Journal of Multimedia Information System VOL. 7, NO. 3, September 2020 (pp. 205-214): ISSN 2383-7632 (Online)
http://doi.org/10.33851/JMIS.2020.7.3.205

Detection of Dangerous Situations using Deep Learning Model with Relational Inference

Sein Jang[1], Lkhagvadorj Battulga[2], Aziz Nasridinov[3*]

Abstract

Crime has become one of the major problems in modern society. Even though visual surveillances through closed-circuit television (CCTV) is extensively used for solving crime, the number of crimes has not decreased. This is because there is insufficient workforce for performing 24-hour surveillance. In addition, CCTV surveillance by humans is not efficient for detecting dangerous situations owing to accuracy issues. In this paper, we propose the autonomous detection of dangerous situations in CCTV scenes using a deep learning model with relational inference. The main feature of the proposed method is that it can simultaneously perform object detection and relational inference to determine the danger of the situations captured by CCTV. This enables us to efficiently classify dangerous situations by inferring the relationship between detected objects (i.e., distance and position). Experimental results demonstrate that the proposed method outperforms existing methods in terms of the accuracy of image classification and the false alarm rate even when object detection accuracy is low.

Key Words: Closed-circuit television, Crime, Danger detection, Smart cities

I. INTRODUCTION

Crime is one of the major problems in modern society. For example, according to the crime statistics of the Korea Supreme Prosecutor's Office [9], there were 3,884.8 crimes per 100,000 people in 2016. The total crime incidence in the past ten years has decreased to 2.6%; however, this rate includes traffic crimes, which account for a significant proportion of all crimes. Excluding traffic crimes, the overall crime rate has increased to 11.2% (refer to Figure 1). In particular, over the past decade, violent crimes (e.g., murder, robbery, arson, assault, sexual assault) have increased compared to other crimes.

One of the most efficient methods of preventing crimes is visual surveillance through closed-circuit television (CCTV). However, there are numerous difficulties in CCTV surveillance by humans. For example, there is an insufficient workforce for performing CCTV surveillance. Additionally, it is difficult for humans to rapidly and accurately detect dangerous situations. In other words, it is almost impossible for a human to focus only on crime surveillance through CCTV by monitoring it for 24 hours. Owing to this problem, there are several limitations in crime prevention, even if a large number of CCTV systems are installed. In other words, CCTV is frequently used for finding evidence or suspects after a crime has occurred rather than for crime prevention. In recent years, deep learning has become a widely used technology for automatic visual surveillance through CCTV. In this study, we utilize deep learning to solve the aforementioned problems.

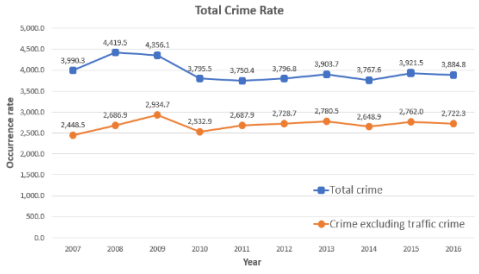


Fig. 1. Korea supreme prosecutors' office crime analysis [9].

Manuscript received September 03, 2020; Revised September 25, 2020; Accepted September 25. 2020. (ID No. JMIS-20M-09-023)

Corresponding Author (*): Aziz Nasridinov, Department of Computer Science, Chungbuk National University, Cheongju 28644, Korea, aziz@chungbuk.ac.kr

[1]Data Fusion Research Center (DFRC), South Korea, sein@dfrc.ch

[2]Department of Mathematics and Computer Science, Eindhoven University of Technology, 5612 AZ, Eindhoven, The Netherlands; lhagvaa1221@gmail.com

[3]Department of Computer Science, Chungbuk National University, Cheongju 28644, Korea, aziz@chungbuk.ac.kr

To accomplish this goal, we first define the problems that can occur in danger detection through CCTV and then apply deep learning to solve these problems. Furthermore, we conduct a comparative analysis of various state-of-the-art methods on dangerous situation classification. This work is the next step towards achieving our final goal of crime prevention through visual surveillance by CCTV without human assistance presented in [7]. More specifically, the contributions of this study are as follows:

- We propose a deep learning model for the autonomous detection of dangerous situations in CCTV scenes.
- We utilize relational inference, which considers the position of objects in CCTV and the distance between the objects. The relational inference enables the model to more accurately classify a situation according to the relationship (i.e., distance and position) of each detected dangerous object.
- Experimental results demonstrate that the proposed model outperforms other baseline models in terms of the accuracy of image classification and the false alarm rate. In addition, the proposed model is efficient even if object detection accuracy is low.

