Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 8870211, 8 pages
https://doi.org/10.1155/2020/8870211

*Research Article*

# A Framework for Detecting Vehicle Occupancy Based on the Occupant Labeling Method

**Jooyoung Lee**,[1] **Jihye Byun**,[2] **Jaedeok Lim**,[3] **and Jaeyun Lee**[3]

[1]*The Cho Chun Shik Graduate School of Green Transportation, Korea Advanced Institute of Science and Technology, Daejeon 34051, Republic of Korea*
[2]*Center for Eco-Friendly Smart Vehicle, Korea Advanced Institute of Science and Technology, Daejeon 34051, Republic of Korea*
[3]*Technical Research Center, GnT Solution, Inc., Seoul 07255, Republic of Korea*

Correspondence should be addressed to Jihye Byun; snowflower@kaist.ac.kr

Received 24 August 2020; Revised 31 October 2020; Accepted 17 November 2020; Published 3 December 2020

Academic Editor: Ladislav Routil



High-occupancy vehicle (HOV) lanes or congestion toll discount policies are in place to encourage multipassenger vehicles. However, vehicle occupancy detection, essential for implementing such policies, is based on a labor-intensive manual method. To solve this problem, several studies and some companies have tried to develop an automated detection system. Due to the difficulties of the image treatment process, those systems had limitations. This study overcomes these limits and proposes an overall framework for an algorithm that effectively detects occupants in vehicles using photographic data. Particularly, we apply a new data labeling method that enables highly accurate occupant detection even with a small amount of data. The new labeling method directly labels the number of occupants instead of performing face or human labeling. The human labeling, used in existing research, and occupant labeling, this study suggested, are compared to verify the contribution of this labeling method. As a result, the presented model's detection accuracy is 99% for the binary case (2 or 3 occupants or not) and 91% for the counting case (the exact number of occupants), which is higher than the previously studied models' accuracy. Basically, this system is developed for the two-sided camera, left and right, but only a single side, right, can detect the occupancy. The single side image accuracy is 99% for the binary case and 87% for the counting case. These rates of detection are also better than existing labeling.

## 1. Introduction

As the vehicle supply increases, the road infrastructure capacity is relatively reduced, so continuous construction of new roads is needed in many areas around the globe. However, increasing the road infrastructure capacity by building more roads is costly and time-consuming, so there is a limit that cannot accommodate the vehicle growth rate. In order to solve this problem, some policies have been implemented to encourage carpooling, such as reducing travel time through HOV lanes or providing discounts on congestion tolls from multipassenger vehicles [1]. To enforce this policy, technology for detecting vehicle occupants is essential. Currently, when enforcing HOV lane control policies or providing congestion toll discounts to multipassenger vehicles, employees visually estimate the number of passengers in each vehicle by checking the video data in management centers [2]. This manual method is labor-intensive, lowers operational efficiency, and increases labor costs. In the United States, which is cracking down on the illegal use of HOV lanes, the actual violation rate is about 50–80%, but the crackdown rate is reported to be less than 10% [3]. In South Korea, where discounts on congestion tolls are provided, congestion is likely to increase even more during peak hours due to inspection of the number of passengers in each vehicle at the toll gates and the collection of the tolls.

To solve this problem, various studies were conducted to automate the vehicle occupant estimate process. The research can be divided into two detection technology areas: using in-vehicle sensors [4–10] and using the image data from outside cameras [11–17]. When using in-vehicle sensors, the accuracy is generally high; however, all vehicles

need to be equipped with devices that can detect the number of passengers. Such devices usually use video cameras, which causes privacy concerns for many people. Therefore, the use of this method is impractical. Moreover, most studies that detect occupants using outside cameras had limited scope. For example, they can only detect the number of passengers in the front seat [12–14], only count the number of children onboard [16], or only determine if two or more passengers have boarded a vehicle. In particular, in [17], an 88% detection accuracy was achieved using image data captured outside the vehicle by one front and one side camera. This accuracy level is applicable to the real world, so pilot services were performed in several regions in the United States.

