations regarding the detection of dumping behavior.

Multi-CNN exhibited high values in terms of *recall* but it exhibited low *precision* values, as shown in Figure 10. According to the (14), the performance of Multi-CNN indicates the presence of several false positives. The critical

TABLE 1 Frame-level average *F*-measure results

Method	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6	Scene 7	Scene 8	Average
Pose SVM [39]	0.1896	0.0303	0.0520	0.2333	0.3627	0.1333	0.4720	0.1000	0.2562
ST-GCN [40]	0.0202	0.0229	0.0076	0.0769	0.4266	0.5660	0.0109	0.3442	0.2804
Multi-CNN [41]	0.4265	0.0882	0.4444	0.2462	0.5217	0.2717	0.5983	0.2134	0.3653
RBD (Proposed)	0.5149	0.8333	0.5000	0.7419	0.7194	0.8235	0.6667	0.5405	0.7136

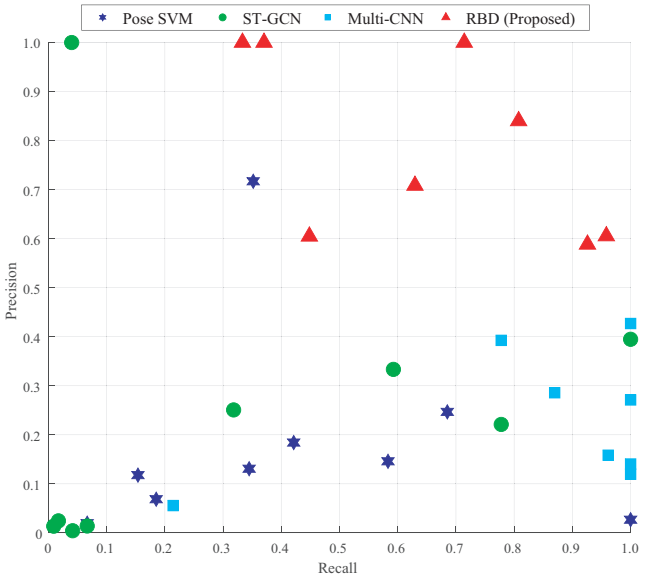


FIGURE 10 *Precision-recall* plot of each test. Marks that are not filled represent the average *precision-recall* of each algorithm. The algorithm performs better if each dot resides closer to the top-right corner

problem in the actual use of intelligent visual surveillance system is the occurrence of false positives because it interferes with the concentration of the observer instead of helping them. By contrast, owing to multiple steps such as carried-object detection and tracking based on reliable hand joint, the proposed method exhibited a high precision through few false positives.

4.4 | Ablation study of the proposed module

Tables 2 and 3 list the contribution of garbage object tracking in Section 3.5 and voting based decision in Section 3.6. As shown in Table 2, garbage object tracking improves the overall performance because it stabilizes the position and size of the object. In Table 3, RBD (w/o voting) indicates the distance between the hand coordinates and trash object in the dumping action decision without (13). RBD (w/voting) indicates that dumping action is detected by the voting scheme using the distances of all joints. As presented in Table 3, the voting method shows a significant improvement because the voting scheme considered the configuration of all joints. In other words, it gives better results when considering the overall relation between the human and object, rather than considering the distance between the human hand and the object.

4.5 | Parameter evaluation

In the proposed framework, several threshold parameters for decision modules were given. Among them, th_h in (7) denotes the carried object and th_v denotes the dumping action in (13). We used parameter sweep experiments to determine the parameters with the highest *F*-measure value. Figure 11A and 11B are the *F*-measure graphs of changing the heatmap threshold parameter th_h and the voting threshold parameter th_v, respectively. As shown in Figure 11A, it shows highest *F*-measure when th_h is 0.5. Thus, we set th_h to 0.5.

In the case of th_v, as shown in Figure 11B, it shows high performance when th_v is 0.3. Unlike th_h, th_v shows a large performance variation with parameter change. th_v determines the joints with changed distance that should be judged as a dumping event. The smaller value of th_v increases the sensitivity but sometimes false positives occur if the estimated joint location is uncertain. Conversely, if th_v is too large, most events are recognized as normal actions. This indicates that the relation model is useless because the dumping action is also considered normal according to (8) and (9). In this case, the model loses the distinction of our target behavior and consequently, most situations are judged to be normal. The parameter th_v can either be increased when the joints are rarely detected to increase the sensitivity or decreased when the relationship between joints and distance slowly changes.

