Computers, Materials & Continua
DOI: 10.32604/cmc.2022.027840
Article

Tech Science Press

Convergence of Stereo Vision-Based Multimodal YOLOs for Faster Detection of Potholes

Sungan Yoon and Jeongho Cho*

Department of Electrical Engineering, Soonchunhyang University, Asan, 31538, Korea
*Corresponding Author: Jeongho Cho. Email: jcho@sch.ac.kr
Received: 27 January 2022; Accepted: 29 April 2022

Abstract: Road potholes can cause serious social issues, such as unexpected damages to vehicles and traffic accidents. For efficient road management, technologies that quickly find potholes are required, and thus researches on such technologies have been conducted actively. The three-dimensional (3D) reconstruction method has relatively high accuracy and can be used in practice but it has limited application owing to its long data processing time and high sensor maintenance cost. The two-dimensional (2D) vision method has the advantage of inexpensive and easy application of sensor. Recently, although the 2D vision method using the convolutional neural network (CNN) has shown improved pothole detection performance and adaptability, large amount of data is required to sufficiently train the CNN. Therefore, we propose a method to improve the learning performance of CNN-based object detection model by artificially generating synthetic data similar to a pothole and enhancing the learning data. Additionally, to make the defective areas appear more contrasting, the transformed disparity map (TDM) was calculated using stereo-vision cameras, and the detection performance of the model was further improved through the late fusion with RGB (Red, Green, Blue) images. Consequently, through the convergence of multimodal You Only Look Once (YOLO) frameworks trained by RGB images and TDMs respectively, the detection performance was enhanced by 10.7% compared with that when using only RGB. Further, the superiority of the proposed method was confirmed by showing that the data processing speed was two times faster than the existing 3D reconstruction method.

Keywords: CNN; YOLO; disparity map; stereo vision; pothole

1 Introduction

Potholes are a type of road damage in which a bowl-shaped depression is formed in the surface of the paved road. In recent years, the number of potholes is increasing rapidly owing to the aging of roads, climate change and increased traffic [1]. A pothole can cause severe damages to a moving vehicle and unexpected major traffic accidents while the process of avoiding the pothole [2–4]. To effectively solve the problems caused by these potholes, it is crucial to quickly detect and repair the potholes. However,

currently, potholes are being detected and reported by public officials or professional technicians, which is quite inefficient for continuous road management [5,6]. Because this method relies on the judgment of individual inspectors, the detection results lack consistency. Therefore, the need for objective, reliable, and robust automatic pothole detection system is increasing [7].

Research on pothole detection is majorly divided into three types: researches using vibration sensor-based method, three-dimensional (3D) reconstruction and two-dimensional (2D) vision-based method. The vibration sensor-based method is a method for detecting a pattern of vibration that occurs when a vibration sensor is attached to a vehicle that passes through a pothole. There is a limitation that this method cannot explicitly infer the shape and volume of the pit from the data obtained from the vibration sensor and the accuracy of detection is very low, such as vibrations caused by joints of roads or manholes are incorrectly detected as potholes [6,8]. The 3D reconstruction method [9,10] uses a laser sensor to generate 3D data regarding the information of the shape of the road surface and the defect area, enabling accurate pothole detection. However, laser sensor has the disadvantage of limited use due to its high initial and maintenance costs [11]. Accordingly, a detection system using the "Kinect" sensor, a low-cost laser scanning device, has been proposed, but because it is not designed for outdoor use, it often does not work properly and measures incorrect values when exposed to direct sunlight [12,13]. As another complement, a 3D reconstruction method using a vision sensor rather than a laser sensor has been proposed [14,15]. It is a technique for 3D reconstruction of the road surface that estimates the depth of an object using the disparity information of two images employing a stereo vision camera instead of using an expensive laser scanning device. Wang [16] showed 3D reconstructed pairs of images modeled after detecting cracks through 2D images of two vision sensors. In the report by Zhang et al. [17], a pothole was detected by the difference between an actual value and an estimated one from the model that was fitted on the road surface after converted into 3D point cloud data (PCD) using the disparity information obtained through two cameras. Using a stereo vision sensor, a 3D reconstruction technique can be implemented at a low cost, but as the task of constructing the surface becomes the main focus, many calculations are required to construct the surface, making the sensor difficult to use in a real-time environment due to its low execution speed. There is a disadvantage that the quality of the camera can be greatly affected if the camera is misaligned by the vibration of the vehicle [18].

The 2D vision method uses a vision sensor such as an RGB camera, and this method can be broadly classified into two types. 1) The computer vision method preprocesses 2D images to separate and detect the damaged and the nondamaged areas on the road surface [19–21] and 2) the object detection method uses a neural network [22,23]. The advantages of the 2D vision method are that a pothole detection system can be built at a low cost using a vision sensor and a higher performance can be expected compared with a vibration-based sensor. The vision sensor can work in an outdoor environment [6] and has the advantage of being easy to mount on a vehicle. In the report by Koch et al. [11], a pothole was detected by comparing the texture of the separated defective and nondefective areas with the surrounding areas based on the threshold value of the histogram of the road surface. Azhar et al. [24] reported the features of potholes that were extracted using the Histograms of Oriented Gradients (HoGs) technique based on the shape characteristics and classified the potholes using the naive Bayes classifier. However, the computer vision method can operate in real time owing to lower computational complexity than the stereo vision method; however, its detection performance is not satisfactory [25,26]. Moreover, it depends on lighting and sunlight and the detection performance will be poor if the pothole surface is filled with water or foreign matters [11]. In practice, the shape and texture of a pothole are highly irregular and the geometric features assumed during the feature extraction step may lose their effect. Therefore, neural networks were used for improving pothole

detection [22,23]. Ukhwah et al. [27] detected a pothole using the You Only Look Once (YOLO) model, which is a convolutional neural network (CNN)-based object detection model, and estimated the surface area of the pothole by comparing the pixel value with the actual distance. Meanwhile, the method using the neural network can achieve detection and high accuracy that can cope with various situations as compared with the conventional computer vision method; however, high accuracy can be expected only when there is a large amount and good quality data, and the data labeling procedures can be very labor-intensive [28].

