International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 4
ISSN: 2088-5334

Reference Class-Based Improvement of Object Detection Accuracy

Raegeun Park[a], Jaechoon Jo[b]

[a] *Department of Smart Information and Telecommunication Engineering, Sangmyung University, Cheonan, 31066, Republic of Korea*
E-mail: sogtvtk@gmail.com

[b] *Division of Computer Engineering, Hanshin University, Osan,18101, Korea, Republic of Korea*
E-mail: jaechoon@hs.ac.kr

***Abstract*— To date, the Frames Per Second (FPS) and accuracy of object detection based on deep learning have made rapid progress. However, the accuracy is limited by issues such as false positive (FP) cases. FP cases can trigger malfunctions in applications requiring high accuracy, such as in autonomous vehicles, where it is essential to ensure driver safety when malfunctions occur. To reduce the occurrences of FP cases, we conducted an experiment to derive the association by separately detecting a highly relevant element called a reference class, in addition to the target class to be detected. To measure the association, we obtained the integrated association by first finding the associations between the bounding boxes of the target and reference classes. Then we generated a reference class-based model by applying the integrated association to a trained model. The reference class-based model achieved approximately 15% higher accuracy than the trained model at iteration 1,000. Besides, the proposed model reduced the FP cases to approximately half of the 18.964% in the conventional method; the FP reduction through an increase in iteration was only 11.008%. The reference class can be applied in various fields, such as security and autonomous vehicle technology. It can be used to reduce the FP cases and improve the accuracy performance limits in object detection. Furthermore, it is possible to reduce the cost of reinforcing the training dataset and using high-performance hardware, and the time cost of increasing training numbers.**

***Keywords*— reference class; target class; FP case; association; improvement of accuracy performance.**

I. INTRODUCTION

The advent of deep learning has enhanced the performance of artificial intelligence in many fields. In the Contest to classify multiple images belonging to a particular category (ISVRC, ImageNet Large Scale Visual Recognition Challenge), the classification error rate remained at 20% for many years. After the introduction of Alexnet [1] in 2012, that led to a sudden drop in the error rate to a 10% range. Although previous deep learning did not work well owing to hardware limitations, object detection has progressed rapidly once object classification became sufficiently accurate. Algorithms such as SIFT [2] and SURF [2] based on conventional feature points face limitations in real-time processing because of their processing speed.

Research on improving the performance of feature-based algorithms with poor real-time performance is continuously being conducted. SSD, RCNN, Fast-RCNN, Faster-RCNN, You Only Look Once (YOLO) [3]–[4], and other deep learning models have been developed and applied to many applications. YOLO has a high frame rate (i.e., frame rate per second, FPS) and reasonable accuracy.

TABLE I
PERFORMANCE TABLE USING YOLOV3

Model	FPS
YOLOv3-320	45
YOLOv3-416	35
YOLOv3-608	20
YOLOv3-tiny	220
YOLOv3-spp	20

Table I [5] shows the performance results of YOLOv3 based on the COCO Dataset [6] that has been used for performance measurement in many studies. YOLOv3 achieves high speeds of 20 FPS (YOLOv3-spp), 35 FPS (YOLOv3-416), and 220 FPS (YOLOv3-tiny). Although the performance varies with the cameras available in the market in actual usage, the excellent real-time performance of current object detection technology can be inferred from the fact that most of the cameras commonly available in the market have frame rates of 30 FPS. With the improvement of object detection FPS performance, movements can be processed at high speeds in devices, such as autonomous vehicles and smartphones. However, to date, full accuracy has not yet been achieved. This has caused related problems.

Typical methods of increasing accuracy include using models that can perform well in the intended application and reinforcing the training dataset. Although these methods can improve accuracy, they also have drawbacks, such as they are costly and can only achieve limited performance improvement.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

Equation 1 is a formula for calculating the accuracy. Among the various performance evaluation elements considered in calculating the accuracy, a crucial element is the false positive (FP) case. For example, an autonomous vehicle traveling at high speed on a highway that detects a person in front at a short distance away is likely to execute sudden braking. Sudden braking at high speeds can cause damage to the vehicle and also endanger the driver. Thus, the FP case is more than just an element to consider for improving the accuracy, but it is also closely related to malfunctions when the system is applied to an actual product. Furthermore, it is highly likely that malfunctions caused by the FP cases will entail various safety issues.

Object detection has made large progress in detection speed and accuracy. Despite the many studies on improving object detection performance [7]–[27], uni-class-based object detection has seen only limited improvement. FP cases can occur in uni-class-based object detection when there are many objects similar to the class to be detected. For example, in attempts to extract only a human face, there may be regions similar to the face, and FP occurs in the non-face region. Because typical methods for preventing FP cases in the detection process involve training using the uni-class, the training dataset can be reinforced or the model changed to one that is suitable for the use case.

Reinforcing the training dataset has drawbacks such that substantial time and monetary costs are incurred in acquiring and pre-processing suitable data, and the range of possible performance improvement is limited. As an alternative, we can employ a model that is suitable for the required accuracy. For example, if the application does not require high speed, a highly accurate but slow model can be used. However, this approach is not applicable in the scenario of a self-driving car travelling at high speed in which the maximum performance limitations of the model cannot be overcome.

In this study, we propose a reference class to reduce the object detection FP cases and increase the accuracy.

II. Material and Method

A. Problems with the FP Case

We aim to reduce the FP cases to improve the accuracy performance limit of the conventional model. FP case can go beyond simply improving the accuracy performance figures, and there is a possibility of causing safety issues for various projects. When obstacles are detected in front of autonomous vehicles traveling at high speeds, control of sudden braking and avoidance will be inevitable. In these situations, the driver's safety cannot be guaranteed. Currently, these errors are corrected by comprehensively using sensors such as radio detection and ranging (radar) and light detection and ranging (LiDAR). However, if the road surface is wet because of weather conditions, diffused reflection can occur, which leads to malfunctions of the LiDAR sensors. Because there are constant reports of malfunctions in distance-based sensors, such as LiDAR, there is a need for vision sensor-based countermeasures as backups when malfunctions occur [28]. Road environment objects such as traffic signs and traffic lights have been designed to meet driver requirements. Thus, research based on vision sensors is necessary.

