Special Paper
방송공학회논문지 제24권 제7호, 2019년 12월 (JBE Vol. 24, No. 7, December 2019)
https://doi.org/10.5909/JBE.2019.24.7.1266
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

Deep Learning Object Detection to Clearly Differentiate Between Pedestrians and Motorcycles in Tunnel Environment Using YOLOv3 and Kernelized Correlation Filters

Sungchul Mun[a)‡], Manh Dung Nguyen[b)], Seokkyu Kweon[b)], and Young Hoon Bae[c)]

Abstract

With increasing criminal rates and number of CCTVs, much attention has been paid to intelligent surveillance system on the horizon. Object detection and tracking algorithms have been developed to reduce false alarms and accurately help security agents immediately response to undesirable changes in video clips such as crimes and accidents. Many studies have proposed a variety of algorithms to improve accuracy of detecting and tracking objects outside tunnels. The proposed methods might not work well in a tunnel because of low illuminance significantly susceptible to tail and warning lights of driving vehicles. The detection performance has rarely been tested against the tunnel environment. This study investigated a feasibility of object detection and tracking in an actual tunnel environment by utilizing YOLOv3 and Kernelized Correlation Filter. We tested 40 actual video clips to differentiate pedestrians and motorcycles to evaluate the performance of our algorithm. The experimental results showed significant difference in detection between pedestrians and motorcycles without false positive rates. Our findings are expected to provide a stepping stone of developing efficient detection algorithms suitable for tunnel environment and encouraging other researchers to glean reliable tracking data for smarter and safer City.

Keywords : Smart City, Surveillance, Deep Learning, YOLOv3, Kernelized Correlation Filter

I. Introduction

Smart City has been emerged as an alternative solution to address the recently raised issues (e.g., aging infrastructure, society, and increasing criminal rates) caused by exponential growth of urbanization and population[1]. Due to the all-connected ICT concept in Smart City, much attention has been given to increasing safety level of surveil-

a) Department of Smart City Research, Seoul Institute of Technology
b) Technical Research Institute, IVS Incorporation
c) Chief Executive Officer, IVS Incorporation
‡ Corresponding Author : Sungchul Mun
E-mail: sungchul.mun@sit.re.kr
Tel: +82-2-6912-0958
ORCID: http://orcid.org/0000-0003-4596-9889
※ This work was supported by Seoul Institute of Technology (SIT) (19-4-5, Development of Crime Detection Technology on CCTV).
· Manuscript received October 31, 2019; Revised December 17, 2019; Accepted December 17, 2019.

lance systems for sustainable Smart City. For the enhanced safety level, it is indispensable to instantly give warnings to security agents in charge of the surveillance system when accidents or crimes occur. Characteristics of extensive camera system and real-time algorithms for processing video data reflecting real-world settings should be considered when developing the selective warning system.

Many studies have been conducted to propose essential algorithms for the real-time surveillance system [2-9]. Xiong et al. [2] proposed a recognition algorithm of different appearances of the same person under dynamic circumstances. They applied a multiple deep metric learning method with modified Softmax regression models, which can calculate probabilities of differences in appearances of the same person. Sharif et al. [3] addressed issues in selecting robust features for recognizing human actions with an innovative method utilizing multi-class correlation and Euclidean distance techniques. Big data framework and camera network system suitable for video surveillance system were proposed by Subudhi et al. [4] and Lee and Kim [6] for the purpose of efficiently storing, retrieving, processing, and analyzing gigantic data coming from real-world security environment. Auto-tracking algorithms using multiple cameras have also been proposed to reflect characteristics of real-world surveillance system [5, 10-14].