The rest of the paper is organized as follows: The related studies that focus on deep learning approaches for danger detection and their shortcomings are discussed in Section 2. The proposed model is described in Section 3. Performance evaluation is presented in Section 4. Finally, Section 5 provides the conclusions and direction of future work.

II. Related Work

There are a variety of approaches on danger detection that use deep learning. We can divide them into two main approaches for danger detection through CCTV, i.e., image classification and object detection. The details of each approach and their shortcomings are described in the following subsections.

2.1. Image Classification Approach for Danger Detection

Image classification approaches generally analyze an image using various techniques and determine if it is a dangerous situation. [12] proposed a convolutional long short-term memory (Conv-LSTM) network for anomaly detection that predicts future frames by reconstructing an image. In experiments, the Conv-LSTM network shows competitive performance compared with other state-of-the-art anomaly detection methods. [1] presented a novel framework for the anomaly detection of a scene. The framework receives a video as an input and performs several processes, such as background subtraction, object tracking, and scene analysis, for detecting abnormal situations. The framework outperforms other frameworks in experiments using actual video data sets. [3] proposes an algorithm to classify safe and dangerous situations according to the existence of a knife in an image. The authors use a sliding window mechanism to obtain the features of the knife in the image and utilize visual descriptors from MPEG-7 for feature data extraction. The extracted feature data are input to Support Vector Machine (SVM) for classification. As a result, the false alarm rate of knife detection is 7%. [20] proposed a real-world abnormal behavior detection system that uses deep learning. This method models normal and abnormal events as instance bags, train these instances using a deep anomaly ranking model, and predict high anomaly scores for anomalous video segments. Here, abnormal behaviors are detected by extracting features related to the amount of change of actions in the whole frame. In [2], we argue that in specific crime behaviors, such as shoplifting, there is not much change in the action of the user. To solve this problem, we proposed to first extract a person object as a region of interest (ROI) using Mask-R-Convolutional Neural Network (Mask-R-CNN) [4] and then determine crime behaviors using the amount of change in the ROI person object using optical-flow. [21] proposed to detection motion anomalies in surveillance videos using Region Association Graph (RAG). The authors extracted statistical features from given scene and classified two types of anomalies (i.e., an object encircles within a particular region or within a set of regions) using SVM algorithm. The experiment results using two benchmark datasets indicate the effectiveness of the proposed method.

2.2. Object Detection Approach for Danger Detection

Object detection approaches perform visual surveillance by identifying dangerous objects and actions from CCTV images. [14] proposes a simple method that detects knives, blood, and guns. The detected objects are closely related to crime; thus, we can predict whether a dangerous situation has occurred. For this purpose, the authors implement a CNN that contains the Rectified Linear Unit (ReLU) nonlinearity, a convolution layer, a fully connected layer, and a dropout. The object detection model exhibits a test accuracy of 90.2%. There are also several issues related to object detection in crime scenes. For example, there is a possibility of occlusion in object tracking, which can significantly reduce crime detection rates. To solve occlusion problems, [19] proposed a scale-adaptive object-tracking algorithm. The authors achieve high accuracy in occlusion detection due to features extracted ResNet, and more efficient update strategy compared with correlation filters (CFs) that are frequently used in occlusion detection.

Recently, aerial video surveillance using various devices, such as drones, are actively studied. For example, [13] proposes “i-SURVEILLANCE,” which can alert a human

operator when a dangerous action and abnormal behavior are detected. The authors focus on reducing the number of false alarms. The detailed process of the i-SURVEILLANCE system is as follows: First, it obtains a frame scene from a CCTV video. Second, the motion in the scene is detected through background subtraction. Third, motion tracking is achieved by utilizing a Kalman filter. Finally, the motion behavior obtained from previous processes is analyzed using a behavioral analysis algorithm to check for any dangerous action or abnormal behavior. [15] uses the TensorFlow's Object Detection API for object detection and proposes an autonomous unmanned aerial vehicle video surveillance system to monitor suspicious activities through human pose estimation. The proposed model detects people, extracts the pose of each person, and then matches it with the pose in a suspicious action dataset. [18] proposes the aerial surveillance of dangerous situations in a public area using object detection through drones. Visual surveillance is performed by defining five violent activities, i.e., punching, stabbing, shooting, kicking, and strangling, and detecting the defined activities through the drone. The authors use a feature pyramid network for person detection and the ScatterNet hybrid deep learning network to estimate the pose of each detected person.