In the existing computer vision method, the geometrical characteristics of the object must be specified in advance to detect an object [15]. However, the characteristics that are crucial to the detection performance change depending on the angle, sunlight, rainwater, etc. By contrast, the method using a neural network can show higher detection performance compared with the former method when exposed to various environments through various data and learning progress. Therefore, in this paper, we propose a pothole detection strategy using stereo vision-based multimodal YOLOs to secure the cost-effectiveness and real-time utilization of the pothole detection sensors. To achieve a high detection performance of YOLO, training must be conducted using a large amount of high-quality data, but it is challenging to obtain such data. Therefore, synthetic pothole data similar to the actual pothole is artificially generated and added to the training data through a data reinforcement process. In addition, high level of detection performance is not expected because the existing pothole detection technologies using a neural network have employed only a single RGB sensor. Therefore, to compensate for the disadvantages of the RGB vision sensors that are sensitive to the external environment, the disparity map calculated using the stereo vision sensor is obtained and converted into a transformed disparity map (TDM), which clears the boundary of the pothole, and then learning is achieved using YOLO, which is separated from RGB. Thus, each pothole is detected using the RGB sensor as well as TDM and the detection performance is improved by complementing the detected results after they are converged.

Potholes formed by factors such as climate change, aging and pressure increase on the road lead to a problem that must be resolved immediately when found because there is a high risk of traffic accidents. For this purpose, the contributions of the newly proposed 2D vision-based method in this context are as follows:

i. A new pothole detection model that is more accurate and reliable than the vibration sensor-based method has been proposed. At the same time, by using only a stereo vision sensor, real-time operation is possible with lower cost and much less complex data processing for pothole detection compared to laser-based 3D reconstruction techniques.
ii. By proposing a method to utilize not only RGB images but also a more refined disparity map through a stereo vision sensor, additional detection of recessed potholes became possible. Through the convergence of the obtained results, the detection accuracy was significantly improved compared to the existing 2D vision-based method.

The rest of this paper is organized as follows: In Section 2, the proposed pothole detection model is described; in Section 3, experimental results are presented. Finally, in Section 4, conclusions are presented.

2 Methodology

RGB-based pothole detection is based on the geometric characteristics of the pothole. For example, a pothole has characteristics such as a darker interior, rougher surface and deeper topography

than the surrounding area; therefore, it is easy to find characteristics such as texture, color and shape of the pothole through RGB images, but they may be difficult to detect because RGB images are sensitive to changes in the external environment. By contrast, the stereo vision-based disparity image uses the disparity between the two images to obtain more stable characteristics compared with the RGB image as a quantity indicating the depth in the image. Thus, the proposed pothole detection method detects each pothole through an independent neural network using the disparity image and the RGB image obtained from the stereo vision sensor and derives the optimal result from the detected results. The block diagram of the proposed pothole detection system is shown in Fig. 1, and it is largely composed of three blocks: 1) calculating the disparity map transformed by TDM, 2) pothole detection using YOLO and 3) decision by non-maximum suppression (NMS).

