Anti-Litter Surveillance based on Person Understanding via Multi-Task Learning

Kangmin Bae*
kmbae@etri.re.kr
Kimin Yun*
kimin.yun@etri.re.kr
Hyungil Kim
hikim@etri.re.kr
Youngwan Lee
yw.lee@etri.re.kr
Jongyoul Park
jongyoul@etri.re.kr

Artificial Intelligence Research Laboratory
Electronics and Telecommunications Research Institute (ETRI)
Daejeon, Republic of Korea

Abstract

In this paper, we propose a new framework for an anti-litter visual surveillance system to prevent garbage dumping as a real-world application. There have been many efforts to deploy an action recognition based visual surveillance system. However, many conventional methods were overfitted for only specific scenes due to hand-crafted rules and lack of real-world data. To overcome this problem, we propose a novel algorithm that handles the diverse scene properties of the real-world surveillance. In addition to collecting data from the real-world, we train the effective model to understand the person through multiple datasets such as human poses, human coarse action (*e.g.*, upright, bent), and fine action (*e.g.*, pushing a cart) via multi-task learning. As a result, our approach eliminates the need for scene-by-scene tuning and provides robustness to behavior understanding performance in a visual surveillance system. In addition, we propose a new object detection network that is optimized for detecting carryable objects and a person. The proposed detection network reduces the computational cost by specifying potential suspects only to the person who carries an object. Our method outperforms the state-of-the-art methods in detecting the garbage dumping action on real-world surveillance video dataset.

1 Introduction

The environmental pollution problem caused by the dumping of garbage continues to be serious. According to a recent report [10], 15 tonnes of plastic waste leaks into the sea every minute, which is equivalent to the capacity of one garbage truck. Many countries and states have been trying to reduce trash and illegal dumping, by establishing a litterer report program [1, 7, 9, 11]. As machine learning and computer vision technologies have exploded in

*They are equally contributed and sorted in alphabetical order.

Figure 1: The overview of our method.

recent years, there are many efforts to make use of these technologies to solve the litter and illegal dumping problem that causes the social and environmental problems. For example, the non-profit research group [29] has been collecting the trash dataset (TACO) and suggested several future applications: drones surveying trash, robots picking up litter, anti-litter surveillance, and so on. We also deeply sympathize with developing an artificial intelligence (AI) and computer vision system that solves littering problem. As a part of solving the littering problem, we propose an anti-litter video surveillance system in the real-world. In conjunction with the litterer report program, our method can help to prevent unauthorized dumping by automatically detecting waste litterers, and reporting them to the authorities.

In this paper, we propose a novel framework for a real-world visual surveillance system to prevent garbage dumping. As shown in Figure 1, the proposed method detects the person who dumps the garbage in a two-stage manner. From the input frame, the proposed method detects the person and carryable objects (denoted as the black bounding boxes), then decides a potential litterer (denoted as the green box). After selecting the target person, the proposed method estimates the human pose and infers the human actions including a dumping action.

The proposed method effectively utilizes multiple datasets in order to avoid overfitting caused by insufficient surveillance datasets through multi-task learning. For more detail, the human pose, human state (coarse action), human detailed behavior (fine action), and the real-world surveillance datasets are jointly used to train our network. Since these datasets provide complementary information to each other, the multi-task learning can improve the performance of garbage dumping action detection. Also, we train a new object detector by re-organizing COCO dataset [25] based on objects that person can carry ('carryable'), which enables to detect various human-carrying objects and limit the target person to be monitored. The experimental results show the improved performance compared to both the conventional image-based recognition methods [6, 12] and the state-of-the-art methods [18, 37, 40].

2 Related Works

2.1 Action recognition

Recently, action recognition research has been actively conducted using large-scale datasets [3, 38]. Most studies have extended image recognition studies using convolutional neural networks (CNN) to video understanding: action classification (*i.e.*, finding action labels at the clip level) [4, 32, 33] and action localization (*i.e.*, finding start and end of the event in the video) [27, 30, 42]. However, there are many difficulties in applying the success of these studies directly to surveillance applications. First, conventional methods focus on sports and movies that have behavioral characteristics through the whole video. On the other hand, the action/event in the surveillance scene occurs in a locally and temporally limited region, so the feature from the whole frames may contain meaningless information. In addition, the surveillance camera shoots the video in continuous streaming while traditional methods han-

dle fixed length video. In contrast to previous works, we design a framework for dumping action detection (as one of the real-world action recognition applications) in online streaming video.