2.3. Problems of Previous Work

Existing surveillance approaches may show good performance. However, they still experience problems in certain cases. One of the problems is that image classification, and object detection approaches rely strongly on the performance of each technique. For example, in image classification approaches, if CCTV resolution is low and dark, the data obtained through an image may be reduced, and thus, the accuracy of visual observation becomes poor. In object detection approaches, if a system fails to detect an object that is defined as dangerous, it will also fail to perform visual surveillance. In other words, the system cannot generate a warning about danger because it does not recognize that a dangerous object exists in a CCTV image. Therefore, object detection models must have high accuracy. According to [5], typically, an object detection model that can provide real-time speed with a low accuracy compared with other models. In addition, object detection performance decreases with image resolution. However, model must simultaneously produce high speed and accuracy to perform automatic danger detection. This paper proposes a deep learning model for danger detection that simultaneously performs image classification and object detection.

Another limitation of the previous work is that they do not consider the relationship between the detected objects in CCTV images. In this paper, we demonstrate that considering the relationship between the detected objects in CCTV images can improve the accuracy of dangerous situation. This is because even if the same objects are detected in different CCTV images, as shown in Figure 2, the situation differs depending on the position of the objects. For example, even if a knife is detected in CCTV images, the situation in which a person is holding the knife and the situation in which the knife is on the floor are different. Therefore, we must infer the relationship (i.e., distance and position) of each detected dangerous object.

(**a**) (**b**)

Fig. 2. Different situations depending on the position of dangerous object (i.e., knife): (a) Crime situation because a person is holding a knife; (b) Possible dangerous situation, but not as critical as (a).

III. Proposed Model

The overall structure of the proposed model is shown in Figure 3. The proposed model can be divided into two steps, i.e., the object detection step and relational inference step. In the first step, we use object detection to determine the factors that make a situation dangerous and determine the danger of the situation through CCTV images. In the second step, the relationship between detected dangerous objects and human actions in CCTV images is inferred through a modified relation network (RN) [17] so that the model can classify the situation more accurately. The composite function of the proposed model is as follows:

$$D = f_{\phi}(\sum_{k} g_{\theta}(i, o_k)), \quad (1)$$

where the inputs are a latent feature of entire image i and the set of the latent features of each detected object $O = \{o_1, o_2, \ldots, o_k\}$. f_{ϕ} and g_{θ} are Multilayer Perceptrons (MLPs) that have the parameters of learnable synaptic weights. Based on Equation 1, the proposed model can consider the potential relations between detected object pairs. In other words, the proposed model learns to infer the existence of the relationships between detected objects. The detailed process of each step is discussed in the next subsections.

