Hindawi
Advances in Multimedia
Volume 2022, Article ID 5283191, 6 pages
https://doi.org/10.1155/2022/5283191

Research Article

Semantic Detection of Vehicle Violation Video Based on Computer 3D Vision

Yue Dai

Chuzhou Vocational and Technical College, Anhui, Chuzhou 239000, China

Correspondence should be addressed to Yue Dai; k17301162@stu.ahu.edu.cn

Received 10 December 2021; Revised 20 February 2022; Accepted 13 March 2022; Published 9 April 2022

Academic Editor: Qiangyi Li

In order to study the semantic detection accuracy of 3D vehicle accident video, an accident detection method combining 2D image and 3D information was proposed. The 3D semantic bounding box generated by the 3D detection and tracking task of the vehicle is used to extract the motion features of the vehicle, it includes the trajectory of the vehicle and the dimension and orientation of the 3D bounding frame, and the 3D semantic bounding frame is used to establish the evaluation index of the accident detection. The experimental results show that the average loss function of each batch of 1000 images is calculated by the stochastic gradient descent method to update the parameter values. The learning rate was set to 0.001 in the first 30,000 iterations and 0.0001 in the last 10,000 iterations. The MOTA of the CEM algorithm is 78.4%, FP is 1.1%, and FN is 3.5%, and the MOTA of the 3-DCMK algorithm is 88.6%, FP is 0.9%, and FN is 1.9%. The MOTA of this method is 89.3%, FP is 0.9%, and FN is 1.2%. The 3D target semantic detection of vehicle accident video has stability and accuracy.

1. Introduction

With the development of autonomous driving industry, relevant research has been promoted. On-board sensors, on-board control systems, high-precision positioning maps, and other technologies have developed vigorously in recent years. However, for autonomous driving technology, safety is the primary factor to consider [1]. To ensure the safety and reliability of autonomous driving, we need a precise understanding of the surrounding environment. Automatic driving system converts sensor information into semantic information, and target detection, as the basis and an important step of semantic information extraction, plays an indispensable role in the automatic driving system. In the automatic driving system, the detection of objects such as people and vehicles is an important part. How to accurately identify objects and quickly locate them is a difficult problem for autonomous driving [2]. Different from the target detection in the picture, the environment of automatic driving is three-dimensional space, and three parameters of the center point and length, width, and height of the object need to be obtained. And for vehicles, it is more necessary to know the current driving direction [3]. Therefore, traditional two-dimensional target detection methods cannot meet the needs of autonomous driving, and a more accurate perception of the three-dimensional space around the vehicle is needed. However, the image data alone cannot well satisfy the perception tasks under various weather and extreme light conditions. A vehicle accident detection method based on 3D object detection to generate 3D semantic bounding box is proposed. The method extracts the vehicle posture from the convolutional neural network and predicts the 3D bounding box of the vehicle. 3D object detection and 2D to 3D extended Kalman filter were combined to recover the 6-dOF 3D vehicle attitude and tracking track in the video sequence, and the 3D semantic bounding frame was used to establish the evaluation index of accident detection. According to different data sources, 3D target detection can be roughly divided into three directions: target detection based on point cloud data, target detection based on image, and target detection based on the combination of heterogeneous data [4]. Point cloud data can be obtained by lidar or binocular camera. These three modes all have their own characteristics, which combine to produce different research methods,

image, as the most common research data in computer vision; although the picture cannot well represent the three-dimensional space, the 3D state of the object can be estimated with the help of the information on the image. Monocular cameras provide detailed information in the form of pixels to display shape and texture features at a greater scale. These features can be used to detect lane lines and traffic lights or other categories of information. However, the drawback of monocular cameras is also obvious: the lack of depth information. Due to the lack of depth information, most methods based on monocular vision are based on two-dimensional image target detection algorithms; firstly, the position of the object on the 2D image is obtained, and then the 3D target border parameter regression is carried out to get the final target position. Some of the previous detection methods on two-dimensional images can be directly performed using model-based or segmentation-based target detection; then, with the application of deep learning in the field of object detection, the deep learning method is gradually adopted for 3D object detection. In view of this research problem, Wang et al. used 3D pixel features to detect vehicles in images; a new model representation 3DVP is proposed, which is segmented after two-dimensional image detection and projected into three-dimensional space to obtain accurate 3D position information [5]. The concept of subcategory is proposed; that is, objects with similar attributes are regarded as the same category, and the 3DVP structure is used to construct subcategories; a convolution network is used to generate a heat map for each subclass of the proposed region; after the region of interest is pooled, the network outputs category classification and two-dimensional border estimation. The algorithm works well for the problems of occlusion and partial missing objects, but these scenes also appear as a category in the model dictionary; if the vehicle attitude does not appear in the model dictionary, it cannot be accurately detected. To overcome this problem, Ma et al. used a multitask network to estimate vehicle state, partial position, and shape. Car shapes generate 3D candidate box shapes from a series of key points. Firstly, the 2D candidate box is obtained through the improved Region Proposal Network (RPN), and then the candidate box is obtained based on the inferred 3D model matching [6]. With the development of deep learning, the algorithm obtains the two-dimensional target border through the neural network and then carries out the three-dimensional parameter regression. Lv et al. proposed Mono3D, which firstly generated numerous detection candidate boxes in three-dimensional space, then feature extraction was carried out for the candidate boxes, and scores were calculated for each candidate box projected on the image plane by using various features such as semantics or shapes, and regression of 3D candidate boxes was carried out [7]. However, due to the lack of depth information, the accuracy of position detection is poor. Based on the current research, a vehicle accident detection method based on 3D object detection to generate 3D semantic bounding box is proposed; this method extracts the vehicle posture from the convolutional neural network and predicts the 3D bounding box of the vehicle. Combining 3D object detection and 2D to 3D extended Kalman filter, the 6 degrees of 3D vehicle attitude and tracking trajectory are restored in the video sequence, and the 3D semantic bounding frame is used to establish the accident detection evaluation index. The test results show that each batch of 1000 images was trained by the stochastic gradient descent method, and their average loss function values were calculated to update the parameter values. The learning rate was set to 0.001 in the first 30,000 iterations and 0.0001 in the last 10,000 iterations. The MOTA of the CEM algorithm is 78.4%, FP is 1.1%, and FN is 3.5%, and the MOTA of the 3-DCMK algorithm is 88.6%, FP is 0.9%, and FN is 1.9%. The MOTA of this method is 89.3%, FP is 0.9%, and FN is 1.2%. The semantic detection of vehicle accident video for 3D targets has stability and accuracy [8].