B. Reference Class: Solution for Improving Performance

To solve the fundamental problem of FP cases in object detection, multi-class-based object detection is performed instead of uni-class-based object detection to detect the target class. The target class refers to the object class to be detected. One of the reasons why the FP case occurs when an object is misclassified as the target class is because of its similarity to the target class, coupled with insufficient training data or limited model performance. As a result, the FP case has a low detection confidence score [3] during the object detection.

This study proposes a reference class as a solution to this problem. Humans do not consider only a single element when determining the class of a particular object. For example, assume that there is a picture of a dog with fur similar to the bread we commonly ate. We do not consider only the object itself to determine its class but also comprehensively consider its background and position and other objects in its vicinity. We focus on the elements that can be mistaken by us to identify the object during this process. In view of these human characteristics, we consider objects above the target class in a sense explained in the next paragraph. If there are many features similar to the human target class, the problem of incorrect detection cannot be resolved even if the training dataset is reinforced. Accordingly, the FP case can be removed when an upper object in the target class is detected together with the object.

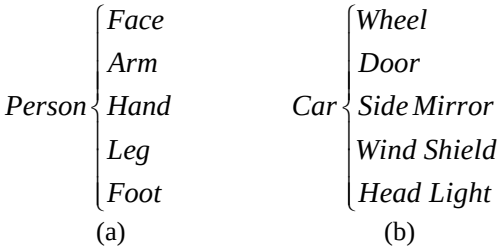


Fig. 1 Association between reference and target classes

Fig. 1 shows an example of the target and reference classes. As shown in Fig. 1(a), an individual's face cannot exist in general independently, that is, Person is the upper element to Face. Assuming that the target class is Face, Person has a high association with Face and is the upper element of Face. In this study, the reference classes are defined as the upper elements. In the same context, assuming that the target class is Wheel in Fig. 1(b), the reference class is Car.

In addition to the compositional relationship between upper and lower elements in Fig. 1, objects that have a high association with the target class can also be used as the reference class. For example, a road mark, in general, has a road lane around it. Assuming that the road mark is the

target class, the road mark itself does not have a direct compositional relationship to the road lane but has a high association. Thus, the road lane can be used as a reference class owing to its high association to the road mark.

C. Performance Improvement Process

To measure the performance of the reference class, it is assumed that an object detection system is created to detect only the front face of a standing person. Hence, the processes illustrated in Fig. 2 are performed. The target class is the front face, and the reference class is the entire human body. The implementation details are as follows.

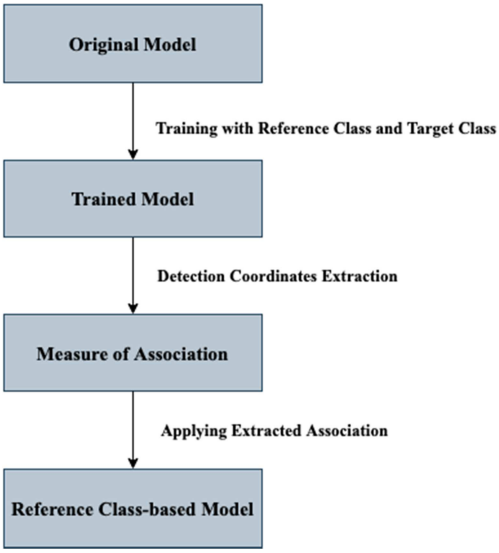
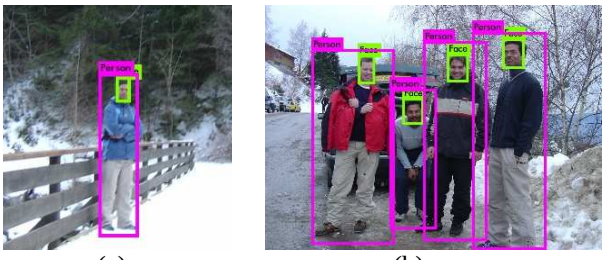


Fig. 2 Performance improvement process

D. Training Model with Reference Class

As shown in Fig. 3, multi-class-based object detection is implemented instead of uni-class object detection to extract only the face looking at the front. YOLOv2 is used for the object detection model. The target class is Face and the reference class is Person, and the trained model is created by training with the two classes.



(a) (b)

Fig. 3 Trained model

FFHQ-Dataset (Flickr-Faces-HQ) [29] and INRIA Person Dataset [30], [31] were used as the training datasets for training the object detection model. In particular, the INRIA Person dataset comprises images captured in various environments. Training with the dataset allows the object detection model to respond to various environments rather than limited specific environments, such as indoor environments.

E. Detection Coordinates Extraction and Application

The face is one of the components of the human body. Regardless of how much a person moves, the range of positions that the face and body can be relative to each other is limited. By looking at the results of Fig. 3(a) and 3(b), we can easily identify the association between Person and Face through their respective bounding boxes. However, there is the problem, as best described in Moravec's Paradox [32], that the system may simply be unable to identify the association. To compensate for this, the coordinates of the bounding boxes generated by detecting the Face and Person classes are extracted to measure the association using the constraint that the posture ranges that the face and body can express are limited.

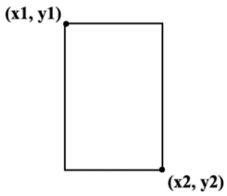


Fig. 4 Extracted coordinates of bounding box

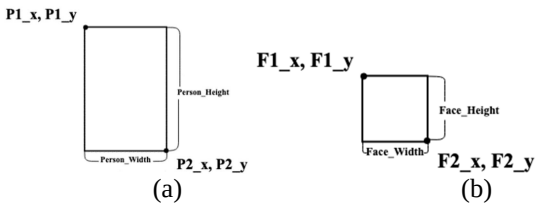
To extract the minimum number of coordinates of the detected boundary boxes in Fig. 3 using the trained model, two coordinates are extracted for each boundary box. As shown in Fig. 4, the extracted coordinates are the upper left and lower right points of the bounding box. These points are used to measure the association using the width, height, area, and coordinates of the bounding box. By applying the association extracted from the bounding boxes to the trained model, a class-based reference model is created. The reference class-based model can avoid the FP cases that occurred previously and improve the accuracy of object detection.

III. Results and Discussion

In this section, we calculate the association between the reference and target classes by extracting the coordinates from the trained model through the processes described in Fig. 2. We then present the improvement in the accuracy of the reference class-based model that results from applying the extracted association to the trained model.