2.2 Intelligent visual surveillance

Research on making an intelligent visual surveillance system has been studied for a long time [16]. In the past, the system aims to detect target events as the rule-based method depending on the scene and camera constraints [26, 39]. The most similar application to our problem is the abandoned luggage detection, which has been developed for the dangerous object detection in major facilities such as airports [17, 28]. However, the scenarios in the abandoned luggage detection are impractical because the abandoned objects in experiments are in a well visible position in the frame. Moreover, although it is important who actually left the object, they only focused on abandoned objects after a person leaved the object.

Due to advances in deep learning and the release of large-scale datasets (*e.g.*, ImageNet [8]), many computer vision algorithms have reduced the domain and scene dependency and have been applied to surveillance scenes [6, 18, 23, 24, 37, 41]. In detecting the garbage dumping action, a few works showed the possibility of the surveillance behavior understanding based on the person and surrounding object [19, 40]. Although they showed improved results over the rule-based methods, they still had the scene dependency problem due to a lack of real-world data.

2.3 Multi-task learning

The proposed method utilizes multi-task learning techniques to overcome the modeling limitations due to a lack of data for the target problem. Multi-task learning can increase generality through multiple datasets that can generate synergies. As a representative work, Mask R-CNN [13] performs classification, detection, and segmentation while sharing the same backbone network. In human understanding, a few studies [2, 15, 34] have showed the improvement through the use of relative information such as human pose and surrounding objects. For understanding actions in surveillance scene, we extend and devise the multi-task network to obtain the generality through simultaneous learning of pose, coarse action, fine action, and target surveillance action.

3 The Proposed Method

As depicted in Figure 2, our framework first finds the person and carryable objects, and then performs multiple tasks including the dumping action classification. For the person and carryable object detector, we re-organize the detection dataset based on objects that person can carry ('carryable') and train the detection network based on the re-organized dataset. It enables to detect human-carrying objects with various shapes and improves the performance and efficiency by limiting the target person to be monitored.

To establish the multi-task classification network, we propose a new structure that utilizes multiple datasets for the person understanding to avoids overfitting. The multi-task network consists of three modules: 1) feature extraction network, 2) feature fusion network, and 3) pose estimation and multiple action classification networks. To be more specific, the proposed classifier is jointly trained by four datasets: the COCO keypoint dataset [25] to infer the locations of human body parts (pose), the MPHB dataset [3] to infer the overall state

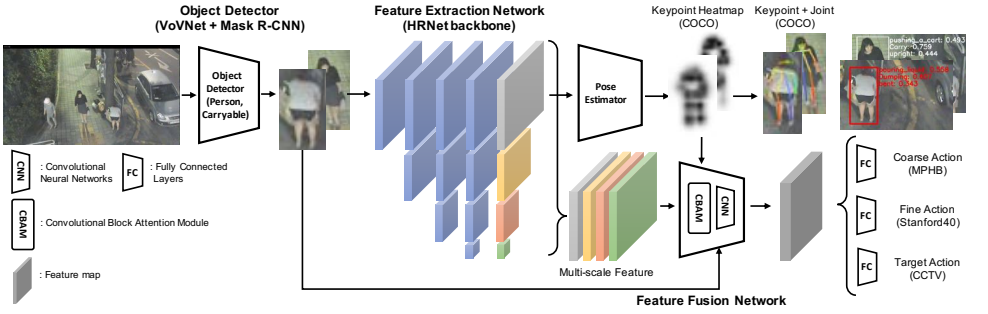

Figure 2: Whole architecture of the proposed network. The proposed network performs four tasks: 1) COCO style pose estimation, 2) coarse action (MPHB), 3) fine action (Stanford40), and 4) target action (CCTV) classification. The multi-scale features and keypoint heatmaps are fused to generate optimized features for dumping action detection.

of a person (coarse action), the Standford40 dataset [38] to find person's detailed behavior (fine action), and the surveillance dataset to model the dumping action patterns from the real-world surveillance camera (target action).