TABLE 2 Frame-level average *precision*, *recall*, and *F*-measure results for the contribution to garbage object tracking

Measure	*Precision*	*Recall*	*F*-measure
RBD (w/o tracking)	0.5390	0.4521	0.4917
RBD (w/ tracking)	0.7933	0.6485	0.7136

TABLE 3 Frame-level average *precision*, *recall*, and *F*-measure results for the contribution to voting-based decision

Measure	*Precision*	*Recall*	*F*-measure
RBD (w/o voting)	0.4413	0.3657	0.4000
RBD (w/ voting)	0.7933	0.6485	0.7136

4.6 | Runtime evaluation

Table 4 lists the computation time of each module measured on Intel Core i7-6700K 4.0 GHz PC and GTX TITAN X GPU. GPU is only used in the Openpose library and all other modules operate at CPU level. Using the multi-GPU function supported by the *Openpose* library, pose estimation up to 20.92 ms (48 fps) was achieved.

For other modules, the foreground extraction took the longest time equaling 46 ms but the rest did not take much time. To summarize, our framework runs at approximately 6.74 fps when using one GPU and at 10.57 fps when using four GPUs. In other words, it is possible to process the image with a grabbing rate of 10 fps in real-time detection in real-world application. Even if the grabbing rate changes depending on computations or GPU resources, our method maintains the performances because the proposed relation

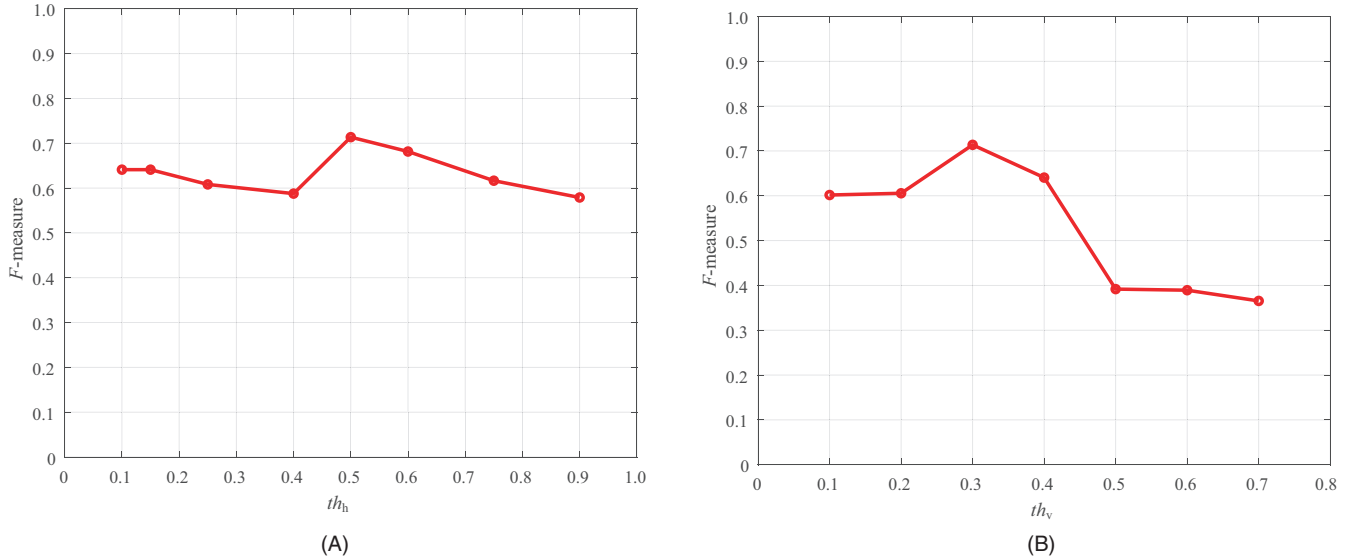


FIGURE 11 *F*-measure graphs for parameter evaluation: (A) the heatmap threshold th_h in (7) and (B) the voting threshold th_v in (13)

model handles the grabbing rate change through the online update method.