In the vehicle occupant detection field, there is another limitation in that only newly acquired images can be used as training data. Therefore, an algorithm is needed to achieve a high detection rate even with a small data set. In previous studies, a two-stage detection algorithm was used to overcome this limitation. Generally, the two-stage detection algorithm first detects the window area in the vehicle images and then detects the number of passengers in the window area only [15]. However, this algorithm has some limitations due to its complicated learning process and the increased network size, which increases the required calculation times.

Therefore, this study proposes an overall algorithmic framework that effectively detects vehicle occupants using left and right side photographic data from the vehicle exterior in a one-step process using a small amount of data. Specifically, we present a new data labeling method to accurately detect the number of occupants. The new labeling method directly labels the number of occupants instead of performing face or human labeling, which is a widely used method for image detection. Based on this advanced labeling method, this study contains only a single-stage detection algorithm. A decrease in the detection stage shrinks the network size, number of samples, and detection time.

The structure of this paper is as follows: the second section introduces an image acquisition system for detecting in-vehicle occupants and describes a new occupancy labeling method and acquired image data set; the third section describes the structure of the deep neural network used to detect occupants; the fourth section presents a discussion of the results of the presented algorithm in this study; and the final section summarizes the conclusions and implications of this study.

## 2. Image Acquisition and New Occupancy Labeling Method

Two infrared ray cameras, infrared ray illuminators, and a laser trigger acquire the images used for training and testing. An overview of the image acquisition system is shown in Figure 1. The cameras are located on the left and right sides of the vehicle. Through various tests, the research team determined the optimal specifications of the locations, heights, and angles of the cameras [18]. The infrared ray illuminators are used to improve the images when there is not enough visible light, such as at night or when the windows of the vehicles are tinted. The laser trigger detects the vehicle's entry into the detection zone that has the cameras. When the trigger recognizes a vehicle, the infrared ray cameras take images of the left and right sides of the vehicle. Then, the cameras send the frames to the server, and the accumulated images are used for training. When detecting vehicle occupancy, the images do not need to be transmitted to the server since they are treated by the on-site system.

As mentioned in the Introduction, previous research has labeled objects, such as faces, humans, and windows, and this labeling method has some benefits: (i) the number of labeling types, as the method needs one or two kinds of labels; (ii) securing a large number of learning samples since every image has to have one or more windows and a human. However, the method needs two stages, such as finding windows and then faces or an algorithm to divide the row of occupants. It leads to more times for calculation and higher error rates. To overcome the limitation, this study adopts a new labeling methodology to determine how many people are in the front and rear passenger seats. Therefore, each image must have two labels among six kinds of labels: one person in the front seat or two people in the front seat, and 0, 1, 2, or 3 people in the rear seat, as shown in Figure 2.

## 3. Vehicle Occupancy Detection Methodology

Figure 3 shows the proposed methodology for detecting occupants using the proposed labeling method in this study. An independently trained occupancy detection model is used for the images on each side, and passengers in the front seat are detected from the right side. As for the detection of occupants in the rear seat, both the left and right side images are used, and the number of occupants in the rear seat is determined using the higher detection score that results from comparing the detection scores obtained from the images of both sides. After that, the numbers of occupants in the front seat and the rear seat are added to obtain the total number of occupants.

This study trained the detection model and tested the results in the MATLAB 2019b environment. We used the Faster RCNN detection method, which has a high detection accuracy, instead of a unified detection algorithm, such as Yolo or an SSD with high speed [19]. The Faster RCNN method was introduced in [20], and it can detect multiple objects in one image with high accuracy and speed. This speeds up processing the regional-based CNN algorithm proposed in [21]. Specifically, the region proposal network (RPN), which is based on a fully convolutional network, was introduced to derive the region proposals from the feature map of the input image, as it replaces the selective search, which was a bottleneck of the training process. The RPN slides a 3 x 3 spatial window on a feature map to predict the region proposals, called multiple anchors, for each window. An anchor is the bounding box of the number of occupants that need to be detected in the input image. As in the previous paper, nine combinations of three sizes (128, 256, and 512) and three ratios (2 : 1, 1 : 1, and 1 : 2) of the anchor box were used for training in this paper. The derived anchors are classified into region proposals if the IoU (Intersection