3.1 Person and carryable object detector

To increase the efficiency, we build a new carryable object detector that limits the target person to the person who carries an object. Thus, we design and train a carryable object detector by creating a new class with a functional classification of a category. More specifically, we re-organized the COCO object detection dataset [25] that contains 80 object classes into three classes: 'person', 'carryable', and 'others'. The 'person' class denotes an ordinary *person* which is the same as the original COCO dataset [25]. However, we reassigned objects such as a *backpack*, *handbag*, and *suitcase* that can be carried by a person to 'carryable' class. And, 'others' class consists of the object classes apart from human interests like a *bus*, *train*, and *bird* class. After that, we train a detection network on the newly classified COCO dataset using the model pretrained on ImageNet [8]. We train the Mask R-CNN [13] with VoVNet [21, 22] backbone, where VoVNet is an energy and GPU memory efficient backbone that meets the resource constraints of the surveillance system.

3.2 Feature extraction network

In order to train the network for multiple tasks, the structure of the feature extraction network is important. Since our target problem should distinguish behaviors through differences in the human image, we need a high-resolution feature representation as a fine-grained classification problem. Therefore, we adopt the HRNet [31] as the feature extraction network instead of other conventional networks such as ResNet [12] by two reasons: 1) HRNet [31] is designed to extract a high-resolution feature for both image classification and keypoint detection, 2) HRNet [31] can extract multi-scale features simultaneously.

We design the feature extraction network as a shared the backbone network structure for multiple tasks. Since our network is trained to estimate the pose and classify multiple actions simultaneously, the extracted feature improves the overall human understanding by avoiding overfiiting. For example, human pose information can enable the network to attend the discriminative joint features such as bent pose or outstretched pose to dump an object. As

depicted in Figure 2, our backbone network based on HRNet is connected to two paths: keypoint estimator and multiple action classifiers. For the keypoint estimator, the high-resoultion feature is used to generate the pose heatmaps like the original HRNet. For the action classifiers, the multi-scale features are used to handle the various sizes of the human and the object that the person carries.

3.3 Feature fusion network

In this module, we try to incorporate multiple features to make a improvements in detecting the garbage dumping action. As shown in Figure 2, the feature fusion network takes the pose feature, multi-scale features, and the input RGB patch as an input. Human pose feature is used to guide to focus the human details. The multi-scale features are the human-centric features because the backbone is trained by human pose and actions simultaneously. The RGB image gives the information on the carrying object and the background. To incorporate these features, we propose the feature fusion network that fuses three different features: the RGB patch ($\mathbf{x}$), the concatenated multi-scale features ($\phi(\mathbf{x})$), and the predicted pose heatmap ($\hat{\mathbf{h}}$). These features are resized to the size of $\phi(\mathbf{x})$ in order to perform concatenation. The fused feature $\rho(\mathbf{x})$ is given as:

$$\rho(\mathbf{x}) = \psi(\mathbf{x} \oplus \phi(\mathbf{x}) \oplus \hat{\mathbf{h}}), \tag{1}$$

where ψ is the feature fusion network and $\oplus$ is the concatenation of features. For the feature fusion network, we stack the convolutional block attention module (CBAM) [35] and an additional convolutional layer, sequentially. Here, CBAM [35] is utilized as an adaptive feature fusion block to weight features with different characteristics. Then, additional convolutional layer is added to incorporate the weighted feature to action feature space.

3.4 Pose estimation and multi-task classification network

Our network consists of a pose estimation network and three action classification networks to achieve multi-task learning. The pose estimation network consists of a single convolution neural network by connecting the highest resolution feature of the backbone network. Each action classification network is devised using a single-layer perceptron to receive features obtained from the feature fusion network.