4.7 | Limitations and future works

The proposed method detects dumping activities by determining the change in relationship between a person and a hand-held object. This approach can handle various dumping actions and discarded objects but it depends on the performance of joint estimation, assuming that the garbage object is visible. If a pedestrian is accurately tracked even in an occluded situation, dumping action can be deduced by determining the person who carried an object and did not have the object after a specific period. In addition, although learning-based methods such as ST-GCN and Multi-CNN do not show satisfactory performance owing to the challenges in this problem, their performances can be improved by combining them with this method through more specific action modeling. Future work will focus on improving the performance through prior knowledge and learning-based reasoning. For example, if people stay in the garbage dump area for a long time even if the object is not visible, it can be a dumping action with high probability and the system informs and asks the monitoring agent for the final judgment.

TABLE 4 Computation time of each part in the proposed method per frame

Module	Time (millisecond)
Joint estimation (1 GPU)	74.62
Joint estimation (2 GPU)	38.02
Joint estimation (4 GPU)	20.92
Foreground region extraction	45.78
Pedestrian tracking	12.87
Object candidate extraction	14.93
Garbage object tracking	0.1367
Voting-based detection	0.0058

5 | CONCLUSION

In this paper, we proposed a new framework to detect dumping behavior by extracting candidates of garbage objects and modeling the distance configuration using human joint information and foreground extraction algorithm. The proposed framework was targeted to detect the dumping action in actual applications. Therefore, because it was designed to cover the various patterns of actual dumping behavior, it worked robustly against various camera views, garbage objects, and acting styles. Experimental results showed the effectiveness of the proposed method in comparison with state-of-the art methods. In addition, because our framework significantly reduced false positives and operated online in real-time, the proposed framework is suitable for real-world surveillance application. Future works will detect garbage dumping when the garbage object is invisible through the inference using a tracking and pretrained inference model.

ORCID

Kimin Yun https://orcid.org/0000-0002-4493-9437

REFERENCES

1. W. Liu et al., *SSD: Single shot multibox detector*, in Proc. Eur. Conf. Comput. Vision (ECCV), Amsterdam, The Netherlands, Oct. 11–14, 2016, pp. 21–37.