However, the previous studies have heavily focused on recognizing and classifying multiple objects in normal situations on the road or public space. In other words, although a lot of studies have been conducted to propose the real-time warning systems, it remains to be seen how researchers and developers address challenging points when recognizing, tracking, and re-identifying specific objects under harsh real world situations. In order to develop intelligent surveillance system covering the whole range of the city, detecting and tracking vehicles and pedestrians in tunnels should be performed while minimizing false positive rates or losing object trajectories. Few trials have been made to develop object detection and tracking techniques in the real-world tunnel environments. In the future Smart City, considering specific characteristics of core sensing technologies in autonomous vehicles, unexpected accident rates might be higher in tunnel environments than in normal driving condition (except for snowy road condition). This study is novel in that we proposed a method to clearly reduce false-positive rates between motorcycles and pedestrians in tunnel environments. High false positive rates in the tunnels are heavily attributed to contamination effects from sudden changes in intensity of illumination caused by headlights reflected on the wall (or on the road) and light intensity distributions of tunnel lamps.

To put it simply, this study was aimed at developing an improved technique to detect pedestrians in a tunnel environment at higher accuracy compared to existing methods, e.g., using YOLO to detect a pedestrian, which falsely detects a motorcycle as a pedestrian sometimes.

II. Methods

Figure 1 illustrates our algorithm. This algorithm includes four major steps, detecting objects, tracking objects, calculating object trajectory smoothness, and finally classifying objects into a pedestrian or motorcycle.

In this paper, we proposed a method of detecting a pedestrian, which combines deep-learning based detection mechanism and object tracking. We employed YOLO v3[15] as a detection mechanism. YOLO is a very popular deep-learning algorithm which performs both detection (locating an object) and classification concurrently. In the network architecture of YOLOv3, input layer of the network is image frame and the outputs are prediction feature map which contains the attributes of a bounding box.

Entire image is divided into grid of cell, and bounding boxes cells can predict are generated.

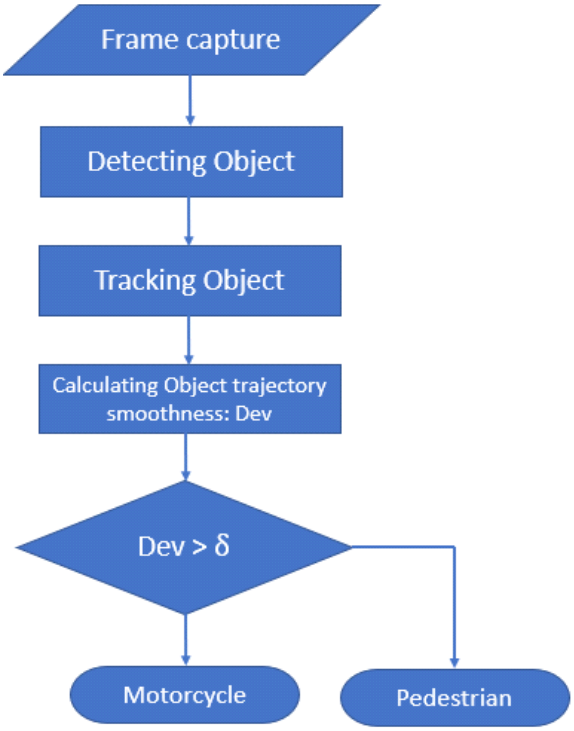

Fig. 1. Algorithm flow chart

Attributes of bounding box include bounding rectangle coordinators, objectness score, and classes score. Once a pedestrian object is detected by YOLOv3, the proposed mechanism starts tracking the object in the next frames to determine whether it is a pedestrian or a motorcycle ridden by a human. Let the nth frame be a frame that an object is detected. A bounding box is defined such that its area contains the object with minimal background.

We trained YOLOv3 model to detect humans only. Humans include pedestrians, human riding motorcycles and human riding bicycles. The dataset included over 50,000 images, which were extracted from COCO dataset and some of which were manually labeled to enhance the detection rate in tunnels.