To train the proposed network, the cross entropy loss (CE) and the mean squared error (MSE) loss are used for the action classification and the keypoint estimation, respectively. Let $\mathcal{X}^{(i)}$ be the i^{th} action dataset among the multiple action datasets. $f(\cdot;\theta)$ denotes the feature from the backbone network f with parameter θ. And, the $g(\cdot;\eta^{(i)},\mathcal{X}^{(i)})$ is defined as the i^{th} classification network with parameter $\eta^{(i)}$ corresponding dataset $\mathcal{X}^{(i)}$. For the j^{th} element of the i^{th} action dataset $\mathcal{X}^{(i)}$, a paired set $\{\mathbf{x}_j, y_j\}$ is given as input image $\mathbf{x}_j$ and action label y_j. The classification loss $\mathcal{L}_{cls}$ is given as:

$$\mathcal{L}_{cls}(\mathcal{X}^{(i)};\theta,\eta^{(i)}) = \frac{1}{n^{(i)}} \sum_{\{\mathbf{x}_j,y_j\}\in\mathcal{X}^{(i)}} CE(g(f(\mathbf{x}_j;\theta);\eta^{(i)},\mathcal{X}^{(i)}),y_j), \tag{2}$$

where $n^{(i)} = |\mathcal{X}^{(i)}|$ and $|\cdot|$ is the cardinality of the set. That is, the proposed network uses the shared parameter θ for the feature extraction f and the different parameters $\eta^{(i)}$ depending on the i^{th} action.

For keypoint estimation, the MSE loss is calculated for each channel of keypoint heatmap. Let $\mathcal{X}^p$ be the keypoint dataset for human pose and $h(\cdot;\eta^p)$ denotes the estimated heatmap of each joint with parameter η^p. The keypoint estimation loss $\mathcal{L}_p$ is given as:

$$\mathcal{L}_p(\mathcal{X}^p;\theta,\eta^p) = \frac{1}{n^p} \sum_{\{\mathbf{x}_j,\mathbf{h}_j\}\in\mathcal{X}^p} ||h(f(\mathbf{x}_j;\theta);\eta^p,\mathcal{X}^p) - \mathbf{h}_j||_2^2, \tag{3}$$

where $n^p = |\mathcal{X}^p|$ and $\{\mathbf{x}_j,\mathbf{h}_j\}$ is a paired set with the input image $\mathbf{x}_j$ and the corresponding ground truth heatmap $\mathbf{h}_j$. Finally, the total loss $\mathcal{L}_{total}$ is given as:

$$\mathcal{L}_{total} = \sum_{\mathcal{X}^{(i)}} \mathcal{L}_{cls}(\mathcal{X}^{(i)};\theta,\eta^{(i)}) + \mathcal{L}_p(\mathcal{X}^p;\theta,\eta^p), \tag{4}$$

where $\mathcal{X}^{(i)} \in \{\mathcal{X}^{coarse},\mathcal{X}^{fine},\mathcal{X}^{target}\}$. Therefore, the whole network parameter Θ of our proposed network is given as:

$$\Theta = \{\theta,\eta^p,\eta^c,\eta^f,\eta^t\}, \tag{5}$$

where η^c, η^f, and η^t denote the parameters of classification network for $\mathcal{X}^{coarse}$, $\mathcal{X}^{fine}$, and $\mathcal{X}^{target}$, respectively. The final goal is to find the optimal network parameter $\hat{\Theta}$ as:

$$\hat{\Theta} = \arg\min_{\Theta} \mathcal{L}_{total}, \tag{6}$$

which enables network to perform pose estimation, and multi-task classification at the same time.

4 Experiments

4.1 Dataset

While training the network, the number of datasets substantially affects the final performance. However, due to privacy and security issues, we could not find any publicly available dumping action detection dataset. Therefore, we have built our own dataset from over 50 closed-circuit television (CCTV) cameras installed by local governments. Raw videos are dealt with only for research purpose under security manuals. We split the raw videos to 899 videos with at least one dumping action in each clip. These videos were taken with resolution of 1920×1080 where the resolution of average human size was 39×64. Among 899 videos, 101 were taken at night and the length of video varies from 3 to 13 seconds. We have labeled not only litterers but also non-litterers to train and validate the proposed method.