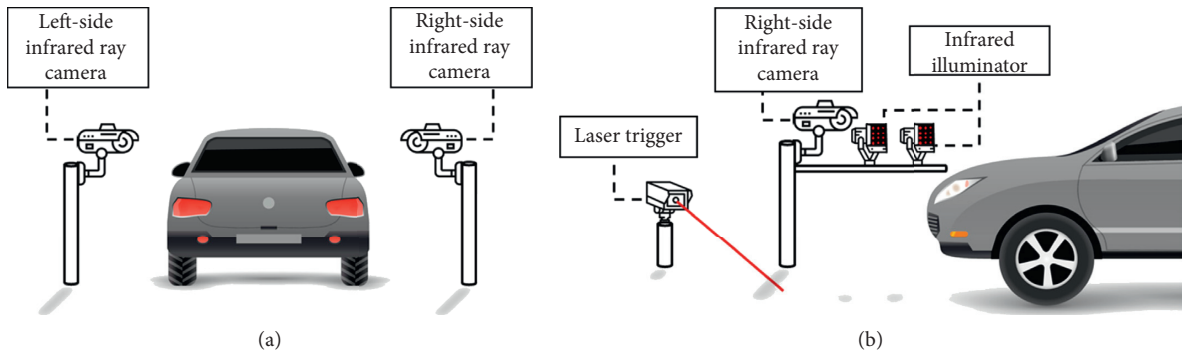


Figure 1: Overview of the image acquisition system for vehicle occupancy detection. (a) Rear view. (b) Side view.


Figure 2: Examples of the new labeling method. (a) Front seat: 1, rear seat: 1 case. (b) Front seat: 2, rear seat: 2 cases. (c) Front seat: 1, rear seat: 0 cases. (d) Front seat: 2, rear seat: 1 case.

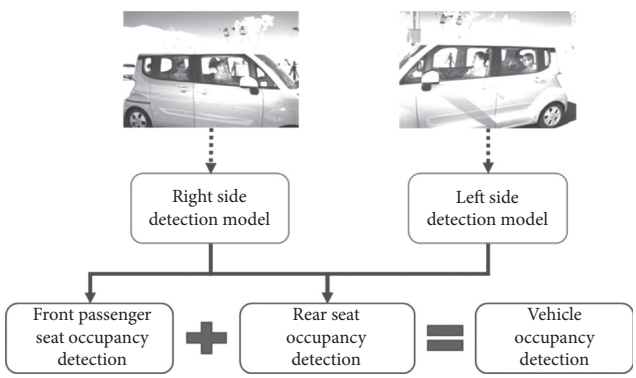


Figure 3: Overview of the vehicle occupancy detection methodology using both right and left side images.

over Union, see 1) with the ground truth box is higher than 0.7 or if it is the highest. If IoU is lower than 0.3, it is classified as background.

$$\text{IoU} = \frac{\text{anchor} \cap \text{ground truth box}}{\text{anchor} \cup \text{ground truth box}}. \quad (1)$$

An RoI (Region of Interest) maxPooling layer is used to fit different size proposed regions that are derived from the RPN to the same size. After the RoI pooling process, the softmax classifier, which classifies the occupants, and the box regressor, which estimates the bounding box, are trained. Therefore, we used the following multitask loss function for training, which is the sum of the $\mathscr{L}_{\text{clf.}}$ (loss function for classification) and the $\mathscr{L}_{\text{bbox}}$ (loss function for bounding box detection).

$$\mathscr{L} = \mathscr{L}_{\text{clf.}} + \mathscr{L}_{\text{bbox}},$$

$$\mathscr{L}(p_i, c_i) = \frac{\sum_i \mathscr{L}_{\text{clf.}}\left(p_i, p_i^{g.t.}\right)}{N_{\text{clf.}}} + \frac{\lambda \sum_i p_i^{g.t.} \cdot \mathscr{L}_1\left(c_i - c_i^{g.t.}\right)}{N_{\text{bbox}}}, \quad (2)$$

where $p_i$ is the predicted probability of anchor $i$, which is an object, and $p_i^{g.t.}$ is the ground truth label of whether anchor $i$ is an object or a background. $c_i$ indicates the predicted four parameterized coordinates of anchor $i$: $x$, $y$ position, width, and height. $c_i^{g.t.}$ is the ground truth coordinate of anchor $i$, and $N_{\text{clf.}}$ and $N_{\text{bbox}}$ represent the normalization term, which is set to be the minibatch size and the number of anchor