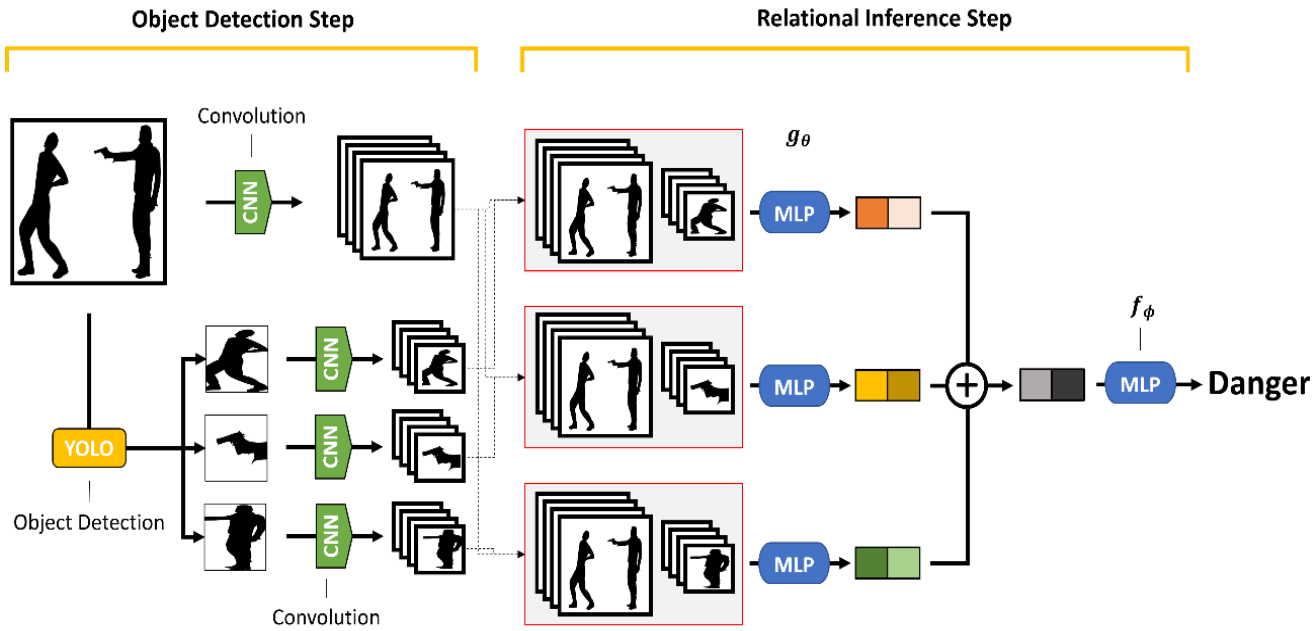


Fig. 3. Overall structure of the proposed method.

3.1. Object Detection Step

We use the YOLO (You Only Look Once) model for object detection [16]. YOLO is one of the fastest object detection models. The output of YOLO contains three features, i.e., the class, coordinate, and boundary box information of a detected object. Thus, we obtain the partial image of a detected object by cropping it from the entire image.

Then, CNN [11] is utilized to obtain a set of latent features from each partial image of the detected object. CNN is one of the most influential deep learning models in the field of computer vision. The performance of CNNs was proven in the 2012 ImageNet competition in terms of reducing classification error from 26% to 15% [10]. Since then, it has become an important model for deep learning research in image processing. Although there are numerous advances in CNN [6, 8], we selected the traditional CNN algorithm due to its simplicity and efficiency in obtaining features from the image. Furthermore, we obtain the latent features for not only the detected object but also the entire image. This process is employed to generate the input features for the relational inference step.

3.2. Relational Inference Step

We use a two each four-layer MLP for inferring the relationship between detected objects. We modify the RN proposed by [17]. The RN was originally proposed for the task of visual question and answering, particularly for answering relational questions. As our model must infer the relationship between detected objects, we use a modified structure of the RN that only considers detected objects and not every part of an image. To infer the relationship between detected objects, we create pairs of the set of the latent features from the entire image, each detected object image object classes and coordinate of detected object. All pairs are the inputs of MLP g_θ, and we perform the element-wise sum of all outputs of g_θ. This result is the input of MLP f_ϕ, which infers the relationship between detected objects. As we input the coordinate of detected objects, the model can learn the relationship between the position of objects and the distance between objects.

3.3. Advantages of the Proposed Model

In this work, we focus on two problems of danger detection. The first is that the model depends on the performance of object detection, and the second is that the relationship between detected objects should be considered.

To solve the first problem, we train the model by inserting not only the images of detected objects but also the entire image. Owing to this, even if object detection fails, the lack of information can be compensated through the added data of the entire image. In addition, we obtain the latent features from each image of detected objects through a CNN, and the model can classify a situation in more detail. As shown in Figure 4, the information about the motion of a detected object can be obtained through its image. The motion of the detected object can be used as the key factor for classifying the danger of a situation. This is because the situation may change according to motion of the detected object even if the same object is detected in different images.

(**a**) (**b**) (**c**)

Fig. 4. Difference in danger due to motion of object.

Furthermore, we propose a modified structure of the RN to infer the relationship between detected objects, where it is possible to classify a dangerous situation according to the relationship between objects.

IV. Performance Evaluation

In this section, we describe the experiments that are conducted to demonstrate the performance of our proposed model. The data set used in the experiment, the baseline models that are compared with the proposed model, and the evaluation methods and results are described in detail. The experiments are conducted in two parts. First, we compare the accuracy of dangerous situation classification through the analysis of image data by the proposed model and baseline models. In addition, we examine the accuracy of dangerous situation classification by the proposed model and baseline models as object detection accuracy decreases. In the second experiment, we compare the false alarm rate of dangerous situation classification by the proposed model and baseline models.

4.1. Experimental Setup

In this subsection, we explain data set used in experiments and the details of competing methods, including their hyperparameter settings.