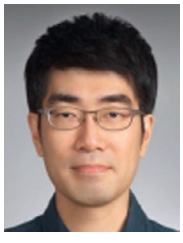
2. J. Redmon et al., *You only look once: Unified, real-time object detection*, in Proc. Comput. Vision Pattern Recogn. (CVPR), Las Vegas, NV, USA, June 27–30, 2016, pp. 779–788.
3. S. Ren et al., *Faster R-CNN: Towards real-time object detection with region proposal networks*, IEEE Trans. Pattern Anal. Mach. Intell. **39** (2017), 1137–1149.
4. K. Simonyan and A. Zisserman, *Two-stream convolutional networks for action recognition in videos*, in Proc. Adv. Nneural Inf. Process. Syst. (NIPS), Montreal, Canada, Dec. 8–13, 2014, pp. 567–576.
5. D. Tran et al., *Learning spatiotemporal features with 3D convolutional networks*, in Proc. Int. Conf. Comput. Vision (ICCV), Santiago, Chile, Dec. 7–13, 2015, pp. 4489–4497.
6. L. Wang et al., *Temporal segment networks: Towards good practices for deep action recognition*, in Proc. Eur. Conf. Comput. Vision (ECCV), Amsterdam, The Netherlands, Oct. 11–14, 2016, pp. 20–36.
7. F. Porikli, Y. Ivanov, and T. Haga, *Robust abandoned object detection using dual foregrounds*, EURASIP J. Adv. Signal Process. **30** (2008), 1–11.
8. R.H. Evangelio and T. Sikora. *Complementary background models for the detection of static and moving objects in crowded environments*, in Proc. Adv. Video Signal-Based Surveillance (AVSS), Klagenfurt, Austria, 2011, pp. 71–76.
9. S. Kim et al., *Intelligent visual surveillance-A survey*, Int. J. Contr. Autom. Syst. **8** (2010), 926–939.
10. G. Chunhui et al., AVA: A video dataset of spatio-temporally localized atomic visual actions, arXiv:1705.08421, 2017.
11. W. Kay et al., The kinetics human action video dataset, arXiv:1705.06950, 2017.
12. J. Moon et al., *Extensible hierarchical method of detecting interactive actions for video understanding*, ETRI J. **39** (2017), 502–513.
13. L. Cewu, J. Shi, and J. Jia. *Abnormal event detection at 150 FPS in MATLAB*, in Proc. Int. Conf. Comput. Vision (ICCV), Sydney, Australia, Dec. 1–8, 2013, pp. 2720–2727.
14. K. Yun, Y. Yoo, and J.Y. Choi, *Motion interaction field for detection of abnormal interactions*, Mach. Vis. Appl. **28** (2017), 157–171.
15. R. Csordas, L. Havasi, and T. Sziranyi, *Detecting objects thrown over Fence in outdoor scenes*, in Proc. Int. Conf. Comput. Vision Theory Applicat. (VISAPP), Berlin, Germany, 2015, pp. 593–599.
16. S. Mahankali et al., *Identification of illegal garbage dumping with video analytics*, in Proc. Int. Conf. Adv. Comput., Commun. Inf. (ICACCI), Bangalore, India, Sept. 19–22, 2018, pp. 2403–2407.
17. H. Begur et al., *An edge-based smart mobile service system for illegal dumping detection and monitoring in San Jose*, in IEEE SmartWorld Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI, San Francisco, CA, USA, Aug. 4–8, 2017, pp. 1–6.
18. A. Dabholkar et al., *Smart illegal dumping detection*, in Proc. IEEE Int. Conf. Big Data Comput. Service Applicat., San Francisco, CA, USA, Apr. 6–9, 2017, pp. 255–260.
19. E. Esen, M.A. Arabaci, and M. Soysal, *Fight detection in surveillance videos*, in Proc. Content-Based Multimedia Indexing (CBMI), Veszprem, Hungary, June 17–19, 2013, pp. 131–135.
20. Z. Zhang, C. Conly, and V. Athitsos, *A survey on vision-based fall detection*, in Proc. ACM Int. Conf. Pervasive Technol. Related Assistive Environ., Corfu, Greece, July 1–3, 2015, pp. 1–7.
21. X. Zhou, C. Yang, and Y. Weichuan, *Moving object detection by detecting contiguous outliers in the low-rank representation*, IEEE Trans. Pattern. Anal. Mach. Intell. **35** (2013), 597–610.
22. S. Javed et al., *Background-foreground modeling based on spatio-temporal sparse subspace clustering*, IEEE Trans. Image Process. **26** (2017), 5840–5854.
23. B. Heo, K. Yun, and J.Y. Choi, *Appearance and motion based deep learning architecture for moving object detection in moving camera*, in Proc. Int. Conf. Image Process. (ICIP), Beijing, China, Sept. 17–20, 2017, pp. 1827–1831.
24. T.P. Nguyen et al., *Change detection by training a triplet network for motion feature extraction*, IEEE Trans. Circuits Syst. Video Technol. **29** (2019), 433–446.
25. K. Yun, J. Lim, and J.Y. Choi, *Scene conditional background update for moving object detection in a moving camera*, Pattern Recogn. Lett. **88** (2017), 57–63.
26. T. Yi Lin et al., *Microsoft COCO: Common objects in context*, in Proc. Eur. Conf. Comput. Vision (ECCV), Zurich, Switxerland, Sept. 6–12, 2014, pp. 740–755.
27. M. Andriluka et al., *2D human pose estimation: new benchmark and state of the art analysis*, in Proc. Comput. Vision Pattern Recogn. (CVPR), Columbus, OH, USA, June 23–28, 2014, pp. 3686–3693.
28. Z. Cao et al., *Realtime multi-person 2D pose estimation using part affinity fields*, in IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Honolulu, HI, USA, July 21–26, 2017, pp. 1302–1310.
29. G. Rogez, P. Weinzaepfel, and C. Schmid, *LCR-Net: Localization-classification-regression for human pose*, in IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Honolulu, HI, USA, July 21–26, 2017, pp. 1216–1224.
30. H. Yoo et al., *Online scheme for multiple camera multiple target tracking based on multiple hypothesis tracking*, IEEE Trans. Circuits Syst. Video Technol. **27** (2017), 454–469.
31. H.W. Kuhn, *The Hungarian method for the assignment problem*, Naval Research Logistics Quarterly **2** (1955), 83–97.
32. P.F. Felzenszwalb, R.B. Girshick, and D.A. McAllester, *Cascade object detection with deformable part models*, in Proc. Comput. Vision Pattern Recogn. (CVPR), San Francisco, CA, USA, June 13–18, 2010, pp. 2241–2248.
33. J. Sklansky, *Finding the convex hull of a simple polygon*, Pattern Recogn. Lett. **1** (1982), 79–83.
34. K. Kang et al., *Invariant-feature based object tracking using discrete dynamic swarm optimization*, ETRI J. **39** (2017), 151–162.
35. M. E. Yildirim et al., *Direction-based modified particle filter for vehicle tracking*, ETRI J. **38** (2016), 356–365.
36. S. Yun et al., *Action-decision networks for visual tracking with deep reinforcement learning*, in IEEE Conf. Comput. Vision Pattern Recogn. (CVPR), Honolulu, HI, USA, July 21–26, 2017, pp. 1349–1358.
37. J. F. Henriques et al., *High-speed tracking with kernelized correlation filters*, IEEE Trans. Pattern Anal. Mach. Intell. **37** (2015), 583–596.
38. K. M. Yi et al., *Detection of moving objects with non-stationary cameras in 5.8ms: bringing motion detection to your mobile device*, in IEEE Conf. Comput. Vision Pattern Recogn. Workshops (CVPRW), Portland, OR, USA, June 23–28, 2013, pp. 27–34.
39. C. Cortes and V. Vapnik, *Support-vector networks*, Mach Learn. **20** (1995), 273–297.