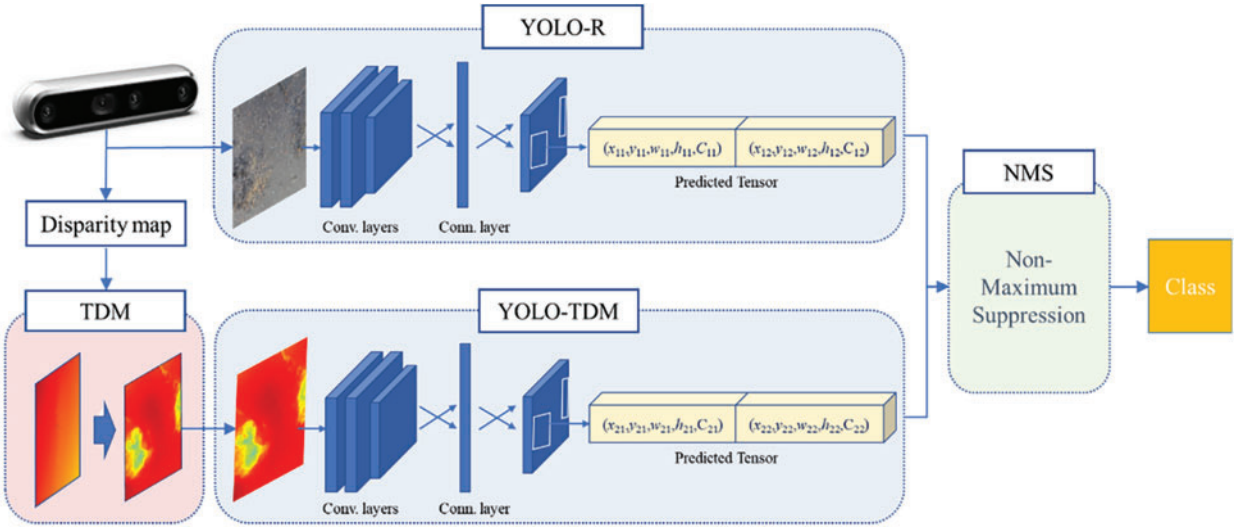

Figure 1: Block diagram of the proposed pothole detection system

2.1 TDM

Disparity map is an image showing the exact depth of an object through various corrections using the disparity of two images and a key element for pothole detection through stereo vision-based 3D reconstruction technique. Disparity map is advantageous for 3D reconstruction-based detection because it understands the road surface well; however, it is somewhat disadvantageous for the 2D vision-based detection because the boundary between the pothole and the surroundings is not clear. Therefore, in this study, the TDM is used to make a clear distinction of the pothole and the surrounding area so that the disparity map is more suitable for 2D vision-based detection [29,30].

Assuming that the road surface and the camera are perfectly horizontal, the disparity projection into the v-disparity region can be expressed as a linear straight line.

$$P(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \alpha_0 + \alpha_1 v \tag{1}$$

Here, $\boldsymbol{\alpha} = [\alpha_0, \alpha_1]^T$ is the disparity projection model coefficients vector, $\boldsymbol{\gamma} = [u, v]$ is the vertical–horizontal coordinate pixel in the disparity map, and when the disparity value of the disparity map is $\boldsymbol{d}$, the optimal $\boldsymbol{\alpha}$ is estimated as follows:

$$\boldsymbol{\alpha}_o = \underset{\alpha}{argmin} [\boldsymbol{d} - \boldsymbol{\gamma}\alpha]^T [\boldsymbol{d} - \boldsymbol{\gamma}\alpha] \tag{2}$$

However, in the real environment, the camera is not perfectly in level with the ground; therefore, the roll angle θ of the camera does not always become 0, causing distortion in the mismatch map. Therefore, if reverse rotation is performed using θ to turn the misaligned angle parallel to the horizontal axis, the original coordinate $\boldsymbol{\gamma} = [u, v]^T$ of the mismatch map is converted to a new coordinate $\boldsymbol{\gamma}' = [g,\ h]^{\mathrm{T}}$. Therefore, the disparity mapping into the v-disparity region is expressed as $P(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \theta) = \alpha_0 + \alpha_1 (vcos\theta + usin\theta)$, and TDM is defined as follows so that the defect area is further emphasized:

$$\mathrm{TDM} = \boldsymbol{d} - P(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \theta) + \delta \tag{3}$$

where δ is a set constant such that all transformed mismatch values are nonnegative. The conversion process of TDM is shown in Fig. 2. It can be seen that the defect area is more clearly visible than the conventional disparity map using TDM, and this is an important factor to increase the object detection performance.

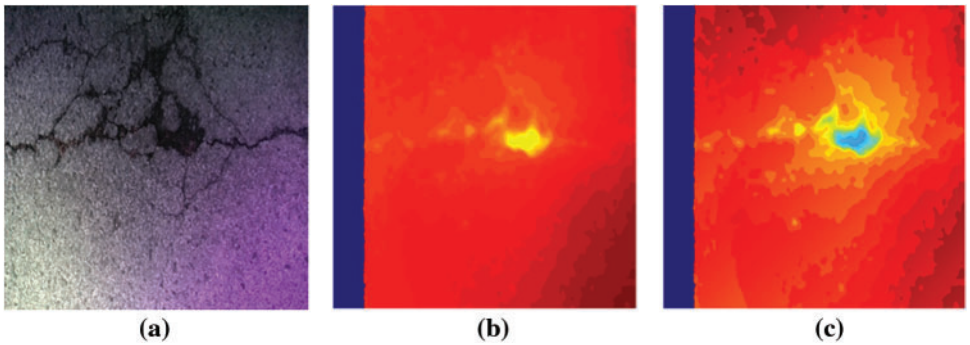

Figure 2: Example of a TDM conversion process: (a) raw image, (b) disparity map and (c) TDM

2.2 Object Detection Using YOLO

Image recognition using CNN has made it possible to recognize images with high accuracy through the development of various learning models and algorithms such as GoogleNet, Residual Net (ResNet), and Visual Geometry Group 16 (VGG16). As high-level recognition became possible, attention was naturally drawn to the object detection problem, which is an old problem in the field of computer vision and determines the location and type of a specific object in an image. The object detection problem has been difficult to access because it is more difficult than simple image classification problems, such as the problem of determining the location of an object and what the object is, and its structure is complicated. However, starting with region-based CNN (R-CNN) using a CNN-based image classifier, with the development of various object detection models such as Fast R-CNN and Faster R-CNN, the detection performance is gradually improving and active research is in progress [31]. These systems calculated the bounding box and class probability of an object separately and learned only the part that classifies the object through the neural network, which took a long time to learn, so they were insufficient for real-time application. The YOLO framework, an object detection system developed with more focus on real-time recognition, converts the bounding box and class probabilities within the image to single regression problem to speed up the detection and the type of object that sees the image once.