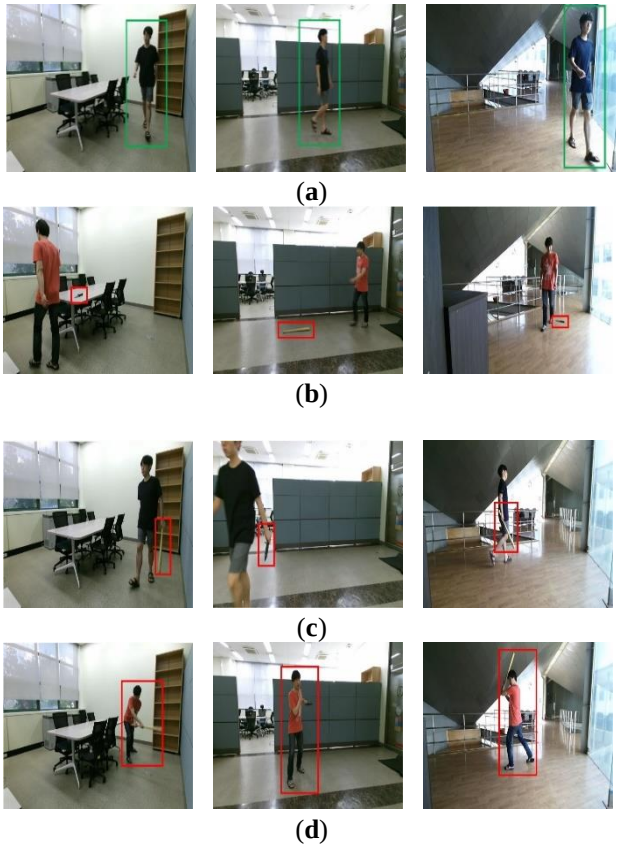
4.1.1. Dataset

We create an experimental dataset to compare the performance of the proposed model and baseline models. The main purpose of the proposed model is to detect dangerous objects or dangerous actions in image data to determine the danger of a situation and classify the situation according to the danger level. Therefore, we first define the dangerous objects and actions that the model should detect in an image.

Dangerous objects are defined by utilizing the crime statistic public data provided by the Korean government [1]. We utilize twelve of the most frequently used objects in crime as dangerous objects. These objects are rifle, handgun, knife, bat, laptop, phone, book, bottle, umbrella, chair, broom, and bag, as shown in Figure 5.

Fig. 5. List of defined dangerous objects.

Further, to classify the danger of a situation for detailed danger assessment. We classify a situation into the following four classes based on the danger of the situation: safe, dangerous tool, potential crime, danger as shown in Figure 6.



(**a**)

(**b**)

(**c**)

(**d**)

Fig. 6. Four classes of data. (a) Safe; (b) Dangerous tool; (c) Potential crime; (d) Danger.

Here, "safe" indicates that there is no danger -> no dangerous object. We create a "safe" data set by recording a video of a person walking around. "Dangerous object-> dangerous tool" implies that dangerous objects exist in an image, and thus, there is a possibility that dangerous situations may occur if a person uses these objects. A "potential crime" is a situation where a person is holding a dangerous object in an image. This class indicates more danger compared to the second class. Finally, a "dangerous -> danger " situation is one where a dangerous object and a dangerous action are detected, for example, a person

wielding a knife or bat, which is classified as the most dangerous situation. We record videos in two locations to prevent overfitting by learning more data about the four defined situations. Additionally, two versions of the videos are recorded by two different people for each situation. Thereafter, we extract the image data for each frame in the videos and label each situation corresponding to the image data. We use 3600 images for the experiment, among which 70% of the images are utilized for training and 30% for validation.

Moreover, we generate data for carrying out an experiment to compare the changes in the accuracy of danger situation classification by the models when there is a decrease in object detection accuracy, which is our second experimental objective. For this purpose, we randomly remove part of the detected object data acquired from the previously obtained image data. More specifically, by randomly removing 10%, 20%, and 30% of data, we assume the accuracy of the object detection model as 90%, 80%, and 70%, respectively.

4.1.2. Learning Method

To search for the best model, we iteratively change the parameters of training algorithms to improve a loss function. We train the deep learning models for the experiment using stochastic gradient descent provided by the TensorFlow open-source machine learning library. The performance reported in our experiment is achieved with the following learning parameters:

- Learning rate: 2.5e-4
- Weight decay: 1e-4
- Momentum: 0.9
- Batch size: 64

We use a GPU that improves the overall speed of model training.

4.1.3. Baseline Model

The performance of the proposed model is compared with that of the following baseline models:

- **SVM**: This is a supervised learning model in machine learning.
- **Decision Tree**: This is a predictive modeling approach in machine learning. A model predicts the value of a target variable based on several input variables. This model is also used for classification.
- **Random Forest**: This is an ensemble learning method for classification. Overfitting can be reduced by utilizing a multiple decision tree model.
- **Gradient Boosting**: This is a machine learning technique for classification, which is the ensemble of weak prediction models.
- **MLP (class)**: This uses only the class of detected dangerous objects (e.g., knife:1, bat:2, etc.).
- **MLP (class + image)**: This uses not only the class of detected dangerous objects but also their partial images.
- **RN**: For relational inference, the modified RN is applied to detect objects only, without using the data of an entire image.