In the $(n+1)^{th}$ frame, the same object was detected using YOLOv3. An object detected in the $(n+1)^{th}$ frame was defined the same as the original object if its bounding box was overlapped with the original object and the color histogram of the bounding box matched that of the original object. Figure 2 illustrates the overlapping the bounding boxes of the same object in two consecutive frames. In Figure 2b, the red bounding box in the center is the bounding box of a human in $(n+1)^{th}$ frame while the green bounding box in the center in Figure 2a is the bounding box of that human in the nth frame. These two bounding boxes were more than 90% overlapped. Two bounding boxes rc_n^{th} and rc_{n+1}^{th} in the two consecutive frames were defined to belong to the same object if they meet following conditions:

$$S(rc_n^{th} \cap rc_{n+1}^{th}) > \alpha \; x \; S(rc_n^{th}) \tag{1}$$

$$S(rc_n^{th} \cap rc_{n+1}^{th}) > \alpha \; x \; S(rc_{n+1}^{th}) \tag{2}$$

$$d(H_n^{th}, H_{n+1}^{th}) > \beta \tag{3}$$

a b

Fig. 2. bounding box overlapped tracking, a (the nth frame) and b (the (n+1)th frame)

where

S is the area of a bounding box,

H_n^{th} and H_{n+1}^{th} are the color histograms of sub images inside bounding boxes,

$d(H_n^{th}, H_{n+1}^{th})$ is the cross correlation of two histograms, and α and β are tuning thresholds, which are determined through experiment.

The cross correlation (delta) of two histograms was calculated by using below equation:

$$d(H_1,H_2) = \frac{\sum_I (H_1(I) - \overline{H_1})(H_2(I)) - \overline{H_2})}{\sqrt{\sum_I (H_1(I) - \overline{H_1})^2 \sum_I (H_2(I) - \overline{H_2})^2}} \quad (4)$$

where

$$\overline{H_k} = \frac{1}{N}\sum_I H_k(I) \quad (5)$$

If the same object was not detected in the $(n+1)^{th}$ frame, i.e., the object was lost, the object was searched in the $(n+1)^{th}$ frame using a tracking algorithm. In this work, we used the KCF (Kernelized Correlation Filters) algorithm [16] as a tracking algorithm. In other words, we used YOLOv3 to detect pedestrians in every frame using object matching algorithms to match the objects in two consecutive frames. We also used three equation (1), (2), (3) to determine whether objects in two consecutive frames are matched or not. If one object is found in the current frame but its matched object is not found in the next frame, it means that YOLOv3 cannot detect this object in the next frame. In this case, we utilized the KCF algorithm to rediscover the object. If KCF cannot find the object in the next frame, the object will be considered as a disappeared one.

Figure 3 illustrates how to use KCF to predict the location of the object in the next frame. The highest correlation score in the confidence map will be estimated as the centroid of the object in the next frame.

In the next step, we calculated the object's trajectory to determine whether it was a pedestrian or a motorcycle ridden by a human. An object's trajectory was defined as {c(0), c(1), ⋯, c(L)} whether c(i) is the center of the object's bounding box in the i th frame. KCF can predict the bounding box as well as centroid c(i) of an object in next frame if we know the object bounding box in the current