Our CCTV dumping action dataset contains more than $220k$ human bounding boxes. In labeling procedure, we divided the dataset into four classes ('Carry', 'Just-before', 'Dumping', 'Normal') rather than a simple binary label (with and without garbage dumping action). This labeling scheme allows the classifier to distinguish details in human images and prevents it from overfitting to the background. The sample images of each class are presented in Figure 3. We manually gave the label of the human bounding box to one of the classes: 'Carry', 'Just-before', 'Dumping', and 'Normal'. A class named 'Carry' denotes all people who carry an object. The object could be a garbage as well as non-garbage such as bag or umbrella. The 'Just-before' class denotes a person who stands just next to the garbage pile with garbage on one's hand. The 'Dumping' class denotes a person who is dumping the

garbage. Finally, the 'Normal' class denotes all the other people who are just going through the CCTV angle. We split the CCTV dataset into two sets (*i.e.*, training and validation) without any scene overlap, where the number of videos for training and validation sets is 813 and 86, respectively.

To overcome the small number of training data, we utilize three additional datasets; 1) Stanford40 [38], 2) MPHB [3], 3) COCO [25]. The Stanford40 [38] dataset labels $10k$ human action images using 40 different classes such as *pushing a cart* and *fixing a car*. Each class is very specific enough to regard it as a fine action dataset. The MPHB [3] dataset has $26k$ images labeled using six classes such as *lying* and *bent*, which is used as a coarse action dataset. Finally, we use COCO [25] dataset as a human pose dataset which contains $59k$ humans pose labeled using keypoints.

(a) (b) (c) (d)

Figure 3: The examples of our surveillance action dataset for each class: (a) Carry, (b) Just-before, (c) Dumping, and (d) Normal.

4.2 Implementation details

Our network is trained on a GPU machine with Intel®Xeon®CPU E5-2660 v4 @ 2.00GHz CPU, 128GB RAM, and four Titan RTX 24GB GPUs. For training the person and carryable object detector with re-organized dataset, we follow the same setting of released code of VoVNet-39 [21, 22] based on Detectron2 [36]. For the multi-task learning for pose estimation and action classification, the batch size is set to 128, and ADAM [20] optimizer is used with 0.00001 initial learning rate to train 40M parameters. Each mini-batch consists of randomly sampled data from the whole multi-task datasets. Since the sampled data have information for the dataset source (*i.e.*, pose or coarse action dataset), only relevant network parameters are updated. We iterate to 210 epochs and decay our learning rate as a factor of 0.1 at epoch 170 and 185. Additionally, we initialize the network using ImageNet pre-trained weights. It takes 5 days to train the proposed multi-task network. The proposed network operates at $110ms$ (carryable object detector: $81ms$ + multi-task action classification network: $29ms$) per frame.

4.3 Performance measure

For performance measure, we have utilized a frame-level accuracy and F-score. When the input frame $\mathbf{x}$ and label y are given, the frame-level accuracy of prediction $\hat{y}$ is given as:

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[y_i = \hat{y}_i], \tag{7}$$

where i is the frame index and n is the total number of frames. The $\mathbb{1}$ is the indicator function that gives 1 if the condition is satisfied and 0 otherwise. To measure F-score, we calculate true positive, false positive and true negative in order to obtain *precision*, *recall*, and F-score. We have compared our methods quantitatively using frame-wise accuracy, *precision*, *recall* and F-score.

5 Results

5.1 Quantitative results

We have compared the baseline and the state-of-the-art methods: SVM [6] is the trained support vector machine using the patch label, RBD [40] is the relation-based detection from a person and carrying object, ST-GCN [37] is a graph convolution network (GCN) using a keypoint information, and OHA-GCN [19] is a GCN using an object-related human pose. In Table 1, our method shows the best performances than the state-of-the-art algorithms in both accuracy and F-score measure. From Table 1, we can infer that others have large false negatives because they show low recall compared to accuracy. This is mainly originated due to a large number of negative samples compared to the positives. Furthermore, the accuracy metric is easily get distorted by data imbalance. Better precision and recall values compared to the other methods prove that our method is less biased to false negatives.

Method	**Accuracy**	*Precision*	*Recall*	***F*-score**
SVM [6]	0.639	0.376	0.269	0.314
RBD [40]	0.658	0.372	0.166	0.229
ST-GCN [37]	0.525	0.325	0.508	0.396
OHA-GCN [19]	0.614	0.395	0.483	0.434
Ours	**0.743**	0.581	0.582	**0.581**

Table 1: Quantitative results of our methods compared to the state-of-the-art methods.