locations, respectively. $\lambda$ is the balancing parameter that makes $\mathscr{L}_{\text{clf.}}$ and $\mathscr{L}_{\text{bbox}}$ of approximately the same weight. In case of the bounding box regression, the coordinates and the training through the $\mathscr{L}_1$ loss function are estimated as follows:

$$
\begin{aligned}
c_x &= \frac{x - x_a}{w_a}, \\
c_y &= \frac{y - y_a}{h_a}, \\
c_w &= \log\frac{w}{w_a}, \\
c_h &= \log\frac{h}{h_a}, \\
c_x^{g.t.} &= \frac{x^{g.t.} - x_a}{w_a}, \\
c_y^{g.t.} &= \frac{y^{g.t.} - y_a}{h_a}, \\
c_w^{g.t.} &= \log\frac{w^{g.t.}}{w_a}, \\
c_h^{g.t.} &= \log\frac{h^{g.t.}}{h_a},
\end{aligned}
\tag{3}
$$

where $x$, $y$, $w$, and $h$ are the coordinates of the anchor and the bounding box: $x$, $y$ position, width, and height, respectively. The variables $x$, $x_a$, and $x^{g.t.}$ indicate the predicted bounding box, anchor box, and ground truth box, respectively, and their meaning is the same as the variables $y$, $w$, and $h$.

In order to train the effective classifier using the Faster RCNN, selecting a pretrained CNN for image feature extraction is important. In this study, we used the Inception-v3 network, which has high accuracy, small model size, and short calculation time, to derive the feature map of the input image used in the RPN and the occupant classification process [22]. In addition, transfer learning was performed using a pretrained Inception-v3 network of over 1 million images in the ImageNet database. The Inception-v3 network is an improved version of GoogLeNet [23], which was released in 2014 with 23.9 million parameters. GoogLeNet features an inception module that allows dense processing of matrix calculations while reducing the connectivity between the nodes in the network configuration. In addition, Inception-v3 improves the kernel used for convolution operations by introducing a new structured inception module that uses the 5 x 5 convolution operation twice for the 3 x 3 convolution operation and replaces the 3 x 3 convolution operation with the 1 x 3 and 3 x 1 convolution operations to reduce the computational complexity. In addition, convolution operations and pooling processes were performed in parallel, and then in concatenation, to improve the representational bottleneck, which is a phenomenon in which the amount of information is greatly reduced when the dimension is reduced excessively in a neural network. Moreover, according to [24], Inception-v3 achieved an accuracy of over 78.1% on ImageNet data sets. To apply Inception-v3 to the Faster RCNN structure, we removed the last three layers, which perform image classification, from the Inception-v3 network and added a feature extraction layer. Afterward, to form the Faster RCNN, a new classification layer and the RPN were added to fit the occupant label defined in this study. The overall structure of the model that detects occupants from single side images is shown in Figure 4.

## 4. Results and Discussion

Randomly sampled from 1,246 image sets, 1,000 image sets were used for model training, and 246 image sets were used for the detection accuracy test to analyze the vehicle occupant detection framework's performance. Model training was performed using a Stochastic Gradient Descent with momentum solver with a momentum of 0.9, and the learning rate was fixed at 0.001 for the entire training process. Previously, a 4-step method was used to train the Faster RCNN; the training of this study model was performed using an end-to-end method, which has improved the training efficiency.