4.1.4. Evaluation Metrics

We use accuracy and false alarm to evaluate the performance of the proposed model. The details of these evaluation metrics are given in Equation 2 and 3, respectively. Here, accuracy is the number of correct predictions made by the model. A false alarm is the sum of the cases where true is identified as false, and false is identified as true, which is the number of cases where the model incorrectly predicts an actual situation. In our experiment, this implies that the model predicts a safe situation as a dangerous situation and a dangerous situation as a safe situation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

$$False\ Alarm = FP + FN \tag{3}$$

In Equation 2 and 3, TP indicates true positive, TN indicates true negative, FP indicates false positive, and FN is a false negative of the confusion matrix, which is frequently used as an evaluation metric of machine learning algorithms.

4.2. Experimental Result

Figure 7 shows the accuracy of dangerous situation classification by the models for object detection accuracies of 100% and 70%. Here, the x-axis indicates competing models described in Section 4.1.3, and the y-axis indicates classification accuracy according to Equation 2.

The performance of all models is high when object detection accuracy is 100%. This is because all models can obtain sufficient data to classify dangerous situations. However, as object detection accuracy decreases by 10% interval, the accuracy of classifying dangerous situations also steadily decreases. Nevertheless, the accuracy of dangerous situation classification by the proposed model is consistently better compared to other models. Furthermore, as object detection accuracy decreases, the classification accuracy of the proposed model decreases only slightly, whereas that of other models decreases considerably. Hence, we can conclude that the proposed model does not depend on object detection.

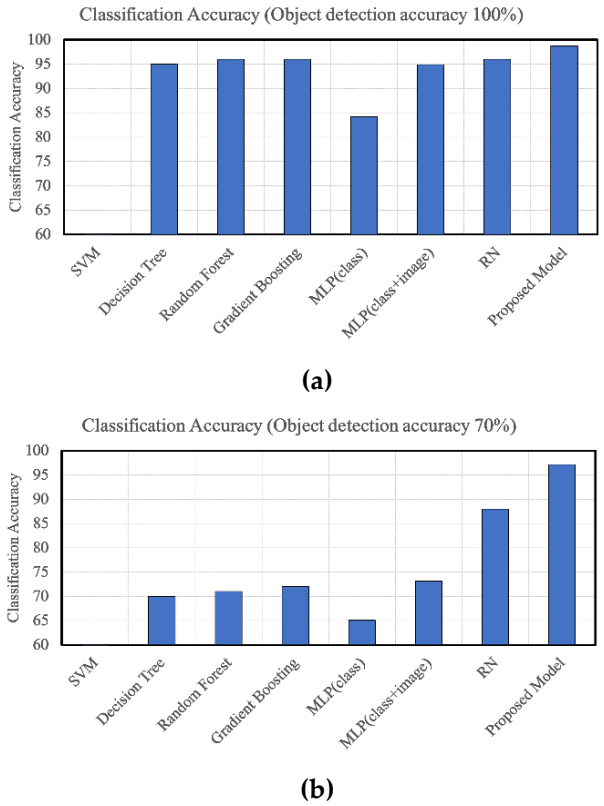


Fig. 7. Performance evaluation of dangerous situation classification when object detection accuracy is: (a) 100%; (b) 70%.

Figure 8 and Table 1 show the variation in the classification accuracy of the proposed model and other models with object detection accuracy. In Figure 8, the x-axis indicates the accuracy of the object detection model, and the y-axis indicates classification accuracy. From the graph, we can observe that the danger situation classification accuracy of the proposed model is not significantly affected by object detection accuracy, as compared to the baseline models.

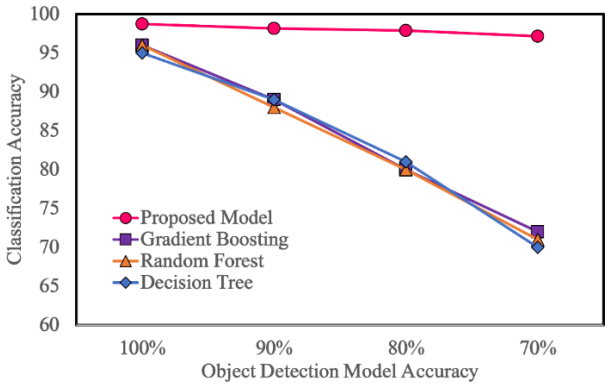


Fig. 8. Changes in accuracy of dangerous situation classification according to accuracy of object detection model.