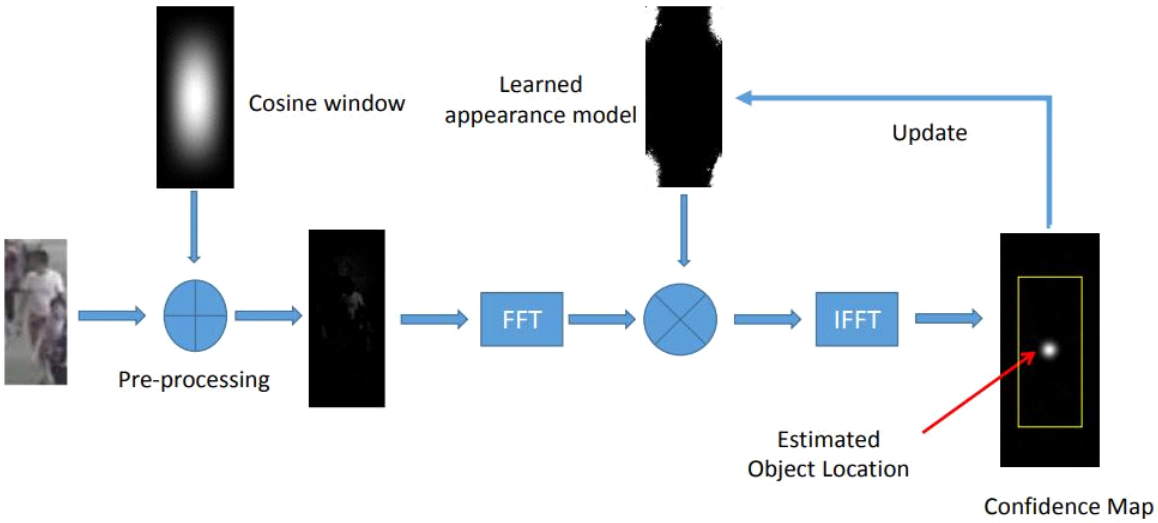

Fig. 3. Kernelized Correlation Filters

frame. Next, we defined two distances, d_{Total} and d_{Real}, as follows:

$$d_{Total} = \sum_{i=0}^{L-1} |c(i+1) - c(i)| \quad (6)$$

$$d_{Real} = |c(L) - c(0)| \quad (7)$$

and next we define deviation, *Dev*, as follows:

$$Dev = d_{Total} / d_{Real} \quad (8)$$

Finally, we defined a criterion to determine whether an object was a pedestrian or a motorcycle as follows:

Object was a pedestrian if

$$Dev \geq \delta \quad (9)$$

Where δ is threshold that was determined by analyzing real trajectories obtained from actual video footages. The meaning of this criterion is self-explanatory since a pedestrian's trajectory is not as smooth as that of a motorcycle as shown in Figure 4. The algorithm was aimed to detect pedestrians in tunnels. We used YOLOv3 algorithm to detect pedestrians, but the YOLOv3 only uses shape information to detect objects. This causes many objects which have very similar shapes as pedestrians such as human riding motorcycles or bicycles to be falsely detected. To reduce the number of false detection rates, we developed a classification algorithm based on object trajectory smoothness. In order to execute the comparative analyses, we used Intel Core -i7 PC and NVIDIA graphic card RTX 2080. The proposed algorithm can apply to tunnel or highway environment, but currently this algorithm only works for the straight way. In case of curved ways, we have to estimate other equations to measure the smoothness of the object trajectory.

Ⅲ. Results

We used 40 videos to evaluate our algorithm, including 20 videos containing pedestrians and the other 20 videos containing motorcycles. The duration of each video is about one minute. We analyzed history of objects for 5 seconds to measure the smoothness of the object trajectory. To put it simply, the runtime of the proposed algorithm took 5 seconds in identifying whether the objects are pedestrians or motorcycles since the algorithm was based on the tracking information of objects during 5 seconds.

Experiment results have shown that pedestrians and motorcycles can be distinguished effectively by comparing smoothness of their movement trajectories. For a motorcycle Dev is close to 1 but for a pedestrian Dev is much

Fig. 4. Trajectories of Pedestrian and Motorcycle

greater than 1. Figure 5 shows some examples of our experiment and Figure 6 is an illustration of experimental summary. The graph in Figure 6 shows that, for pedestrian Dev is always greater than 1.1 but for motorcycle Dev is almost equal 1.