In addition, to evaluate the efficiency of the labeling method presented in this study, we compared it to a model that uses human labeling methods, using the same data set and the same network structure. The human labeling method is a technique for labeling each person present in an image as an individual object. This method is generally used in vehicle occupant detection area and human detection tasks [12–17]. Two scenarios were used to compare the detection accuracy between the two labeling methods. The first scenario uses both side cameras, assuming an environment that requires high accuracy. The second scenario only uses one camera, assuming that the installation environment and cost are limited. In general, to use the HOV lane enforcement system, it is possible to simply calculate accuracy as a binary case that determines whether the total number of occupants in a vehicle is more than two or more than three, depending on the HOV lane types. If detailed seat occupant detection is possible, the system use increases. Therefore, in this study, the accuracy of the binary case, as well as the accuracy of the detected number of occupants in both the front and rear seats, was also calculated and compared. In the case of the model using the occupant labeling method proposed in this study, the detection result is derived from the number of occupants in the front and rear seats without additional postprocessing. However, in the case of the model trained by the comparative labeling method, the number of occupants in the front and the rear seats is recalculated using the human detection results. To distinguish between the front and rear seats, the B-pillar position in each image is calculated from the distance between the detection results. All the methods in this study were implemented using MATLAB 2019b and trained and

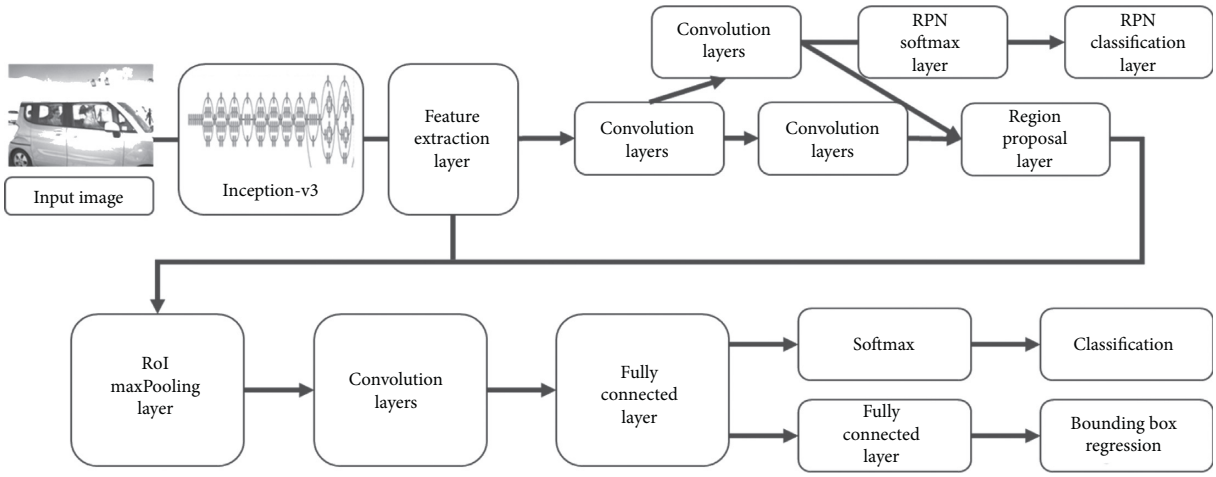


Figure 4: Description of Faster RCNN for one side image.

tested in a Dual Intel® Xeon® Silver 4114 CPU @ 2.20 GHz, 32 GB ram, and single NVIDIA GeForce RTX 2080 Ti computing environment.

The model training time of both labeling methods was about 7 hours for 1 K iterations. When testing these models, the occupant labeling model took an average of 3.4 seconds to output the detection results per an image set; however, it took the human labeling model about 7.6 seconds, more than twice the time of the occupant labeling method. An example of the vehicle occupancy detection test results for each model is shown in Figure 5. The occupant labeling model shows how many people were in the front and rear seats, while the human labeling model shows all the detected people and distinguishes the front seat from the rear one by the virtual B-pillar.

Table 1 compares the results of the occupant detection accuracy of the two models for 246 left and right side image sets. The presented occupant labeling method had a relatively high accuracy in all cases. The human labeling model was also highly accurate in the binary case when detecting two or more persons, but its accuracy was very low when detecting the actual number of occupants. There was an especially big difference in the detection rate of the number of passengers in the rear seat; the proposed labeling method robustly detects the passengers, even when parts of them are hidden in the captured images. In the human labeling method, the neural network learns a person's head and shoulders. When many people occupy a vehicle, especially in the rear seat, some parts of the passengers are often blocked, so it is difficult to identify accurate features. If there are several people riding in a vehicle, the rear seat often covers a part of one or more passengers. Thus, it is difficult to identify accurate human features. The detection accuracy of occupants in the proposed model in this study is 98% for the binary case and 91% for the counting case, which is higher than the accuracy level of the proposed model in [17], which was considered a state-of-the-art occupant detection accuracy.