YOLO divides the input image into S × S grid regions and predicts B bounding boxes predetermined in the region where there is an object using the CNN. The bounding box of each zone represents

five pieces of information in (x, y, w, h, *and* C). (x, y) are the center coordinates and (w, h) are its width and height of the bounding box, respectively, and C is the probability that the bounding box is included in a specific class. C is expressed as follows as the product of the probability of including the object Pr(object) and IOU_{pred}^{truth} which is the area where the actual and predicted values overlap each other (IOU), which determines how accurately the bounding box is predicted:

$$C = \Pr(object) * IOU_{pred}^{truth} \tag{4}$$

if the actual value and the predicted center coordinates of the bounding box exist in the same region, the bounding box is considered to contain an object and Pr(object) is calculated as 1; otherwise it is calculated as 0. The probability of determining which object among the classified N objects is $\Pr(Class_i|Object)$, the total number of bounding boxes is S × S × B, and N CP_{class} for each bounding box is obtained as follows:

$$CP_{class} = \Pr(object) * IOU_{pred}^{truth} * \Pr(Class_i|object) = \Pr(Class) * IOU_{pred}^{truth} \tag{5}$$

Finally, the bounding box with the highest CP_{class} among the predicted bounding boxes B is selected as the bounding box of the object [32].

2.3 NMS

Recently, in the case of object detection in a complex environment such as autonomous driving, if only RGB images are used, the image may be distorted or damaged by external light sources, such as sunlight and lighting, and there is a disadvantage that they cannot be used at night. To compensate for this, a multisensor fusion method that overcomes the disadvantages of digital cameras by additionally using various sensors, such as Light Detection And Ranging (LiDAR) and Radar, has been proposed [33]. The early fusion method is characterized by the fusion of preprocessed sensor data to fully utilize the information of the raw data. However, it is sensitive to spatiotemporal data misalignment between sensors, such as calibration errors, different sampling rates and sensor defects. The late fusion method combines the output of the network and has high flexibility and modularity but involves high computational cost. The intermediate fusion method is a compromise between the early and late fusion methods, and it is possible to learn various features through the expression of various features of the network; however, it is difficult to find the optimal fusion method to accomplish the task of changing the network structure.

Therefore, in this study, we propose the fusion of an RGB camera and a stereo vision sensor to extract the depth feature of an object that cannot be obtained using an RGB camera alone and a late fusion method to minimize the interference between the sensors and derive an optimal object bounding box using NMS. NMS is mainly combined in the latter part of the detection model to improve object detection performance in models such as YOLO and Sing Shot Multibox Detector (SSD) and used for extracting the optimal bounding box.

First, for one class, we select the bounding box with the highest class probability and add it to the final. bounding box list. Second, the selected bounding box is compared with all other bounding boxes and IOUs, and if it is larger than the pre-defined threshold, the corresponding box is removed. Third, among the remaining bounding boxes, the bounding box with the highest class probability is selected

and added to the final bounding box list. Fourth, after comparing the IOU of the reselected bounding box with the remaining bounding box, if the IOU is larger than the threshold, the corresponding box is removed. Finally, steps 1 to 4 are repeated until there no bounding box remains.

In general, NMS in multi-object detection algorithm plays an important role in accurately discriminating the overlapping objects when the overlapping of objects occurs, which greatly affects the performance of the model. However, when the object to be detected is flat, such as during pothole detection, the possibility of overlap is less; therefore, an appropriate IOU threshold setting is required for NMS.

3 Experimental Results

The proposed pothole detection system was installed on an Intel RealSense D455 camera, NVIDIA GTX 1080ti, and Intel Core i7-8700 CPU. The detection model using only RGB images was defined as YOLO-R, the detection model using only the transformed inconsistency map was defined as YOLO-TDM and the model using the converged proposed RGB image and the transformed inconsistency map was defined as YOLO-R/TDM. We evaluated the level of improvement in detection performance via the fusion of data augmentation and multimodal detection results to improve the YOLO learning performance. The performance evaluation of the proposed model was conducted based on mean average precision (mAP), and for its comparison with the research results obtained using the existing state-of-the-arts method, the results were extracted from the report by Fan et al. [34] and used for performance comparison. mAP expresses the area under the curve of the Precision–Recall (PR) curve as a single value and indicates how confident the model is about the detected object. Precision is the ratio that is correctly detected among the results detected by TP/(TP + FP), and recall is the ratio that is correctly detected among the objects that should be detected with TP/(TP + FN), where TP is a true positive, that is, correct detection, FP is a false positive, that is, false detection, and FN is a false negative.

It is not easy for the CNN model to collect enough data to learn depending on the type of data; therefore, if there are less data, the model is trained using data reinforcement techniques such as rotating some secured data or adding noise [35]. The training data of Pothole 600 [36] used in this study was insufficient to train the YOLO network, so synthetic data (Figs. 3a and 3b) were artificially created and added to the training data for learning. The artificially created synthetic data were used as high-quality data for YOLO learning as they expressed the shape of the pothole more accurately and clearly. In general, it is known that the detection performance increases when training YOLO by adding approximately half of the existing training data [35]. Tab. 1 presents the model's performance before and after adding synthetic data by changing the IOU threshold. It can be seen that the overall detection performance improved after adding the data. Among them, when the IOU threshold was 0.6, the difference was at the most improved by 6%. Therefore, it was confirmed that the addition of synthetic data greatly helped in improving the detection performance and accuracy of CNN-based models.