A. Experimental Environment

Each person has unique characteristics, such as face size, height, and shoulder width. Hence, it is impossible to derive an equation that provides an individual's exact height simply by using his face size. Accordingly, overfitting to a specific group is highly likely to occur if only data from people of certain age groups or ideal body shapes, such as models, are used in an experiment.



(a) (b)

Fig. 5 (a) Coordinates of Person class and (b) coordinates of Face class

To avoid this, we conducted an experiment based on varied data comprising infants, children, adolescents, adults of both genders, and professional models. For the convenience of explanation, the coordinates of the Person and Face classes are represented as shown in Fig. 5.

B. Ratios-based Associations

The width, height, and area of each detected person and face were calculated using the coordinates extracted in Fig. 5(a) and 5(b). The results of the calculations to find the associations from the experimental images are shown in graphs (Fig. 6).

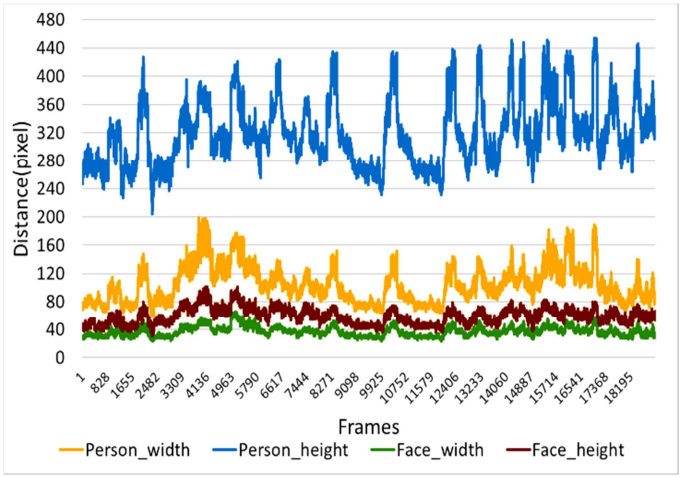


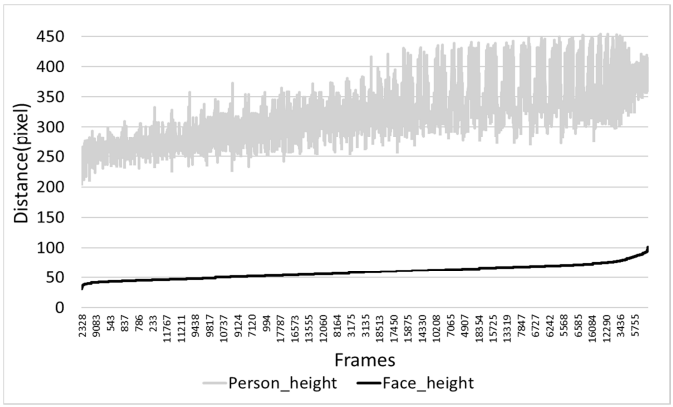
Fig. 6 Distance calculation results according to frames

Fig. 6 shows a graph of the results extracted from images of people with various body shapes. The numerical values of the Person_width, Person_height, Face_width, and Face_height extracted from the image frames were calculated in terms of image pixels. Because the pixel distances differ with the camera specifications and images, the association between the target class Face and the reference class Person was extracted using the ratios of the values in this graph.

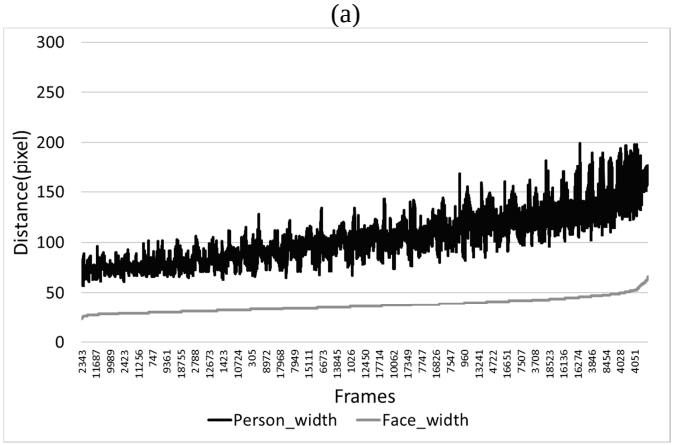
1) Limitations of Linear Equation Derivation

Fig. 7 shows the calculated values of the height, width, and area of the target and reference classes. Each quantity is sorted based on the values for Face, which is the target class. Because each graph is sorted by the target class, the data points do not appear in frame order. To derive linear equations relating the target and reference classes, each element should have a linear relationship between the target and reference classes, but the results in the graphs do not meet this requirement. Each of the respective graphs for the height, width, and area shows the same trend. Although each element of the reference class has been sorted in anticipation of a linear increase in the target class element, the actual experimental results are more similar to noise than to linear increases. Therefore, it is difficult to tell that Face and Person have linear associations to each other from these graphs.

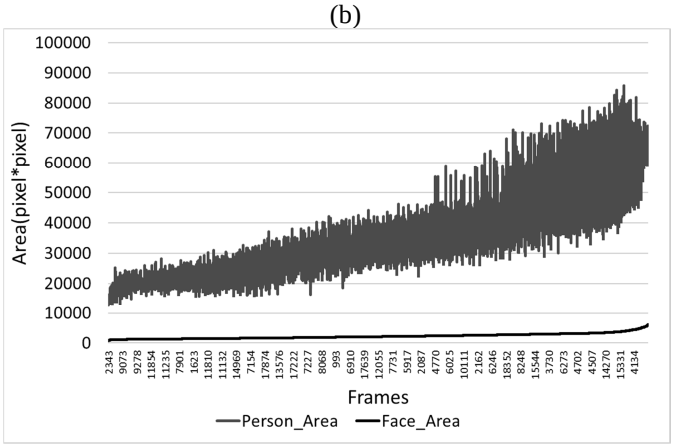
The reason for this is because humans are not homogeneous objects, and each individual has a different face, shoulder width, and height. Thus, as shown in Fig. 7, there is a limit to the feasibility of deriving linear equations relating Face and Person. Similarly, although it might be possible to derive fitting equations to these graphs, there is a limit to the practical application because the error rates for the fitting equations will be high. Therefore, we extract the association by defining section limits rather than linear equations.