40. S. Yan, Y. Xiong, and D. Lin, *Spatial temporal graph convolutional networks for skeleton-based action recognition*, in Proc. AAAI Conf. Artif. Intell., New Orleans, LA, USA, Feb. 2–7, 2018.
41. C.-B. Jin et al., *Real-time action detection in video surveillance using sub-action descriptor with multi-cnn*, arXiv preprint, (2018), arXiv:1710.03383.

AUTHOR BIOGRAPHIES

Kimin Yun received his BS degree and unified MS and PhD degrees in Electrical Engineering and Computer Science from the Seoul National University, Seoul, Rep. of Korea, in 2010 and 2017, respectively. Since 2017, he has been working with the Visual Intelligence Research Group in the SW & Contents Research Laboratory at the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Rep. of Korea, where he has been involved in projects on large-scale visual BigData analysis. His current research interests include computer vision, visual event analysis, moving object detection, and video analysis based on machine learning.

Yongjin Kwon received his BS degree in Computer Science and Engineering from the Pohang University of Science and Technology, Rep. of Korea, in 2009 and his MS degree in Computer Science and Engineering from the Seoul National University, Rep. of Korea in 2012. Since 2012, he has been working with the Visual Intelligence Research Group in the SW & Contents Research Laboratory at the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Rep. of Korea. His research interests include database systems, information retrieval, and machine learning.



Sungchan Oh received his BS, MS, and PhD degrees in Electronic Engineering from the Sogang University, Seoul, Rep. of Korea, in 2006, 2008, and 2014, respectively. From 2014 to 2016, he worked as a senior researcher with LG Electronics, Seoul. He is currently working with the Visual Intelligence Research Group in the SW & Contents Research Laboratory at the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Rep. of Korea. His research interests are computer vision and image processing.

Jinyoung Moon received her BS degree in Computer Engineering from the Kyungpook National University, Daegu, Rep. of Korea, in 2000. She received her MS degree in Computer Science and PhD in Industrial & Systems Engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 2002 and 2018, respectively. Since 2002, she has been working with the Visual Intelligence Research Group in the SW & Contents Research Laboratory at the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Rep. of Korea. Her research interests include human behavior understanding, human attention analysis, and machine learning.

Jongyoul Park received his BS degree in Computer Engineering from the Chungnam National University, Daejeon, Rep. of Korea, in 1996, and his MS degree and PhD in Information and Communication Engineering from the Gwangju Institute of Science and Technology, Rep. of Korea, in 1999 and 2004, respectively. From 2001 to 2002, he was a visiting researcher at the School of Computing, University of Utah, UT, USA. Since 2004, he has been working with the Visual Intelligence Research Group in the SW & Contents Research Laboratory at the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Rep. of Korea. His research interests include visual intelligence, machine learning, and behavior and scene understanding.