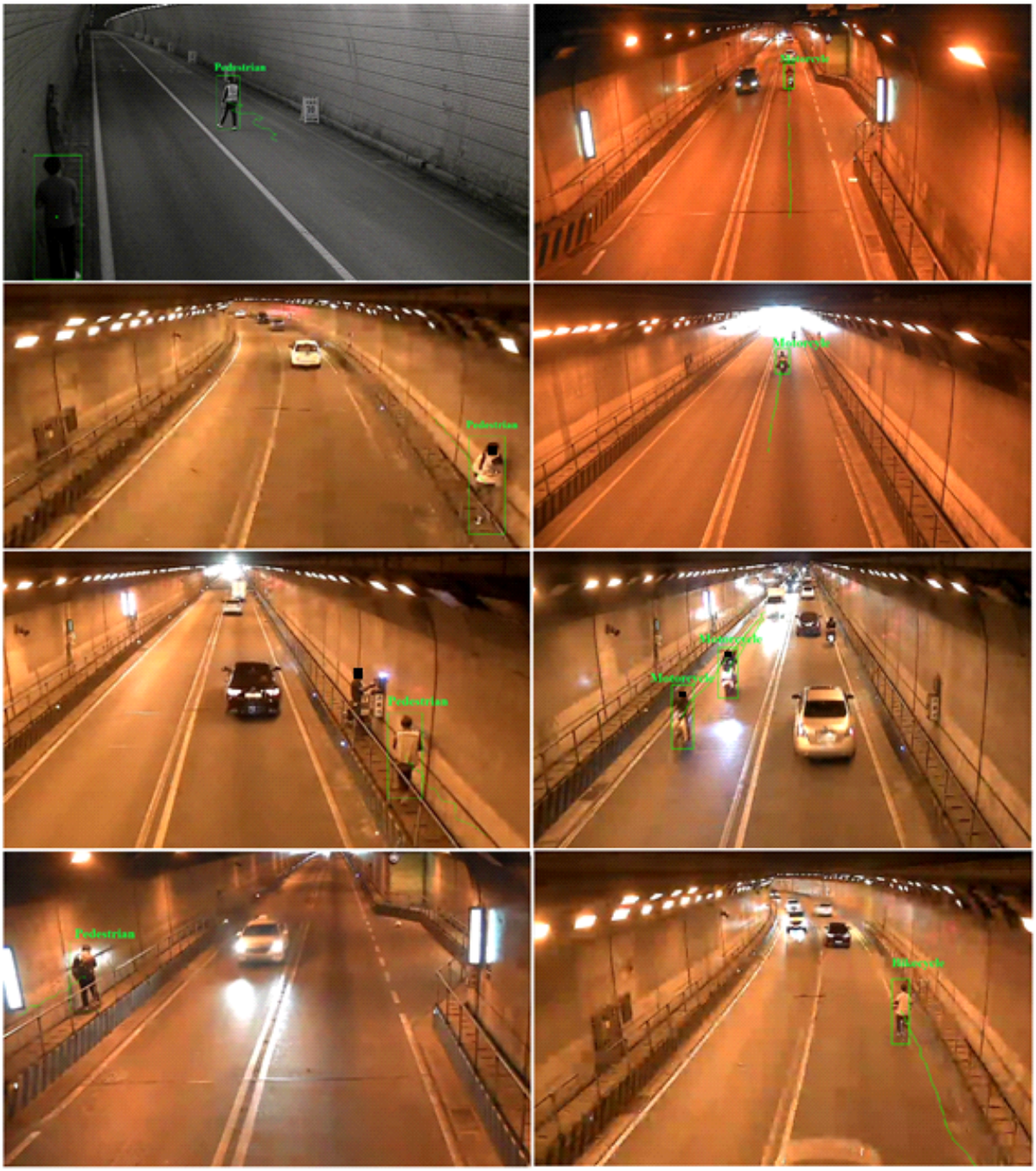

Fig. 5. Experiment Results: sample videos

No	Pedestrian	Motorcycle
1	2.52	1.001
2	1.22	1.002
3	1.24	1.001
4	1.37	1.02
5	1.36	1.003
6	1.83	1.012
7	1.12	1.007
8	1.45	1.008
9	1.17	1.005
10	1.35	1.001
11	1.21	1.004
12	2.16	1.005
13	1.3	1.002
14	1.27	1.006
15	1.34	1.01
16	1.18	1.008
17	2.7	1.003
18	1.9	1.011
19	1.38	1.001
20	1.36	1.022