(a)



(b)



(c)

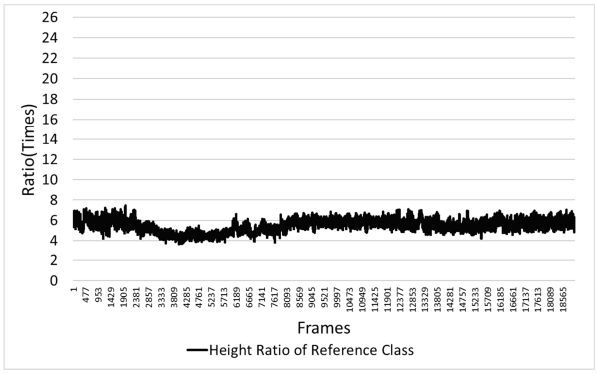
Fig. 7 (a) Association graph of human height and face height in different frames sorted by target class. (b) Association graph of human width and face width in different frames sorted by target class. (c) Association graph of human area and face area in different frames sorted by target class

2) Measurement Ratio Graphs

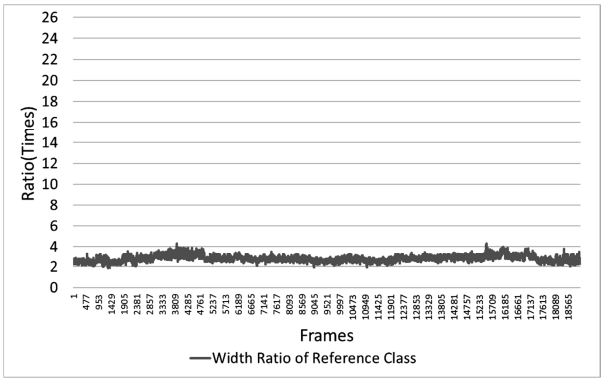
$$Height\ Ratio = \frac{Reference\ Class\ Height}{Target\ Class\ Height} = \frac{|P2\_y - P1\_y|}{|F2\_y - F1\_y|} \quad (2)$$

$$Width\ Ratio = \frac{Reference\ Class\ Width}{Target\ Class\ Width} = \frac{|P2\_x - P1\_x|}{|F2\_x - F1\_x|} \quad (3)$$

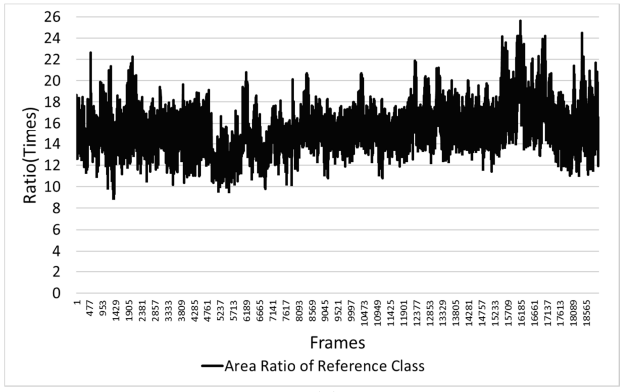
$$Area\ Ratio = \frac{Reference\ Class\ Area}{Target\ Class\ Area} = \frac{|P2\_x - P1\_x| \times |P2\_y - P1\_y|}{|F2\_x - F1\_x| \times |F2\_y - F1\_y|} \quad (4)$$



(a)



(b)



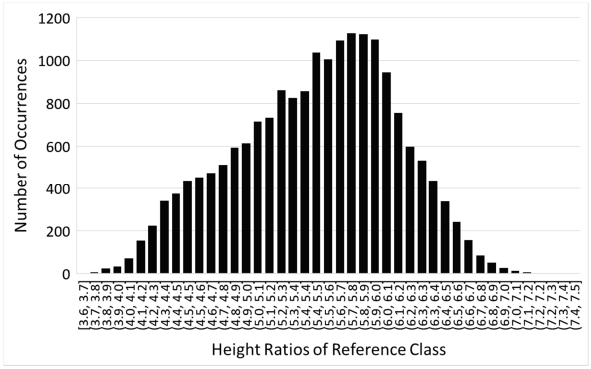
(c)

Fig. 8 (a) Height ratio graph according to frame. (b) Width ratio graph according to frame. (c) Area ratio graph according to frame

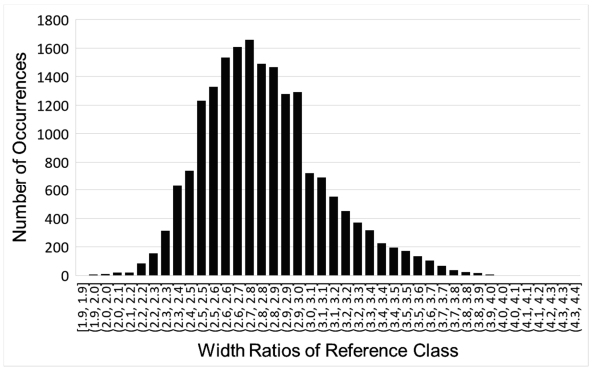
Figs. 8(a), 8(b), and 8(c) are graphs showing the ratios of the target and reference classes calculated through Equations 2, 3, and 4, respectively. As previously described, humans are not objects with uniform features. The ratios of the heights, widths, and areas calculated for each frame are based on various categories of people such as infants, adolescents, adults, and professional models, and not just a specific category. It can be confirmed that these ratios lie within certain limited ranges. Although each person has different physical dimensions, the postures and physical dimensions expressed by the human body are limited to specific ranges that can be used for formulating the association.

Among the three plots in Fig. 8, Fig. 8(b) shows the smallest range of variation because the width of the human shoulder, which is one of the elements affecting a person's width, has a relatively narrow variation range. Although the physical shoulder width of each person may differ, when clothes are worn, the degree of variation of the bounding box extracted from the actual object detection is small. It has a narrower ratio range than the other elements. The height element has a wider variation range than the width. The area is calculated as the product of the width and the height and is influenced by the height. As a result, the ratio ranges of the height and the area, which is affected by the height element, have relatively large variations. Thus, the width ratio is considered to be an element that shows a high association with humans.