5.2 Ablation studies

We perform an ablation study on our method. Table 2 shows the result of the ablation study. The "Carry" denotes the usage of the re-organized object class to train the carryable object detection network. If the "Multi-task" is not checked, only target data (*i.e.*, CCTV dataset) is used for training the network, where the backbone network is directly connected to the target action classifier. When the "Feature Fusion" is not checked, the operation of the feature fusion network is replaced with the simple concatenation. Since the feature fusion network requires features extracted from multi-task datasets, the "Feature Fusion" can be valid only if the "Multi-task" is checked.

Method	Backbone	Carry	Multi-task	Feature Fusion	**Accuracy**	*Precision*	*Recall*	***F*-score**
Ours	ResNet-50	-	-	-	0.641	0.434	0.561	0.489
	ResNet-50	✓	-	-	0.667	0.463	0.526	0.492
	HRNet-W32	-	-	-	0.716	0.539	0.525	0.532
	HRNet-W32	✓	-	-	0.740	0.587	0.511	0.546
	HRNet-W32	-	✓	-	0.694	0.502	0.627	0.557
	HRNet-W32	✓	✓	-	0.720	0.539	0.608	0.571
	HRNet-W32	-	✓	✓	0.723	0.544	0.603	0.572
	HRNet-W32	✓	✓	✓	**0.743**	0.581	0.582	**0.581**

Table 2: Ablation study for our proposed method. The "Carry", "Multi-task", and "Feature Fusion" denote the use of the carryable object detector, multi-task learning, and the feature fusion network, respectively.

The backbone selection, object detector for carrying object and person, multi-task dataset, and feature fusion are all contributed to the performance improvement. The HRNet structure, which can extract multi-scale features rather than ResNet for classification purpose, helps improve performance. Multi-task learning and fusion of features also helped to improve performance through a better understanding of a person. In Table 2, HRNet without multi-task showed a lower recall value as 0.525 compared to the recall of HRNet with multi-task as 0.627. That is, the proposed multi-task learning shows the effect of reducing bias towards false negatives. In addition, the carryable object detector improves the precision score, which reflects the actual demand to reduce the fatigue of the controller caused by a number of false

positives. For the overall performance, the accuracy and F-score were improved by more than 15% and 18%, respectively compared to the baseline network based on ResNet-50 [12] backbone.

5.3 Qualitative results

Figure 4(a) and (b) show the example results of the conventional object detector and the proposed carryable object detector, respectively. With the conventional object detector, objects carried by humans are rarely detected as a similar class (*i.e. handbag*), however they are usually missed due to low confidence. On the other hand, the proposed detector with the 'carryable' class improves performance even with small objects and unclear shapes.

(a) (b)

Figure 4: The qualitative results of person and carryable object detection on surveillance scene: (a) results trained with the original COCO label, and (b) results trained with the reorganized label for the carryable object.

(a) (b) (c) (d)

Figure 5: Qualitative results of the proposed method: (a) input frame with target person, (b) person and carryable object detection result, (c) keypoints estimation result, and (d) results of action classifiers denoted as red for litterers and white for non-litterers boxes with estimated action labels and confidences over three lines (fine action, target action, coarse action)

Figure 5 shows the estimation results of four different tasks simultaneously. In Figure 5(b), the results of person and carryable object detection are shown using black bounding boxes. As shown in Figure 5(c) and (d), our method can provide the predicted action labels as well as the pose estimation result. Since there is no *'walk'* label in the coarse action dataset, the coarse action classifier tends to predict *'bent'* whenever person's leg is bent. Furthermore, our network can handle frames with multiple persons without noticeable performance degradation by filtering out people that do not hold the object. Figure 6 shows the temporal prediction result of the proposed network. As we expected, the network temporally changes the prediction label in order of 'Carry', 'Just-before', 'Dumping' and 'Normal'.