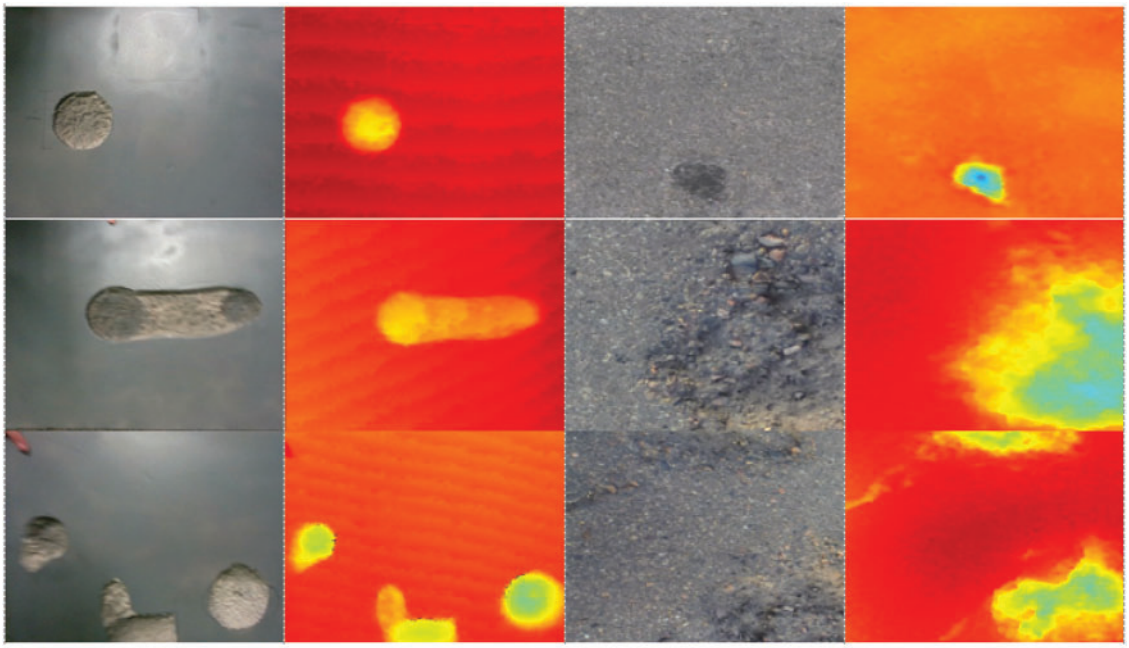

Figure 3: Examples of pothole synthetic data and Pothole 600 data: (a) RGB synthetic data, (b) TDM synthetic data, (c) RGB Pothole 600 data and (d) TDM Pothole 600 data

Table 1: Comparison of the performance of YOLO-R models before and after the addition of synthetic data

Model (YOLO-R)	mAP				
	IOU = 0.2	IOU = 0.3	IOU = 0.4	IOU = 0.5	IOU = 0.6
Before addition of data	83.15%	83.15%	83.15%	81.51%	75.18%
After addition of data	85.20%	85.20%	84.02%	82.94%	81.24%

YOLO-R, YOLO-TDM and YOLO-R/TDM were trained based on the dataset reinforced through synthetic data, and the performance was evaluated by comparing the mAP according to the change of the IOU threshold. The inference of the detection result depends on the IOU threshold, and if the detection result is above the threshold, it is considered to be correctly detected. In general, an IOU threshold of 0.5 is used because it focuses on how accurately an object is found while detecting an object. However, when detecting a pothole, the purpose of quickly finding a pothole is larger; therefore, the IOU threshold value was set to 0.2, and the IOU threshold value was additionally increased and the change in detection accuracy was examined. Looking at the performance comparison results for each detection model according to the IOU threshold change presented in Tab. 2, when the IOU is 0.2 or 0.3 (obtained through the proposed model), YOLO-R/TDM, improved up to 10.71% more than YOLO-R, and up to 3.08% more than YOLO-TDM. Even when the IOU threshold increased to 0.5, the accuracy of the proposed model was observed to be 8.88% and 4.31% higher than those of YOLO-R and YOLO-TDM, respectively, and the detection performance of the proposed model was

excellent. Therefore, it can be said that the proposed model is more suitable for quickly determining the approximate location and number of potholes rather than detecting the exact location of the potholes.

Table 2: Performance comparison by detection model according to IOU threshold change

Model	mAP				
	IOU = 0.2	IOU = 0.3	IOU = 0.4	IOU = 0.5	IOU = 0.6
YOLO-R	85.20%	85.20%	84.02%	82.94%	81.24%
YOLO-TDM	92.83%	92.83%	91.17%	87.51%	77.86%
YOLO-R/TDM	95.91%	95.91%	94.31%	91.82%	84.93%

Fig. 4 shows a sample of the Pothole-600 test result for each model when the IOU is 0.5. Looking at rows 1–3, the potholes that were not detected by YOLO-R were detected through its fusion with YOLO-TDM. In rows 4 and 5, on the contrary, potholes that were not detected by YOLO-TDM can be detected through its fusion with YOLO-R. It was confirmed that the detection performance was improved using the converged model. In the figure, the white bounding box is the ground truth and the blue bounding box is the detection result.