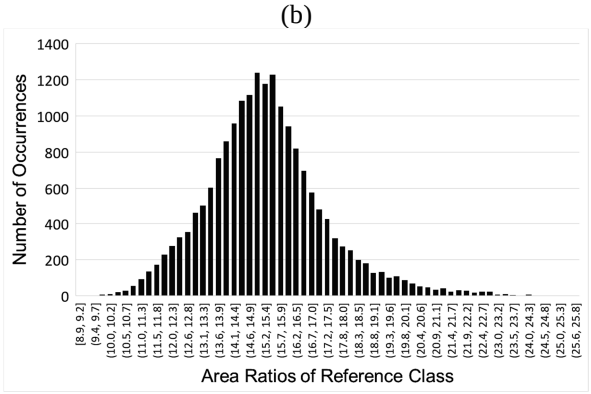
3) Distribution Graphs



(a)



(b)



(c)

Fig. 9 (a) Distribution chart of height ratio. (b) Distribution chart of width ratio. (c) Distribution chart of area ratio

Fig. 9 shows the graphs obtained by converting the results in Fig. 8 into distribution charts, which have Gaussian forms. We can observe that the distribution of the width ratio is narrower than the distributions of the other elements. Distributions less than 0.01% in each distribution chart were negligible as noise and not considered in the section limit:

$$Height\ Association(HA) = \begin{cases} 1, 3.808 \le Height\ Ratio \le 7.291 \\ 0, Otherwise \end{cases} \quad (5)$$

$$Width\ Association(WA) = \begin{cases} 1, 1.861 \le Width\ Ratio \le 4.301 \\ 0, Otherwise \end{cases} \quad (6)$$

$$Area\ Association(AA) = \begin{cases} 1, 9.18 \le Area\ Ratio \le 25.04 \\ 0, Otherwise \end{cases} \quad (7)$$

Through this process, we derived the association between the target and reference classes, as shown in Equations 5, 6, and 7. The area ratio is limited to a specific range but has the drawback of having a wide range distribution.

C. Coordinate Associations

To further investigate the associations using the ratios above, the lengths and sizes of the body and face bounding boxes were analyzed. Through this analysis, we found an additional association.

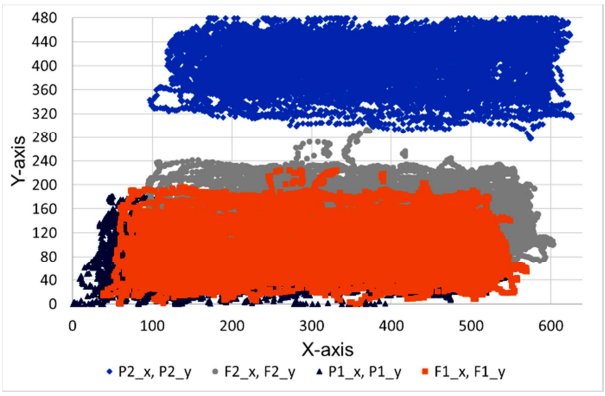


Fig. 10 Coordinate distribution graph of the bounding boxes

Fig. 10 shows the distribution graph of the coordinates of all the bounding boxes detected in the experiment. Notably, among the four coordinates, (P2_x, P2_y) is distributed far away from the other three coordinates. This is because the length of the entire human body is longer than the face, and the face is located at the upper part of the body. The distribution graph indicates that the association of the (P2_x, P2_y) coordinates is lower than that of other coordinates. Using this feature, we obtain the association between the coordinates of the target and reference classes.

As shown in Fig. 10, the distributions of the (P1_x, P1_y) and (F1_x, F1_y) coordinates are quite similar compared to the other coordinates, indicating a high association between them. Accordingly, we extract the association based on these two coordinates. The distance measured from the coordinates on each image was scaled by the total image resolution to correct for the difference in pixel counts owing to camera settings and image resolution:

$$\text{Coordinate Ratio} = \frac{\left|\text{Coordinate of Reference Class - Coordinate of Target Class}\right|}{\text{Resolution Pixels}} \quad (8)$$

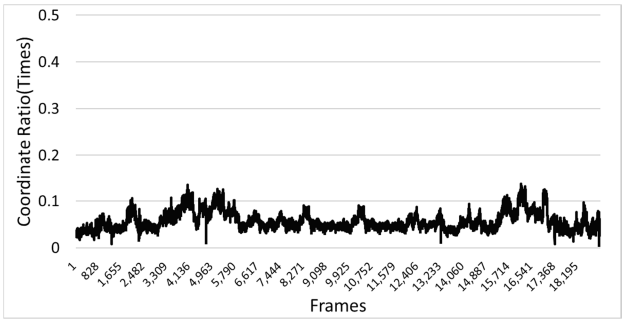


Fig. 11 Coordinate ratio based on the x-coordinate

Fig. 11 shows the coordinate ratio based on the x-coordinate calculated using Equation 8. Although the coordinate ratio has a limited distribution range, many postures can be expressed by the body moving from side to side, thus resulting in a relatively large range when the coordinate ratios are calculated based on the x-coordinate. To address this issue, we calculated the coordinate ratio based on the y-coordinate.

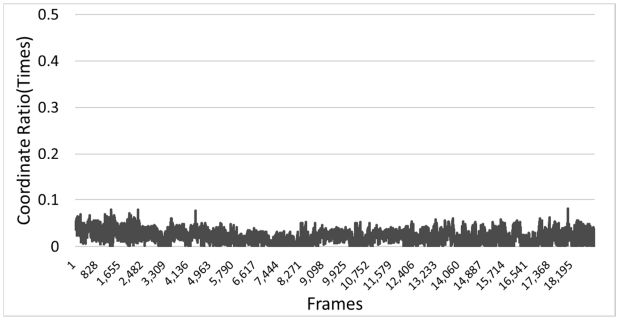


Fig. 12 Coordinate ratio based on the y-coordinate

Fig. 12 shows the coordinate ratios calculated based on the y-coordinate to overcome the problem of the x-coordinate ratios with a large range. A higher association is shown when we comprehensively consider both the distribution graph of Fig. 10 and the graph of Fig. 12.