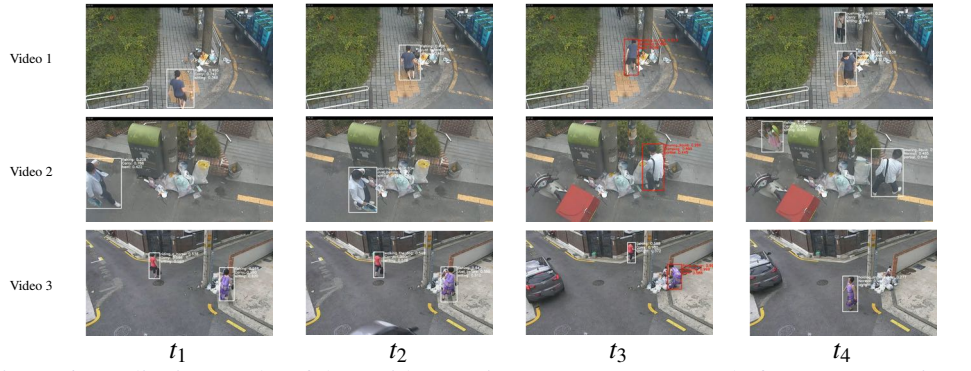

Figure 6: Qualitative results of three videos at time t_1 (Carry), t_2 (Just-before), t_3 (Dumping) and t_4 (Normal). *Zoom in for better view.*

5.4 Limitations and future works

As an initial study of the anti-litter surveillance, we have tried to detect typical dumping action based on the collected dataset, but there are unresolved issues. For example, due to the intrinsic nature of object detectors, small sized garbage may not be detected, so it may have difficulty while being selected as a target. Furthermore, the current model does not use temporal information. Not considering temporal information may cause false alarm in cases like "Someone drops the bag temporally" or "Someone pick up the garbage from the litter pile". If the temporal information are considered in model, anti-litter detection system would reduce these false positives. For further training of network and augmenting the number of training set, some publicly available datasets could be utilized such as ICVL action dataset [14] that includes the dumping action.

6 Conclusion

In this paper, we proposed a novel dumping action detection method to help the monitors for preventing the illegal dumping in the real-world visual surveillance system. We utilized multiple datasets (*i.e.*, coarse action, fine action, and human pose) as well as the surveillance dataset including the dumping action to overcome the lack of real-world dataset. The proposed network was trained by multi-task learning and had produced the synergy of multiple datasets for understanding the human and detecting the dumping action. Furthermore, the newly trained detector for a person and carryable object enhanced the efficiency and the performance in detecting the dumping action. The experimental results showed the best performance compared to the state-of-the-arts in quantitative measures. Since the proposed method not only detected the garbage dumping action, but also provided an understanding of people through a multi-tasking network, it could be used extensively for analyzing events related to people in surveillance scenes.

Acknowledgment

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.B0101-15-0266, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis and No.2020-0-00004, Development of Previsional Intelligence based on Long-term Visual Memory Network).

References

[1] Queensland Litter Prevention Alliance. Report a litter. http://www.keepqueenslandbeautiful.org.au/report-a-litterer. Accessed: 2020-04-22.

[2] B Bhandari, G Lee, and J Choi. Body-Part-Aware and Multitask-Aware Single-Image-Based Action Recognition. *Applied Science*, 10(4), 2020.

[3] Yawei Cai and Xiaoyang Tan. Weakly supervised human body detection under arbitrary poses. In *ICIP*, 2016.

[4] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017.

[5] Jongwon Choi, Hyung Jin Chang, Tobias Fischer, Sangdoo Yun, Kyuewang Lee, Jiyeoup Jeong, Yiannis Demiris, and Jin Young Choi. Context-Aware Deep Feature Compression for High-Speed Visual Tracking. In *CVPR*, 2018.

[6] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, January 1995.

[7] Cardiff Council. Littering. https://www.cardiff.gov.uk/ENG/resident/Rubbish-and-recycling. Accessed: 2020-04-23.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] Environment Protection Authority for New South Wales (NSW EPA). Litter and illegal dumping. https://www.epa.nsw.gov.au/your-environment/litter-and-illegal-dumping. Accessed: 2020-04-22.