The 2D vision method has a high execution speed, but the detection accuracy is not satisfactory. By contrast, the 3D reconstruction method has high accuracy but has a disadvantage that the execution speed is low due to the large amount of computation. The proposed algorithm showed high execution speed and good detection performance using CNN among the 2D vision methods. Tab. 3 compares the detection performance and data processing speed of the pothole detection state-of-the-art models and the proposed model. In the case of the models proposed by Mikhailiuk et al. [14], Fan et al. [29], Zhang [37], and Fan et al. [34], all the 3D reconstruction-type pothole detection methods show high detection performance but indicate disadvantage in some models with quite low data processing speed due to the large amount of computation for 3D reconstruction. The detection rates the models were 73.4%, 84.8%, 98.7% and 98.7%, respectively, while the proposed system showed a much higher detection rate than the models proposed by Mikhailiuk et al. [14] and Zhang [37] at 96.2% and showed a slightly lower detection performance than the models proposed by the other researchers [29,34]. However, the execution speed of the proposed model was ∼6 times faster than that reported by Fan et al. [29] and showed that the data could be processed two times faster than that of proposed by Fan et al. recently [34]. The difference in execution speed between the model proposed by Mikhailiuk et al. [14] and YOLO-R/TDM was very small, 0.1 ms, but YOLO-R/TDM showed 11.4% higher detection performance. In addition, the experimental environment of the previously reported models [14,29,34,37] is NVIDIA RTX 2080ti, while that for the proposed model is NVIDIA 1080ti, which provides much faster execution speed, although data are processed in a relatively poor GPU environment. It was confirmed that the proposed model is useful for real-time use. Through the fusion of RGB and TDM, it was possible to improve the detection accuracy, which is a disadvantage of the 2D vision method, and through this, the performance similar to the 3D reconstruction method was secured.

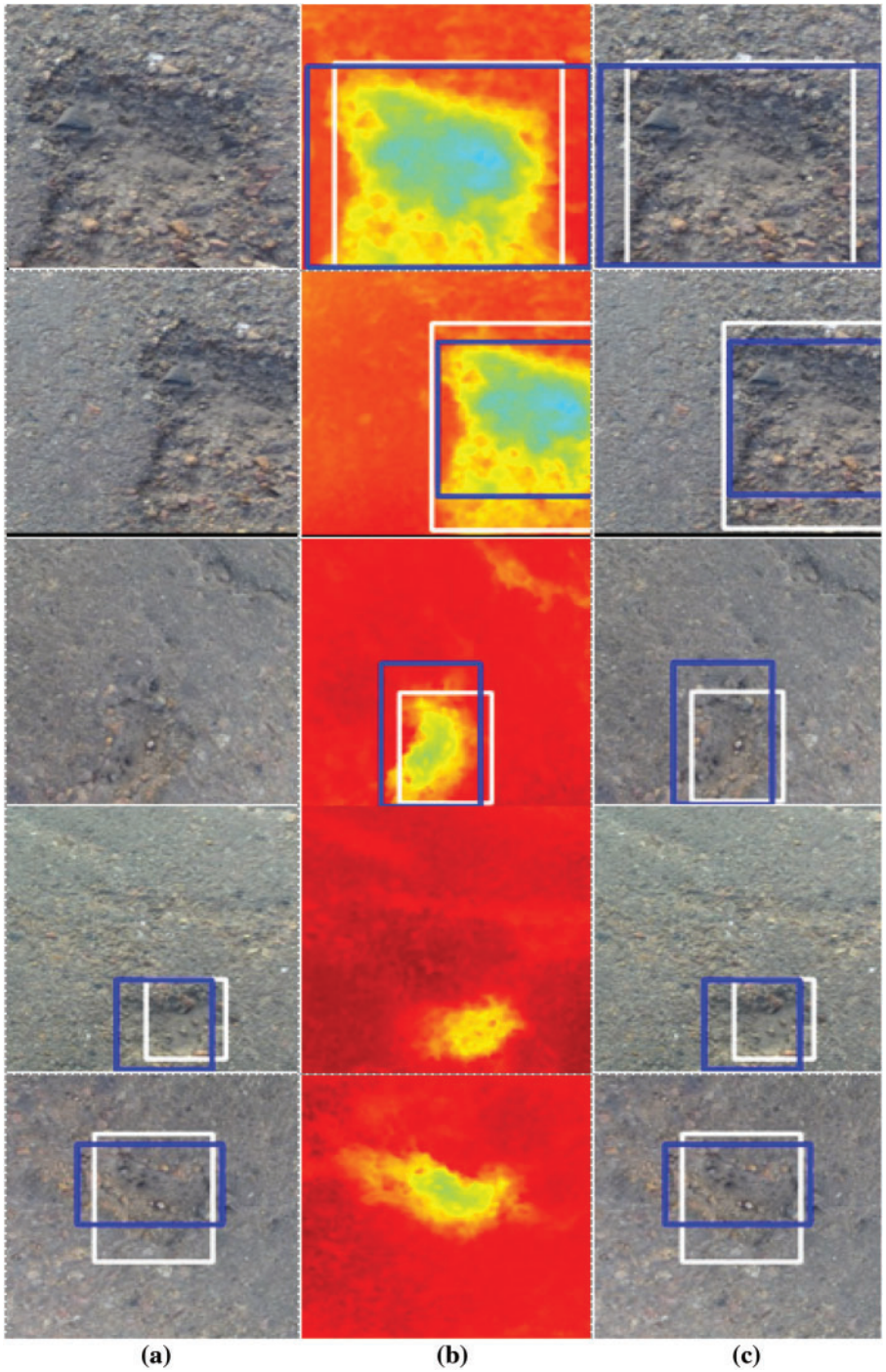

Figure 4: Pothole-600 test result sample by model when IOU = 0.5: (a) YOLO-R, (b) YOLO-TDM and (c) YOLO-R/TDM