The confusion matrix allows a more detailed analysis of the detection results of each model. In Figure 6, we present the confusion matrix of the test results for both models. The two matrices on the left are the model results using the occupant labeling method presented in this study, and the two matrices on the right are the model results using the human labeling model. The front and rear seat detection results for each model are shown in two confusion matrices. First, the front seat results are compared with 99.59% and 82.93%, respectively. In the occupant labeling model, one person was incorrectly detected as two people in one instance. However, there were four cases in which the control group detected that two people boarded while one person actually boarded, but 38 cases detected that one person boarded when two people boarded. A person in the passenger seat might be assumed to be a part of the vehicle or hidden by the driver and not be correctly detected as a person. Furthermore, the difference between the rear seat detection accuracy of the two models was 91.06% and 66.26%, respectively, which is greater than the front seat detection accuracy difference. In most cases, the proposed model in this study accurately detects the number of occupants, and the false detection results are maintained at ±1 person in comparison with the actual number. Therefore, it is evident that this model can robustly detect the results for the binary case. On the contrary, in the control model, the detection accuracy was very low when 3 people or more were on board, and there were many results that showed more than 2-person differences from the actual number of passengers. This is similar to the front seat detection result; the occupant labeling method was more effective when learning the appearance of part of the rear seat passengers. Generally, when using human labeling methods, it is difficult to detect people if some parts of them are hidden.

Instead of using both left and right images, the scenario performed detection using only one image on the right side, and the results are presented in Table 2. In the case of detecting occupants using only one camera image, the proposed model showed better results than the human detection method, similar to those in the case of using two camera images. Besides, when using one camera instead of two cameras, the accuracy of the rear seat decreased because the rear seat occupants are often concealed when using


(a) (b)

Figure 5: Examples of vehicle occupancy detection results for both models. (a) Occupant labeling result. (b) Human labeling result.

Table 1: Detection accuracy for both models using two images.

| Binary case | | | |
|---|---|---|---|
| Labeling method | 2+ | | 3+ |
| Occupant | **99.2%** | | **97.2%** |
| Human | 97.6% | | 82.1% |
| **Counting case** | | | |
| Labeling method | Front seat | Rear seat | Total |
| Occupant | **99.6%** | **91.1%** | **90.7%** |
| Human | 82.9% | 66.3% | 61.8% |