[10] World Economic Forum. The new plstics ecomomy: Rethinking the future of plastics. http://www3.weforum.org/docs/WEF_The_New_Plastics_Economy.pdf, 2016. Accessed: 2020-04-23.

[11] Louisville-Jefferson County Metro Government. Report a litter. https://louisvilleky.gov/government/brightside/report-litterer. Accessed: 2020-04-22.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[14] Cheng-Bin Jin, Shengzhe Li, and Hakil Kim. Real-time action detection in video surveillance using sub-action descriptor with multi-cnn. *arXiv preprint arXiv:1710.03383*, 2017.

[15] Jaewon Jung and Jongyoul Park. Improving visual relationship detection using linguistic and spatial cues. *ETRI Journal*, 2019.

[16] In Su Kim, Hong Seok Choi, Kwang Moo Yi, Jin Young Choi, and Seong G Kong. Intelligent visual surveillance — A survey. *International Journal of Control, Automation and Systems*, 8(5):926–939, 2010.

[17] Jiman Kim and Daijin Kim. Accurate abandoned and removed object classification using hierarchical finite state machine. *Image and Vision Computing*, 44:1–14, 2015.

[18] Minji Kim and Sungchan Kim. Robust appearance feature learning using pixel-wise discrimination for visual tracking. *ETRI Journal*, 41(4):483–493, 2019.

[19] Sunoh Kim, Kimin Yun, Jongyoul Park, and Jin Young Choi. Skeleton-Based Action Recognition of People Handling Objects. In *WACV*, 2019.

[20] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

[21] Youngwan Lee and Jongyoul Park. CenterMask: Real-time anchor-free instance segmentation. In *CVPR*, 2020.

[22] Youngwan Lee, Joong-won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and gpu-computation efficient backbone network for real-time object detection. In *CVPR Workshops*, 2019.

[23] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High Performance Visual Tracking With Siamese Region Proposal Network. In *CVPR*, 2018.

[24] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious Attention Network for Person Re-identification. In *CVPR*, 2018.

[25] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[26] Jinyoung Moon, Junho Jin, Yongjin Kwon, Kyuchang Kang, Jongyoul Park, and Kyoung Park. Extensible hierarchical method of detecting interactive actions for video understanding. *ETRI Journal*, 39(4):502–513, 2017.

[27] Phuc Nguyen, Bohyung Han, Ting Liu, and Gautam Prasad. Weakly Supervised Action Localization by Sparse Temporal Pooling Network. In *CVPR*, 2018.

[28] Fatih Porikli, Yuri Ivanov, and Tetsuji Haga. Robust Abandoned Object Detection Using Dual Foregrounds. *EURASIP Journal on Advances in Signal Processing*, 2008 (1):197875–11, 2007.

[29] Pedro F. Proença and Pedro Simões. Taco: Trash annotations in context dataset, 2019. URL http://tacodataset.org.

[30] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. AutoLoc: Weakly-Supervised Temporal Action Localization in Untrimmed Videos. In *ECCV*, 2018.

[31] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[32] Du Tran, Lubomir D Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks,. In *ICCV*, 2015.

[33] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *CVPR*, 2018.

[34] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-Aware Multi-Level Feature Network for Human Object Interaction Detection. In *ICCV*, 2019.

[35] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In-So Kweon. CBAM: Convolutional block attention module. In *ECCV*, 2018.

[36] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[37] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI*, 2018.

[38] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011.

[39] Kimin Yun, Hawook Jeong, Kwang Moo Yi, Soo Wan Kim, and Jin Young Choi. Motion Interaction Field for Accident Detection in Traffic Surveillance Video. In *ICPR*, 2014.

[40] Kimin Yun, Yongjin Kwon, Sungchan Oh, Jinyoung Moon, and Jongyoul Park. Vision-based garbage dumping action detection for real-world surveillance platform. *ETRI Journal*, 39:21–12, 2019.

[41] Kimin Yun, Jongyoul Park, and Jungchan Cho. Robust Human Pose Estimation for Rotation via Self-Supervised Learning. *IEEE Access*, 8(1):32502–32517, 2020.

[42] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal Action Detection with Structured Segment Networks. In *ICCV*, 2017.