Table 3: Comparison of the detection performance of pothole between state-of-the-art models and the proposed model

Dataset	Method	Correct detection	False alarm	Missed detection	Runtime (ms)
Dataset 1	[37]	11	11	0	33.19
	[14]	22	0	0	22.90
	[29]	22	0	0	117.72
	[34]	21	1	0	47.21
	YOLO-R/TDM	22	0	0	23.45
Dataset 2	[37]	42	10	0	30.77
	[14]	40	8	4	21.39
	[29]	51	1	0	124.53
	[34]	52	0	0	45.32
	YOLO-R/TDM	49	9	3	24.15
Dataset 3	[37]	5	0	0	35.72
	[14]	5	0	0	26.24
	[29]	5	0	0	132.44
	[34]	5	0	0	49.90
	YOLO-R/TDM	5	0	0	23.25
Total	[37]	58	21	0	33.23
	[14]	67	8	4	23.51
	[29]	78	1	0	124.90
	[34]	78	1	0	47.48
	YOLO-R/TDM	76	9	3	23.61

4 Conclusions

A pothole is a type of damage to the road surface, and it can cause traffic accidents and serious damage to vehicles. To solve these problems, it is necessary to quickly detect and repair the pothole. The existing pothole detection system involves public officials or professional inspectors and is very subjective and inefficient; therefore the requirement for a safe, objective and powerful pothole detection system is increasing. Among the various pothole detection methods, the 3D reconstruction method is in the spotlight because it provides high accuracy; however, it is not widely used owing to high equipment cost. It has low detection speed and there have been difficulties in its real-time utilization. Furthermore, although the detection performance of the 2D vision method has been improved owing to the recent development of CNNs, there are limitations that require a lot of data learning. Therefore, in this paper, an efficient real-time pothole detection system based on the fast 2D vision method is presented. First, the lack of training data increased the learning performance of the model by creating synthetic data, adding them to the training data, and converting them into TDM. Afterward, each pothole was detected based on YOLO using RGB images and TDM, and the detection performance was improved by complementing the detection results by performing fusion

and optimization based on NMS. Consequently, a 10.71% improvement in performance was observed when using YOLO-R/TDM compared with that using only YOLO-R and showed an execution speed of 23.61 ms, which showed at least two times faster detection performance than the existing 3D reconstruction method.

NMS was used for the fusion of detection results to find the optimal bounding box through complementation among the detected bounding boxes, but there were cases where some incorrectly detected bounding boxes could not be removed. This was reflected in the fusion result, and the false alarm rate increased, resulting in performance degradation. Therefore, in the near future, we plan to optimize the conditions required for extracting the optimal bounding box and design an algorithm that simultaneously determines and removes the erroneously detected bounding box.

Acknowledgement: The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Funding Statement: This research was funded by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MOE) (No. 2021R1I1A3055973) and the Soonchunhyang University Research Fund.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. S. Miller and W. Y. Bellinger, "Distress identification manual for the long-term pavement performance program," in Office of Infrastructure Research and Development Federal Highway Admin, Washington, DC, USA: Tech Rep FHWA-HRT-13-092, 2014. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/28612.

[2] R. Fan, "Real-time computer stereo vision for automotive applications," Ph. D. Dissertation, University of Bristol, 2018.

[3] D. Dongre, R. Jangam, A. Gosavi, M. Sonekar, R. Kumbhalkar *et al.,* "Advanced drainage and pothole navigation system based on IoT," in *Int. Conf. on Trends in Electronics and Informatics*, Tirunelveli, India, pp. 395–398, 2021.

[4] H. K. Kakmal and M. B. Dissanayake, "Pothole detection with image segmentation for advanced driver assisted systems," in *IEEE Int. Women in Engineering Conf. on Electrical and Computer Engineering*, Bhubaneswar, India, pp. 308–311, 2020.

[5] T. Kim and S. K. Ryu, "Review and analysis of pothole detection methods," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 5, no. 8, pp. 603–608, 2014.

[6] R. Fan, X. Ai and N. Dahnoun, "Road surface 3D reconstruction based on dense subpixel disparity map estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3025–3035, 2018.

[7] S. Mathavan, K. Kamal and M. Rahman, "A review of three-dimensional imaging technologies for pavement distress detection and measurements," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2353–2362, 2015.

[8] J. Erikson, L. Girod and B. Hull, "The pothole patrol: Using a mobile sensor network for road surface monitoring," in *Int. Conf. on Mobile Systems, Applications, and Services*, New York, NY, USA, pp. 29–39, 2008.

[9] K. T. Chang, J. R. Chang and J. K. Liu, "Detection of pavement distress using 3D laser scanning technology," in *Int. Conf. on Computing in Civil Engineering*, Cancun, Mexico, pp. 1–11, 2005.

[10] Q. Li, M. Yao, X. Yao and B. Xu, "A real-time 3D scanning system for pavement distortion inspection," *Measurement Science and Technology*, vol. 21, no. 1, pp. 15702–15709, 2009.