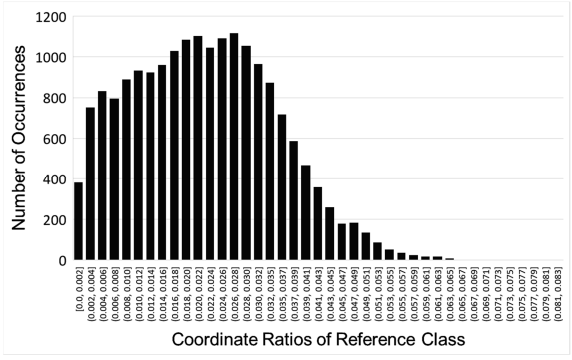


Fig. 13 Distribution chart of the coordinate ratio

Fig. 13 shows the distribution chart of the coordinate ratio based on the y-coordinate in Fig. 12, which has a Gaussian form. Coordinate ratios with distributions of 0.01% or less were regarded as noise and not included in the section limits.

As a result, the association based on the coordinate ratio was defined as follows:

$$Coordinate\ Association(CA) = \begin{cases} 1, 0 \le Coordinate\ Ratio \le 0.065 \\ 0, Otherwise \end{cases} \quad (9)$$

Through this process, we obtained the coordinate ratio association between the target and reference classes shown in Equation 9. As shown in Figs. 10 and 13, the coordinate ratio has a very high association compared to other associations. It thus highly contributes to accuracy improvement.

D. Detection Results Using Reference Class

To find the association between the target class and the reference class, we extracted the height, width, area, and coordinate data. It was impossible to derive linear equations between the different datasets owing to the diversity of human body shapes. Therefore, associations defined through limited ranges of the elements were used in place of linear equations. Combining the extracted associations, the integrated association is obtained:

$$Integrated\ Association(IA) = \begin{cases} Detection, HA \times WA \times AA \times CA = 1 \\ None, Otherwise \end{cases} \quad (10)$$

Extracting the integrated association between the target and reference classes in Equation 10 and applying it to the trained model can reduce the FP cases, which are one of the causes of degraded accuracy. To evaluate the application of the integrated association to the trained model, we used part of the COCO dataset containing various objects, such as people, animals, and food. By using the COCO dataset consisting of various objects as shown in Fig. 14 as the evaluation dataset rather than a dataset biased towards specific objects, we were able to increase the reliability of the measurement accuracy in varied environments and avoid overfitting to a specific environment.

Fig. 14 Sample of the COCO dataset

(a)

(b)

Fig. 15 (a) Before: trained model (b) After: reference class-based model

Fig. 15 [33] shows the results of detecting the target class, i.e., Face. In Fig. 15(a), the results from the trained model show a FP case in which a bear doll at the bottom with similar features to the human face was misidentified as a human face. Fig. 15(b) shows the results from the reference class-based model for which the integrated association in Equation 10 was applied to the trained model to address this issue. The association analysis for the misidentified bear doll at the bottom showed that the doll does not have any association with the reference class, thus eliminating the FP case. In addition to Fig. 15, we were able to reduce the number of FP cases caused by features similar to the target class in other images, some of which are shown in Fig. 16, and hence improve the accuracy.

The derivation of a linear equation is not guaranteed in the association measurement between the target and reference classes. As shown in this work on the association between the face and the entire human body, there is the problem that a linear association cannot be derived because of the presence of numerous variables unless the objects are uniform. This problem can be solved by extracting range limits for the associations of the suggested elements.

We can prevent fraudulent face recognition in smartphones using the owners' photographs by using reference classes to improve accuracy by reducing the FP cases. Assuming the target and reference classes to be the human face and the entire body respectively, as in the current study, comprehensive association with the body can be applied when facial recognition data are input. This can counter illegal input in facial recognition using only face data in applications sensitive to malfunctions caused by FP cases. In addition, using the posts of traffic signs and traffic lights as the reference class for recognizing traffic signs and traffic lights can improve the accuracy in a wide range of areas.