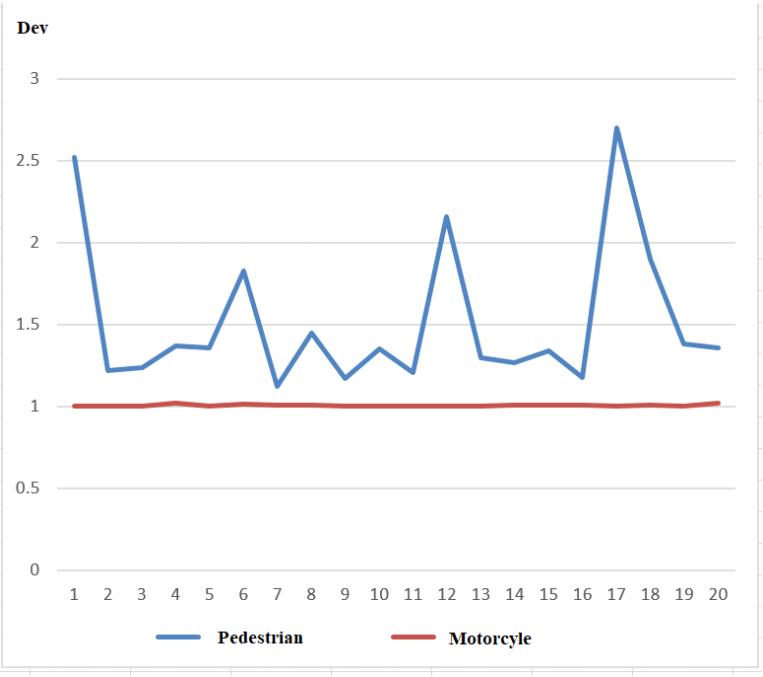

Fig. 6. Experimental summary

To identify statistical reliability of our results, Wilcoxon's Matched-Pairs Signed-Ranks test was used to statistically compare the detection performance between pedestrians and motorcycles because normality assumption was not tenable ($p < .001$). The analysis showed significantly lowered Dev in motorcycle detection compared to that in pedestrian detection ($Z = -3.920$, $p < 0.001$, $r = 0.876$), indicating large effect size (Cohen (1988) criteria of .1 = small, .3 = medium, .5=large).

The experiment results showed that the proposed algorithm can classify pedestrian and motorcycle at 100% success rate. In fact, motorcycles or bicycles are rarely moving following zigzag trajectory because they move at higher speeds. The trajectory of a motorcycle or bicycle usually is a line or curve. In contrast, the trajectory of a pedestrian looks similar to zigzag, so that the deviation of a pedestrian is higher than that of a motorcycle or bicycle. In performance test, we executed our experiment using intel i7 machine with windows 10 and NVIDIA RTX2080. Our system could process 60 frame per second on average. The experiment also record that proposed algorithms require minimum 1.8G CPU RAM and 2.5 GPU RAM.

Ⅳ. Discussion and Conclusion

In this study, we proposed a combined model to detect and track pedestrians accurately in a tunnel environment using YOLOv3, Kernelized Correlation Filter, and empirical rule bases in the real-world settings. Our study has significant points in terms of considerably reducing errors in detecting pedestrians in a tunnel environment where deep learning algorithms very famous for object detection, such as YOLOv3 and SSD often falsely detect motorcycles and bicycles as pedestrians.

Recently, many studies have proposed real-time tracking algorithms and advanced optical flow descriptors to track magnitude changes in pixels between each frame of a spe-

cific video clip. Padmalatha et al. [9] investigated a possibility to detect violent behaviors in real time using an revised Violent Flow Descriptor in which AdaBoost algorithm was used to improve classification of contributing features and detection accuracy. Hasan et al. [17] extracted the conventional spatio-temporal local features and tried to detect anomaly events by learning a fully connected autoencoder, but the proposed model could predict regular patterns with limited supervision circumstances. The previous studies tested their proposed models only in detecting moments of changes in optical flows characterizing violent behaviors mainly happened in public space. Under the actual surveillance environment, accidents, fires or crimes have been frequently occurred in tunnel environment and intelligent surveillance system suitable for tunnel environment to ensure golden time should be developed. Although the methods we proposed did not include all of the technical components for the real time surveillance system in the tunnels, our study is meaningful in that we provided a stepping stone to the future studies for obtaining reliable track data and developing optical flow descriptors for tunnel environments. As a matter of fact, in Republic of Korea, there are 2566 tunnels (1896 km) throughout the nation [18]. Thus, further studies are encouraged to examine feasibility of detecting optical flows of anomaly behaviors and accidents in tunnel environment for selective intelligent surveillance system to monitor harsh real world settings (e.g., dead zones or tunnels).