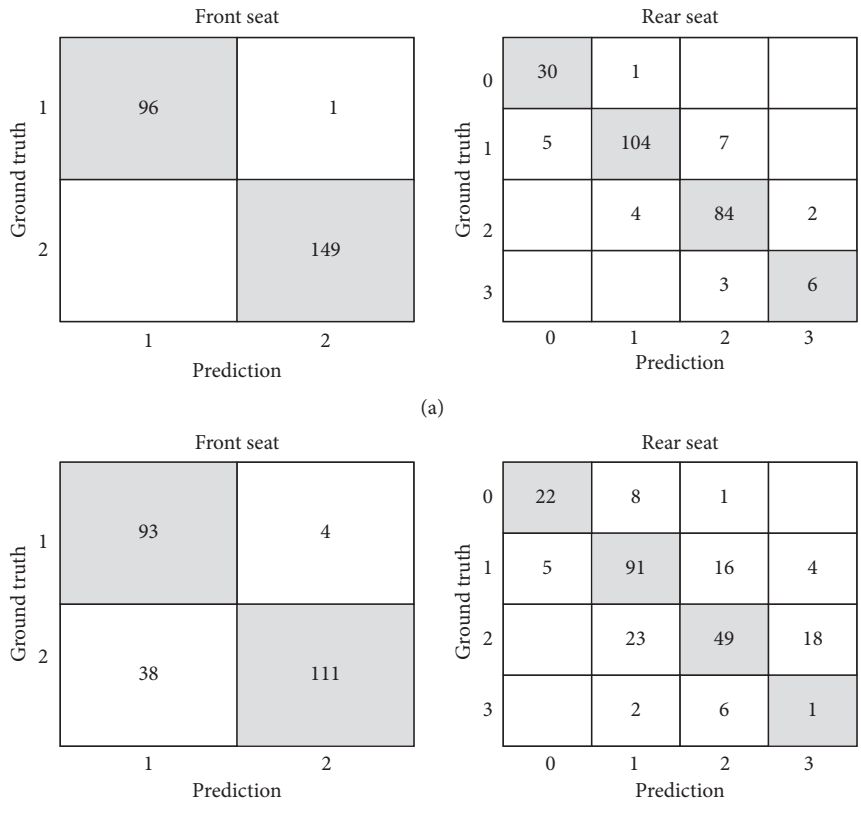


Figure 6: Confusion matrices for tested models. (a) Occupant labeling result. (b) Human labeling result.

Table 2: Detection accuracy for both models using one image

| Binary case | | | |
|---|---|---|---|
| Labeling method | 2+ | | 3+ |
| Occupant | **98.8%** | | **93.5%** |
| Human | 89.4% | | 74.8% |
| **Counting case** | | | |
| Labeling method | Front seat | Rear seat | Total |
| Occupant | **99.6%** | **87.4%** | **87.0%** |
| Human | 82.9% | 54.1% | 47.2% |

images from only one side, and the image from the opposite side cannot compensate for the smaller number of images. Nevertheless, the single-camera model in this study showed a level of accuracy of 87%, which is similar to that in [17], which showed the highest accuracy (88%) when using two cameras. In particular, in the binary case, the model's accuracy is more than 90%, so a single-camera detection model could be used effectively in an HOV enforcement system. Therefore, according to the purpose and environment of use, it is possible to use the proposed occupancy detection algorithm flexibly in this study.

## 5. Conclusions

To overcome increasing traffic and encourage carpooling, many governments use HOV lanes and provide discounted toll prices for cars that have multiple passengers. However, such systems usually determine the number of passengers in each vehicle by employing police officers or employees at the roadsides or near the toll booth cashiers. Thus, such human-resource-based occupancy detection systems lead to an operating budget burden and lower accuracy. Due to these limitations, several studies have attempted to achieve automated vehicle occupancy detection systems in a variety of ways, including the use of in-vehicle sensors or out-of-vehicle images. However, the image acquisition difficulty and the weakness of image processing technologies make implementing such detection systems hard to achieve.

To compensate for the shortages of previous research, this study suggests a new labeling method that detects passengers based on the number of occupants in each row of the vehicle instead of using human (or face) and window labeling. This new labeling method achieves Faster RCNN detection in a short time and with high accuracy. Also, this study had two scenarios: (i) using two cameras; (ii) using a one side camera due to the possible difficulties of setting two cameras on each side of the road in some areas. Each scenario has two cases: (i) binary: 1 or 2 and more ('2+')/1 to 2 or 3 and more ('3+'); (ii) counting the actual passenger numbers. Synthetically, the 2+ case had a similar detection accuracy to that of the occupant labeling method (99%), which this study suggests, and to that of the human labeling (97%) method, which is the usual detection method. However, the 3+ case showed a bigger gap (15%) between the two labeling methods, and the counting case had a huge difference between the two methods: occupants (91%) and humans (62%). The counting case is the actual number of passengers and the actual detection accuracy of the automated detection systems. The one side camera scenarios had similar patterns when it came to the detection results, but generally the accuracy was lower than when two cameras were used. In order, 2+, 3+, and the counting case scenarios had bigger differences with the labeling method, the occupant label had a detection accuracy of 87%, and the human labeling method had an accuracy of 46% at the counting case.

Since higher detection accuracy was achieved with the actual system, this study is important for further research on the way to increase the accuracy ratio. In the future, we will try various machine learning methodologies and neural networks to get more advanced results based on the new labeling method.

## Data Availability

The data used to support the findings of this study have not been made available because of GnT Solution's policy.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Acknowledgments

This research was financially supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D Program (Project no. 0002246).

## References

[1] K. Jang, K. Chung, D. R. Ragland, and C.-Y. Chan, "Safety performance of high-occupancy-vehicle facilities," *Transportation Research Record Journal of the Transportation Research Board*, vol. 2099, no. 1, pp. 132–140, 2009.