Table 1. Accuracy of dangerous situation classification.

Model	Object Detection Model Accuracy			
	100%	90%	80%	70%
SVM	31%	27%	25%	23%
Decision Tree	95%	89%	81%	70%
Random Forest	96%	88%	80%	71%
Gradient Boosting	96%	89%	80%	71%
MLP (class)	84%	74%	71%	65%
MLP (class + image)	94%	89%	81%	73%
RN [17]	96%	91%	90%	88%
Proposed Model	98%	98%	97%	97%

The experimental results for the false alarm rate are shown in Table 2. It is important to minimize the false alarm rate to use the proposed model in real life. In the experiment, we consider the occurrence of two types of false alarms out of 720 test data sets. In the first type, the model predicts a safe situation as a dangerous situation. In the second type, the model predicts a dangerous situation as a safe situation. As shown by the results, the proposed model is the most stable because it exhibits the lowest false alarm rate.

Table 2. False alarm rate of dangerous situation classification.

Model	Object Detection Model Accuracy			
	100%	90%	80%	70%
MLP (class)	4.3%	6.3%	5.1%	9.8%
MLP (class + image)	8%	10.1%	15.8%	15.4%
RN [17]	2.5%	2.2%	3.3%	5.3%
Proposed Model	0.9%	1.6%	1.8%	2.1%

In summary, we generate experimental data to evaluate the performance of the proposed model and compare it with several baseline models. Experimental results show that the proposed model outperforms the baseline. Even when object detection accuracy is low, the classification accuracy of the proposed model is better compared to the baseline models. Consequently, the proposed model can overcome the limitation of the dependence of dangerous situation classification on object detection.

V. Conclusion and Future Work

We propose a method that utilizes deep learning. The proposed method overcomes the problem of existing danger detection methods, i.e., the strong dependence of dangerous situation classification on object detection. Experiments performed using actual video data sets show that the proposed method outperforms other methods. In addition, when object detection accuracy deteriorates, the proposed method exhibits better accuracy of danger situation classification compared with other methods. We performed experiments using actual video data sets that we generated.

In future, experiments must be performed using CCTV datasets to evaluate the suitability of the proposed model for real-world applications. For example, there is a publicly available UFC Crime dataset. In addition, it would be interesting to develop an end-to-end learnable model for danger detection through automatic visual surveillance, instead of using an external object detection model (i.e., YOLO).

Acknowledgments

This work was supported by the Institute for Information and Communications Technology Promotion (IITP) Grant funded by the Korean Government (MSIT) (SIAT CCTV Cloud Platform) under Grant 2016-0-00406. Also, this work was supported by the Industrial Strategic Technology Development Program (No.200003991, Development of Korean Wave Convergence Service for AI-based Motion Evaluation and Learning Technology) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea). Further, this work is an extension of a student abstract presented in AAAI2019 [7]. Lastly, we would like to extend our gratitude to the researchers from Multidimensional Insight Lab of Yonsei University for the dataset that we used in performance evaluation.

REFERENCES

[1] A. Basharat, A. Gritai, M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, USA, pp. 1–8, 2008.

[2] U.-J. Gim, J.-J. Lee, J.-H. Kim, Y.-H. Park, A. Nasridinov, "An Automatic Shoplifting Detection from Surveillance Videos," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI2020)*, New York, USA, pp. 13795-13796, February 7-12, 2020.

[3] M. Grega, A. Matiolański, P. Guzik, M. Leszczuk, "Automated detection of firearms and knives in a CCTV image," *Sensors*, vol. 16, no. 1, pp. 1-16, 2016

[4] K. He, G. Gkioxari, P. Doll´ar, R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 2961–2969, October 22-29, 2017.

[5] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 7310–7319, 2017.

[6] D. Jeong, B.-G. Kim, S.-Y. Dong, "Deep Joint Spatiotemporal Network (DJSTN) for Efficient Facial Expression Recognition," *Sensors*, vol. 20, no. 7, 2020.

[7] S. Jang, Y.-H. Park, A. Nasridinov, "AVS-Net: Automatic Visual Surveillance Using Relation Network," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI2019)*, Honolulu, USA, pp. 9947-9948, January 27-February 1, 2019.

[8] J. –H. Kim, B. -G. Kim, P. P. Roy, D. Jeong, "Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure," *IEEE Access*, vol. 7, pp. 41273-41285, 2019.