It will be very helpful to aid security agents to respond quickly to the accidents or crimes, leading to reducing accidental and criminal rates in Smart City environment. However, it should be noted that excessive specification of deep learning servers of intelligent CCTVs causing heavy computational load and time-consuming repetition processing, and excessive price rise of the system should be avoided for the rapid extension of value chains in intelligent CCTV ecosystem. Even though it heavily depends on environmental factors, appropriate technology rather than the cutting-edge technology may be more effective in accurately detecting unexpected accidents or abnormal behaviors in some cases.

Our study has a limitation of experiments in which comparative trials were not enough to fully validate the experimental results because we used actual video footages taken in a real-world tunnel environment. It is very difficult to glean video clips from real-world settings due to Personal Information Protection Act applied on CCTVs.

References

[1] B. N. Silva, M. Khan, and K. Han, "Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities," Sustainable Cities and Society, Vol.38, pp.697-713, Apr, 2018.

[2] M. F. Xiong, D. Chen, J. Chen, J. Y. Chen, B. Y. Shi, C. Liang, and R. M. Hu, "Person re-identification with multiple similarity probabilities using deep metric learning for efficient smart security applications," Journal of Parallel and Distributed Computing, Vol.132, pp.230-241, Oct, 2019.

[3] A. Sharif, M. A. Khan, K. Javed, H. G. Umer, T. Iqbal, T. Saba, H. Ali, and W. Nisar, "Intelligent Human Action Recognition: A Framework of Optimal Features Selection based on Euclidean Distance and Strong Correlation," Control Engineering and Applied Informatics, Vol.21, pp.3-11, Sep, 2019.

[4] B. N. Subudhi, D. K. Rout, and A. Ghosh, "Big data analytics for video surveillance," Multimedia Tools and Applications, Vol.78, pp.26129-26162, Sep, 2019.

[5] R. Iguernaissi, D. Merad, K. Aziz, and P. Drap, "People tracking in multi-camera systems: a review," Multimedia Tools and Applications, Vol.78, pp.10773-10793, Apr, 2019.

[6] G. Y. Lee, and H. J. Kim, "Optimum Configuration of Surveillance Camera System Based on Real Time Image Recognition Server," The Journal of Korean Institute of Communications and Information Sciences, Vol.44, pp.1124-1127, 2019.

[7] A. K. Chandran, L. A. Poh, and P. Vadakkepat, "Real-time identification of pedestrian meeting and split events from surveillance videos using motion similarity and its applications," Journal of Real-Time Image Processing, Vol.16, pp.971-987, Aug, 2019.

[8] M. Lotfi, S. A. Motamedi, and S. Sharifian, "Time-based feedback-control framework for real-time video surveillance systems with utilization control," Journal of Real-Time Image Processing, Vol.16, pp.1301-1316, Aug, 2019.

[9] E. Padmalatha, K. A. S. Sekhar, and D. R. R. Mudiam, "Real Time Analysis of Crowd Behaviour for Automatic and Accurate Surveillance," International Journal of Advanced Computer Science