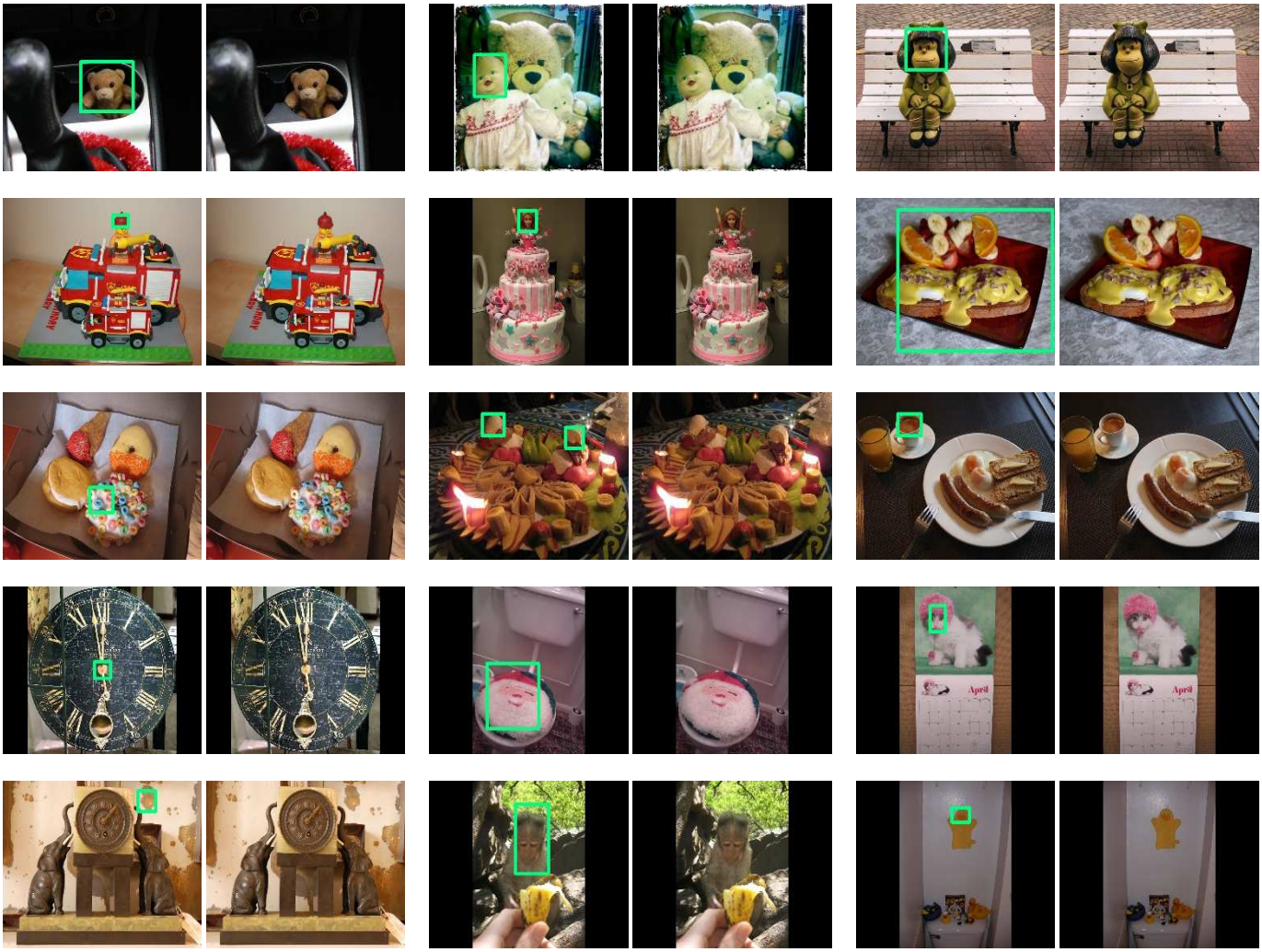


Fig. 16 Left: trained model, right: reference class-based model

E. Accuracy Performance Evaluation Using Reference Class

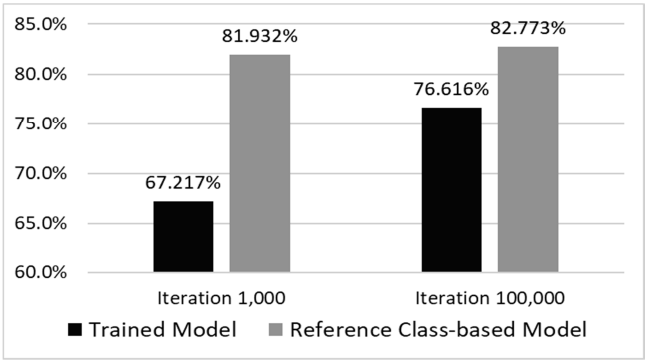


Fig. 17 Accuracy performance comparison of trained and reference class-based models

Fig. 17 shows a graph comparing the accuracy performance of the reference class-based model to which the extracted integrated association was applied and a model trained in a generic way.

The face detection accuracy of the trained model was approximately 67.217% at iteration 1,000 and approximately 76.616% at iteration 100,000, which constitutes 100 times more training. Even after a 100-fold investment increase in training, the performance improvement was limited with an accuracy increase of only approximately 9.399%. To solve the problem of small performance improvement, an integrated association based on the reference class was applied to the trained model at iteration 1,000. The accuracy was approximately 81.932%, an improvement of 14.715%. The accuracy surpassed that of the trained model at iteration 100,000. When the integrated association was applied to the trained model at iteration 100,000, the accuracy was 82.773%, which is an improvement of 6.157% compared to the trained model.

Despite the difference of more than 90,000 training iterations, the trained model at iteration 100,000 has a lower accuracy than the reference class-based model. The results from applying the integrated association based on the reference class show that it is possible to reduce the cost and time incurred in using high-performance hardware for a long time and that the accuracy performance limit of the conventional method can be improved.

Fig. 18 shows a graph of the FP case percentage, i.e., the ratio of the FP cases to the total number of cases, calculated using Equation 11. The FP case percentage was 29.972% at iteration 1,000 and 18.964% at iteration 100,000, showing that the number of FP cases decreases with increased training. However, despite the 100-fold increase in training, the FP cases were reduced by only 11.008%.

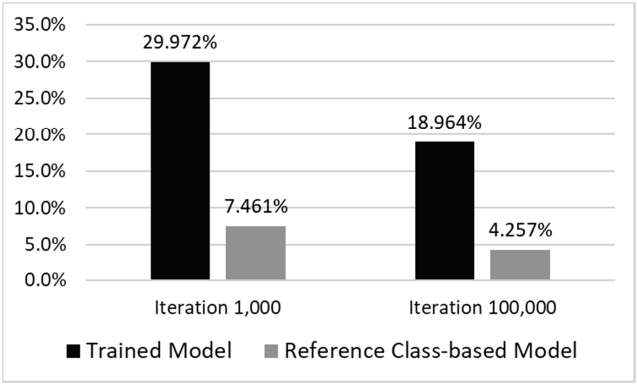


Fig. 18 FP case comparison of trained and reference class-based models

$$FP\,case\,Percentage = \frac{FP}{TP + FN + FP + TN} \qquad (11)$$

The FP case percentage decreased by 22.511% at iteration 1,000 and by 14.707% at iteration 100,000 when the integrated association based on the reference class was applied. Although there were 90,000 fewer training iterations, the reduction of the FP percentage in the reference-class-based model at iteration 1,000 was more than twice that of the trained model at iteration 100,000. The application of reference class-based integrated association for the reduction of FP cases, which are a cause of degraded accuracy, shows that FP cases can be reduced at low monetary and time costs.

IV. Conclusions

This study aimed to improve the accuracy of object detection. To date, object detection based on deep learning has made rapid progress in terms of accuracy and FPS, but there is still a limit to accuracy improvement. To improve the accuracy, we proposed the reference class as a method of reducing FP cases beyond the conventional methods of augmenting the data of the training dataset or using the appropriate model for the project. FP cases are one of the fundamental causes of lowered accuracy. In this study, using the reference class, we extracted the association between the target class and the reference class for multi-class, rather than uni-class, object detection.

We performed the experimental process shown in Fig. 2. We extracted the association using the height, width, area, and coordinates of the bounding boxes of each detected class. Since we were unable to derive linear equations relating these elements to one another because of the varying characteristics of the target class in this experiment, we extracted the associations based on range limits obtained by analyzing the association distribution of each element. We obtained the integrated association by combining the associations of the height, width, area, and coordinates extracted from the reference class and were able to reduce the FP cases by applying the integrated association to the training model. The integrated association can be further applied to applications sensitive to malfunctions caused by FP cases.