[9] Korea supreme prosecutors' office crime analysis, http://www.spo.go.kr/, 2017.

[10] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceeding of Neural Information Processing Systems (NIPS)*, Nevada, USA, pp. 1106-1114, December 3-8, 2012.

[11] Y. LeCun, P. Haffner, L. Bottou, Y. Bengio, "Object recognition with gradient-based learning," *Shape, Contour and Grouping in Computer Vision*, vol. 1681, pp. 319–345, 1999.

[12] J. R. Medel, A. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," arXiv preprint arXiv:1612.00390.

[13] M. Nair, G. Mathew, J. Neethu, J. Davies, "i-Surveillance crime monitoring and prevention using neural networks," *International Research Journal of Engineering and Technology*, vol. 5, no. 3, pp. 1231–1236, 2018.

[14] M. Nakib, R. T. Khan, M. S. Hasan, J. Uddin, "Crime scene prediction by detecting threatening objects using convolutional neural network," in *Proceedings of the International Conference on Computer, Communication, Chemical, Material and Electronic Engineering*, Rajshahi, Bangladesh, pp. 1–4, 2018.

[15] S. Penmetsa, F. Minhuj, A. Singh, S. N. Omkar, "Autonomous UAV for suspicious action detection using pictorial human pose estimation and classification," *Electronic Letters on Computer Vision and Image Analysis*, vol. 13, no. 1, pp. 18–32, 2014.

[16] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: Unified, real-time object detection," in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nevada, USA, pp. 779–788, June 26 - July 1, 2016.

[17] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, "A simple neural network module for relational reasoning," in *Proceeding of Neural Information Processing Systems*, California, USA, pp. 4967–4976, Dec 4-9, 2017.

[18] A. Singh, D. Patil, S. N. Omkar, "Eye in the Sky: Real-time Drone Surveillance System (DSS) for Violent Individuals Identification using ScatterNet Hybrid Deep Learning Network," arXiv preprint arXiv:1806.00746.

[19] Y. Yuan, J. Chu, L. Leng, J. Miao, B.-G. Kim, "A scale-adaptive object-tracking algorithm with occlusion detection," *EURASIP Journal on Image and Video Processing*, vol. 7, no. 1, pp. 1-15, 2020.

[20] Y. Zhu, S. Newsam, "Motion-aware feature for improved video anomaly detection," in *Proceedings of British Machine Vision Conference (BMVC2019)*, Cardiff, UK, pp. 1-12, September 9-12, 2019.

[21] M. Chebiyyam, R. D. Reddy, D. P. Dogra, H. Bhaskar, L. Mihaylova, "Motion anomaly detection and trajectory analysis in visual surveillance," *Multimedia Tools and Applications*, vol. 77, pp. 16223–16248, 2018.

Authors

Sein Jang received his Bachelor and Master of Science in Computer Science from Chungbuk National University, South Korea. During his master's studies, he specialized in the field of recommendation system, machine learning and deep learning. In addition, he worked as a System Developer/Data Analyst on the Smart CCTV project, which was funded by the Korean government. Currently, he works as a data scientist at Data Fusion Research Center (DFRC) in Korea.

Lkhagvadorj Battulga received his Bachelor of Science in the field of Information Technology from the National University of Mongolia in 2015. After receiving his bachelor's degree, He worked as an Information Security Analyst in a telecommu-nication company for a year. In 2018, he received his Master of Science in Computer Science from Chungbuk National University, South Korea. During his master's studies, he specialized in the field of Data Analysis, Machine Learning, and Artificial Intelligence. In addition, he worked as a System Developer / Data Analyst on the Smart Factory 4M project, which was funded by the Korean government. Currently, he is pursuing his Professional Doctorate of Engineering degree in Eindhoven University of Technology, The Netherlands. He is mainly focused on creating a relation between his scientific and practical knowledge with industrial state-of-the-art architectural solutions and designs.

Aziz Nasridinov is currently an Associate Professor of computer science with Chungbuk National University. He received the B.Sc. degree from the Tashkent University of Information Technologies in 2006 and the M.Sc. and Ph.D. degrees from Dongguk University in 2009 and 2012, respectively. Prof. Aziz Nasridinov has published over 30 papers in various high-ranked international journals and conferences. He has also served as a program committee member and co-organizer for numerous top-tier conferences, including ACM SAC, IEEE Big Data, IEEE Globecom and AAAI, and also serves in the editorial board of several international journals. His research interests include traditional databases, big data analytics with machine learning, and computer vision.