and Applications, Vol.10, pp.492-496, Mar, 2019.
[10] R. Eshel, and Y. Moses, "Tracking in a Dense Crowd Using Multiple Cameras," International Journal of Computer Vision, Vol.88, pp.129-143, May 01, 2010.
[11] P. M. Roth, C. Leistner, A. Berger, and H. Bischof. Multiple instance learning from multiple cameras. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 13-18 June 2010 2010. 17-24.
[12] A. T. Y. Chen, M. Biglari-Abhari, and K. I. K. Wang, "Investigating fast re-identification for multi-camera indoor person tracking," Computers & Electrical Engineering, Vol.77, pp.273-288, Jul, 2019.
[13] C. C. Sun, M. H. Sheu, J. Y. Chi, and Y. K. Huang, "A Fast Non-Overlapping Multi-Camera People Re-Identification Algorithm and Tracking Based on Visual Channel Model," Ieice Transactions on Information and Systems, Vol.E102D, pp.1342-1348, Jul, 2019.
[14] F. Previtali, D. D. Bloisi, and L. Iocchi, "A distributed approach for real-time multi-camera multiple object tracking," Machine Vision and Applications, Vol.28, pp.421-430, May, 2017.
[15] J. Redmon, and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv preprint arXiv:1804.02767, April, 2018.
[16] Y. Li, and J. Zhu, " A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration," Lecture Notes in Computer Science, Vol.8926, pp.254-265, March, 2015.
[17] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning Temporal Regularity in Video Sequences," arXiv preprint arXiv:1604.04574 April, 2016.
[18] KOSIS National Tunnel Statistics, http://kosis.kr/statHtml/statHtml.do?orgId=116&tblId=DT_MLTM_1040 (accessed Oct. 29, 2019).

저 자 소 개

Sungchul Mun

- Feb. 2012 : M.S. Emotion Engineering, Sangmyung University
- Feb. 2015 : Ph.D. HCI & Robotics (Data Science for Ergonomics), Korea Institute of Science and Technology (UST)
- Mar. 2015 ~ Nov. 2016 : Post Doc. National Agenda Research Division, Korea Institute of Science and Technology
- Dec. 2016 ~ Jun. 2017 : Chief Research Engineer, Strategy Business Office, Golfzon Newdin Group
- Jul. 2017 ~ Jun. 2019 : General Manager, Future Engine Lab., CJ Hello
- Jun. 2019 ~ Present : Chief Research Scientist, Department of Smart City Research, Seoul Institute of Technology
- ORCID : https://orcid.org/0000-0003-4596-9889
- Research interest : Deep Learning, Video Analytics, Human Factors, HCI, Smart Healthcare

Manh Dung Nguyen

- Feb. 2009 : M.S. Information and Communication, Kongju University
- Dec. 2019 : Ph.D. Information and Communication, Kongju University
- Feb. 2011 ~ Present : Senior Software Engineer, IVS Inc.
- ORCID : https://orcid.org/0000-0001-6165-4137
- Research interest : Deep Learning, Sound Analytics, Video Analytics

Seokkyu Kweon

- Feb. 1991 : M.S. Electronics, Seoul National University
- Dec. 1998 : Ph.D. Electrical Engineering and Computer Science, University of Michigan, Ann Arbor
- May 2000 ~ Feb. 2003 : Software Engineer, Cisco Systems
- Mar. 2003 ~ Feb. 2013 : Team Leader, Samsung Electronics
- May. 2013 ~ Sep. 2015 : VP of Software Development, Samsung Techwin
- Sep. 2015 ~ Dec. 2018 : Head of R&D Center, SMC Networks
- Dec. 2018 ~ Present : Head of R&D Center, IVS Inc.
- ORCID : https://orcid.org/0000-0002-3525-2801
- Research interest : Deep Learning, Sound Analytics, Video Analytics

저 자 소 개

Young Hoon Bae

- Feb. 1981 : M.S. Mechanical Design Department, Seoul National Graduate school
- Oct. 1981 ~ Aug. 1986 : Team leader, Hyundai Engineering ICAE Team
- Jun. 1986 ~ Jun. 1991 : Team leader, Samsung Engineering ICAE Team
- Jun. 1991 ~ Aug. 1998 : General Manager, Samsung SDS CAD/CAM Division
- Mar. 2008 ~ Mar. 2010 : CEO, KiryungElectronics Co., Ltd.
- May 2010 ~ Present : CEO, IVS Inc. / CEO
- ORCID : https://orcid.org/0000-0003-2507-0438
- Research interest : Deep Learning, Sound Analytics, Video Analytics