[11] C. Koch and I. Brilakis, "Pothole detection in asphalt pavement images," *Advanced Engineering Informatics*, vol. 25, no. 3, pp. 507–515, 2011.
[12] D. Joubert, A. Tyatyantsi, J. Mphahlehle and V. Mphahlehle, "Pothole tagging system," in *Robotics and Mechatronics Conf. of South Africa*, Pretoria, South Africa, pp. 1–4, 2011.
[13] I. Moazzam, K. Kamal, S. Mathavan, S. Usman and M. Rahman, "Metrology and visualization of potholes using the microsoft kinect sensor," in *IEEE Int. Conf. on Intelligent Transportation Systems*, The Hague, Netherlands, pp. 1284–1291, 2013.
[14] A. Mikhailiuk and N. Dahnoun, "Real-time pothole detection on TMS320C6678 DSP," in *IEEE Int. Conf. on Imaging Systems and Techniques*, Chania, Greece, pp. 123–128, 2016.
[15] A. Dhiman and R. Klette, "Pothole detection using computer vision and learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3536–3550, 2020.
[16] K. C. Wang, "Challenges and feasibility for a comprehensive automated survey of pavement conditions," in *Int. Conf. on Applications of Advanced Technologies in Transportation Engineering*, Beijing, China, pp. 531–536, 2004.
[17] Z. Zhang, X. Ai, C. K. Chan and N. Dahnoun, "An efficient algorithm for pothole detection using stereo vision," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Florence, Italy, pp. 564–568, 2014.
[18] P. D. Riya, K. R. Nakulraj and A. A. Anusha, "Pothole detection methods," in *Int. Conf. on Inventive Computation Technologies*, Coimbatore, India, pp. 120–123, 2018.
[19] S. Li, C. Yuan, D. Liu and H. Cai, "Integrated processing of image and GPR data for automated pothole detection," *Journal of Computing in Civil Engineering*, vol. 30, no. 6, pp. 04016015, 2016.
[20] R. Fan, M. J. Bocus and N. Dahnoun, "A novel disparity transformation algorithm for road segmentation," *Information Processing Letters*, vol. 140, pp. 18–24, 2018.
[21] S. K. Ryu, T. Kim and Y. R. Kim, "Image-based pothole detection system for ITS service and road management system," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–10, 2015.
[22] J. Dharneeshkar, D. V. Soban, S. A. Aniruthan, R. Karthika and P. Latha, "Deep learning based detection of potholes in Indian roads using YOLO," in *Int. Conf. on Inventive Computation Technologies*, Coimbatore, India, pp. 381–385, 2020.
[23] P. A. Chitale, K. Y. Kekre, H. R. Shenai, R. Karani and J. P. Gala, "Pothole detection and dimension estimation system using deep learning (YOLO) and image processing," in *Int. Conf. on Image and Vision Computing New Zealand*, pp. 1–6, 2020.
[24] K. Azhar, F. Murtaza, M. H. Yousaf and H. A. Habib, "Computer vision-based detection and localization of potholes in asphalt pavement images," in *IEEE Canadian Conf. on Electrical and Computer Engineering*, Vancouver, BC, Canada, pp. 1–5, 2016.
[25] M. R. Jahanshahi, F. Jazizadeh, S. F. Masri and B. Becerik-Gerber, "Unsupervised approach for autonomous pavement-defect detection and quantification using an inexpensive depth sensor," *Journal of Computing in Civil Engineering*, vol. 27, no. 6, pp. 743–754, 2013.
[26] C. Koch, G. M. Jog and I. Brilakis, "Automated pothole distress assessment using asphalt pavement video data," *Journal of Computing in Civil Engineering*, vol. 27, no. 4, pp. 370–378, 2013.
[27] E. N. Ukhwah, E. M. Yuniarno and Y. K. Suprapto, "Asphalt pavement pothole detection using deep learning method based on YOLO neural network," in *Int. Seminar on Intelligent Technology and its Applications (ISITIA)*, Surabaya, Indonesia, pp. 35–40, 2019.
[28] C. Koch, K. Georgieva, V. Kasireddy, B. Akinci and P. Fieguth, "A review on computer vision-based defect detection and condition assessment of concrete and asphalt civil infrastructure," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 196–210, 2015.
[29] R. Fan, U. Ozgunalp, B. Hosking, M. Liu and I. Pitas, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Transactions on Image Processing*, vol. 29, pp. 897–908, 2019.
[30] U. Ozgunalp, R. Fan, X. Ai and N. Dahnoun, "Multiple lane detection algorithm based on novel dense vanishing point estimation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, pp. 621–632, 2017.

[31] W. Sun, L. Dai, X. Zhang, P. Chang and X. Z. He, "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Applied Intelligence*, Online First Article, pp. 1–16, 2021.
[32] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 779–788, 2016.
[33] D. Feng, C. Haase-Schutz, L. Rosenbaum, H. Hertlein, C. Glaeser *et al.,* "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2021.
[34] R. Fan, U. Ozgunalp, Y. Wang, M. Liu and I. Pitas, "Rethinking road surface 3-D reconstruction and pothole detection: From perspective transformation to disparity map segmentation," *IEEE Transactions on Cybernetics*, Early Access Article, pp. 1–10, 2021.
[35] J. C. Tsai, K. T. Lai, T. C. Dai, J. J. Su, C. Y. Siao *et al.,* "Learning pothole detection in virtual environment," in *Int. Automatic Control Conf.*, Hsinchu, Taiwan, pp. 1–5, 2020.
[36] R. Fan, "Pothole-600 dataset," 2021. [Online]. Available: https://sites.google.com/view/pothole-600/dataset?authuser=0.
[37] Z. Zhang, "Advanced stereo vision disparity calculation and obstacle analysis for intelligent vehicles," Ph. D. Dissertation, University of Bristol, 2013.