The reference class-based model generated by applying the reference class-based integrated association to the trained model increased the accuracy by approximately 15% at iteration 1,000 compared to the trained model. This result surpasses the accuracy of the latter at iteration 100,000 despite the 100 times difference in the training time. Besides, the occurrence rate of the FP case at iteration 1,000 was 7.461%, which is a reduction of approximately 23% compared to the conventional trained model. The proposed model reduced the FP cases to less than half of the 18.964% FP rate in the conventional method at iteration 100,000. The latter is a reduction of only 11.008% compared to the FP rate at iteration 1,000 despite the increase in the iteration count to 100,000.

Using the reference class, the FP cases in object detection can be reduced and the accuracy performance limits improved. Furthermore, the cost of reinforcing the training dataset and using high-performance hardware and the time cost of increasing training numbers can be reduced.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government [grant number NRF – 2018 R1D1A1B07051369]

References

[1] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *In Advances in neural information processing systems*, pp. 1097-1105, 2012

[2] E. Karami, S. Prasad and M. Shehata, "Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images," *arXiv preprint arXiv:1710.02726*, Oct. 2017

[3] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *The IEEE conference on computer vision and pattern recognition*, pp. 7263-7271, 2017

[4] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv:1804.02767*, 2018

[5] YOLO. (2020), YOLO:Real-Time Object Detection. [Online]. Available: https://pjreddie.com/darknet/yolo/

[6] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, "Microsoft COCO: Common Objects in Context", *European conference on computer vision. Springer*, pp. 740-755, 2014

[7] Y. Long, Y. Gong, Z. Xiao and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing, 55(5)*, pp. 2486-2498, Jan. 2017

[8] X. Wang, A. Shrivastava and A. Gupta, "A-fast-rcnn: Hard positive generation via adversary for object detection," *the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2606-2615, 2017

[9] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," *the IEEE conference on computer vision and pattern recognition*, pp. 3578-3587, 2018

[10] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," *the IEEE conference on computer vision and pattern recognition*, pp. 6154-6162, 2018

[11] Y. Chen, W. Li, C. Sakaridis, D. Dai and L. V. Gool, "Domain adaptive faster r-cnn for object detection in the wild," *the IEEE conference on computer vision and pattern recognition*, pp. 3339-3348, 2018

[12] J. Jeong, H. Park and N. Kwak, "Enhancement of SSD by concatenating feature maps for object detection," *arXiv preprint arXiv:1705.09587*, May. 2017

[13] X. Sun, P. Wu and S. C. H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing, 299, 42-50*, 2018

[14] X. Zhu, Y. Wang, J. Dai, L. Yuan and Y. Wei, "Flow-guided feature aggregation for video object detection," *the IEEE International Conference on Computer Vision*, pp. 408-417, 2017

[15] T. Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," *the IEEE international conference on computer vision*, pp. 2980-2988, 2017

[16] T. Y. Lin and S. Maji, "Improved bilinear pooling with cnns," *arXiv preprint arXiv:1707.06772*, 2017

[17] L. Tychsen-Smith and L. Petersson, "Improving object localization with fitness nms and bounded iou loss," *the IEEE conference on computer vision and pattern recognition*, pp. 6877-6885, 2018

[18] X. Wang, X. Hua, F. Xiao, Y. Li, X. Hu and P. Sun, "Multi-object detection in traffic scenes based on improved SSD," *Electronics, 7(11), 302*, 2018

[19] G. Bertasius, L. Torresani and J. Shi, "Object detection in video with spatiotemporal sampling networks," *the European Conference on Computer Vision (ECCV)*, pp. 331-346, 2018

[20] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng and S. Yan, "Perceptual generative adversarial networks for small object detection," *the IEEE conference on computer vision and pattern recognition*, pp. 1222-1230, 2017

[21] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu and Y. Chen, "Ron: Reverse connection with objectness prior networks for object detection," *the IEEE conference on computer vision and pattern recognition*, pp. 5936-5944, 2017

[22] Y. Li, Y. Chen, N. Wang and Z. Zhang, "Scale-aware trident networks for object detection," *the IEEE International Conference on Computer Vision*, pp. 6054-6063, 2019

[23] P. Zhou, B. Ni, C. Geng, J. Hu and Y. Xu, "Scale-transferrable object detection," *the IEEE conference on computer vision and pattern recognition*, pp. 528-537, 2018

[24] Z. Zhang, S. Qiao, C. Xie, W. Shen, B. Wang and A. L. Yuille, "Single-shot object detection with enriched semantics," *the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5813-5821, 2018

[25] S. Zhang, L. Wen, X. Bian, Z. Lei and S. Z. Li, "Single-shot refinement neural network for object detection," *the IEEE conference on computer vision and pattern recognition*, pp. 4203-4212, 2018

[26] N. Bodla, B. Singh, R. Chellappa and L. S. Davis, "Soft-NMS--improving object detection with one line of code," *the IEEE international conference on computer vision*, pp. 5561-5569, 2017

[27] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang and W. Ouyang, "T-cnn: Tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology, 28(10)*, pp. 2896-2907, 2017

[28] R. G. Park, H. J. Yun, E. G. Han, S. W. Kang, J. H. Park, E. J. Lee, D. H. Jeon, K. Y. Jung, S. B. Cho and T. K. Kang, "A Study on Countermeasures for ADAS Malfunction Based on YOLO", *Korean Institute of Electrical Engineers(KIEE) Conference*, pp.188-190, Nov. 2019

[29] T. Karras, S. Laine and T. Aila, "A style-based generator architecture for generative adversarial networks," *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401-4410, 2019

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *The IEEE computer society conference on computer vision and pattern recognition*, pp. 886-893, Jun. 2005

[31] M. Everingham, A. Zisserman, C. K. I. Williams, L. V. Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dork´o, S. Duffner, J. Eichhorn, J. D. R. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus and B. Leibe, "The 2005 PASCAL Visual Object Classes Challenge", *In Machine Learning Challenges Workshop*, pp. 117-176, Apr. 2005

[32] V. S. Rotenberg, "Moravec's paradox: consideration in the context of two brain hemisphere functions," *Activitas Nervosa Superior, 55(3)*, pp. 108-111, 2013

[33] (2020) Joint Base Langley-Eustis website. [Online]. Available: https://www.jble.af.mil/News/Photos/igphoto/2000197216/