[2] L. Markkula, "HOV lanes: issues and options for enforcement," Tech. Rep. FHWA-AZ-04-552, Arizona Department of Transportation, Phoenix, AZ, USA,, 2004.

[3] S. Schijns and P. Mathews, "A breakthrough in automated vehicle occupancy monitoring systems for hov/hot facilities," in *Proceedings of the 12th HOV Systems Conference*, vol. 1, Houston, TX, USA, April 2005.

[4] S. Gautama, S. Lacroix, and M. Devy, "Evaluation of stereo matching algorithms for occupant detection," in *Proceedings of the International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV'99 (Cat. No.PR00378)*, pp. 177–184, Corfu, Greece, September 1999.

[5] M. Devy, A. Giralt, and A. Marin-Hernandez, "Detection and classification of passenger seat occupancy using stereovision," in *Proceedings of the IEEE Intelligent Vehicles Symposium 2000*, pp. 714–719, Dearborn, MI, USA, October 2000.

[6] M. Klomark, "Occupant detection using computer vision," M. S. thesis, Department of Electrical Engineering, Linköping University, Linköping, Sweden, 2000.

[7] Y. Lu, C. Marschner, L. Eisenmann, and S. Sauer, "The new generation of BMW child seat and occupant detection system SBE2," *International Journal of Automotive Technology*, vol. 3, no. 2, pp. 53–56, 2002.

[8] Y. Owechko, N. Srinivasa, S. Medasani, and R. Boscolo, "High performance sensor fusion for vision-based occupant detection," in *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*, vol. 2, pp. 1128–1133, Shanghai, China, October 2003.

[9] M. E. Farmer and A. K. Jain, "Occupant classification system for automotive airbag suppression," in *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, Madison, WI, USA, June 2003.

[10] F. Erlik Nowruzi, W. A. El Ahmar, R. Laganiere, and A. H. Ghods, "In-vehicle occupancy detection with convolutional networks on thermal images," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, June 2019.

[11] J. W. Wood, G. G. Gimmestad, and D. W. Roberts, "Covert camera for screening of vehicle interiors and HOV enforcement," in *Proceedings of the Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Defense and Law Enforcement II*, pp. 411–420, Orlando, FL, USA, September 2003.

[12] X. Hao, H. Chen, and J. Li, "An automatic vehicle occupant counting algorithm based on face detection," in *Proceedings of the 2006 8th International Conference on Signal Processing*, vol. 3, Beijing, China, November 2006.

[13] B. Xu, P. Paul, Y. Artan, and F. Perronnin, "A machine learning approach to vehicle occupancy detection," in *Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, vol. 2014, pp. 1232–1237, Qingdao, China, November 2014.

[14] Y. Artan, P. Paul, F. Perronin, and A. Burry, "Comparison of face detection and image classification for detecting front seat passengers in vehicles," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 1006–1012, Steamboat Springs, CO, USA, March 2014.

[15] Y. Artan, O. Bulan, R. P. Loce, and P. Paul, "Passenger compartment violation detection in HOV/HOT lanes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 395–405, 2016.

[16] B. Balci, B. Alkan, A. Elihos, and Y. Artan, "Front seat child occupancy detection using road surveillance camera images," in *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1927–1931, Athens, Greece, October 2018.

[17] A. Kumar, A. Gupta, B. Santra et al., "VPDS: an AI-based automated vehicle occupancy and violation detection system," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 9498–9503, 2019.

[18] J. Lim, S. Kim, S. Kang, and C. Kim, "Development occupancy detection in moving vehicle using computer vision technology," in *Proceedings of the US-Korea Conference on Science, Technology and Entrepreneurship*, Chicago, IL, USA, August 2019.

[19] A. Sachan, "Zero to hero: guide to object detection using deep learning: faster R-CNN," 2020, https://cv-tricks.com/object-detection/faster-r-cnn-yolo-ssd.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*vol. 39, no. 6, , pp. 1137–1149, NIPS, 2015.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[22] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," 2017, http://arxiv.org/abs/1605.07678.

[23] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Boston, MA, USA, June 2015.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, "Rethinking the inception architecture for computer vision," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, NV, USA, June 2016.