2. Methods

2.1. Algorithm Framework. Figure 1 shows the framework of the proposed algorithm based on monocular vision for 3D object detection and multitarget tracking; the algorithm is mainly composed of three stages. Stage 1: a 3D object detector is trained using convolutional neural network to detect vehicles on the road. Stage 2: using extended Kalman filter to predict and update the status of the detected vehicles, 2D-3D multitarget tracking is generated, so as to realize 6 degrees of 3D target attitude recovery in the whole video sequence. In the last stage, a vehicle behavior function is established based on the trajectory and appearance of the moving vehicle to determine the occurrence of the accident.

2.2. Vehicle Accident Detection

2.2.1. 3D Vehicle Detection. The three-dimensional bounding frame of the vehicle can provide information such as position, dimension, and orientation, i.e., 9 degrees of freedom represent the motion state of the vehicle. (x, y, z) represents the position of the vehicle center point in the 3D coordinate system; (ω, h, L) represents the width, height, and length of the vehicle; (α, β, γ) are the deflection angle, heading angle, and pitch angle of the vehicle. Based on the darknet-53 network model, the model is trained end-to-end, and then 3D bounding box is formed by degrees of parameters and camera projection matrix. A feature description module similar to the feature pyramid network (FPN) is extracted from the convolution layer, which includes deeper convolution layer and deconvolution layer to capture image details. Higher resolution output tensors can capture richer image details, such as distant objects in the field of view, while lower resolution output tensors can capture more image context information. The normalized size of each input image is $416 * 416$, and the designed output category is vehicle, the output parameters are 9 degrees of freedom parameters plus a confidence parameter, three feature maps with different sizes are output, and then the tensor depth is shown in formula as follows:

$$3 * (10 + 1) = 33. \quad (1)$$

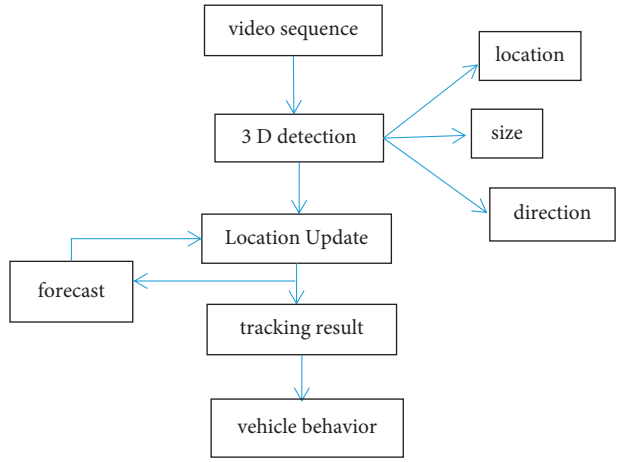

Figure 1: Schematic diagram of algorithm framework.

For these output parameters, a 3D Box Proposal Network (3DBPN) loss function is designed; 3DBPN loss function mainly includes position L_{loc}, dimension L_{size}, and orientation L_θ. x_i, y_i, z_i and $\hat{x}_i$, $\hat{y}_i$, $\hat{z}_i$ represent the real space coordinates and the predicted coordinates in space, respectively; w_i, l_i, h_i and $\hat{w}_i$, $\hat{l}_i$, $\hat{h}_i$ represent the real size and the predicted size of the object, respectively; θ_i and $\hat{\theta}_i$ represent the true and predicted directions, respectively. Different weight α is set for each loss function to further improve the prediction performance of parameters. The loss function is shown in formulas (2)–(5):

$$L_{\text{loc}} = a_{\text{loc}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} L_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{x}_i)^2 + (z_i - \hat{x}_i)^2 \right], \tag{2}$$

$$L_{\text{size}} = a_{\text{size}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} L_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{l_i} - \sqrt{\hat{l}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{x}_i} \right)^2 \right], \tag{3}$$

$$L_\theta = a_{\text{yaw}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} L_{ij}^{\text{obj}} \left(\theta_i - \hat{\theta}_i \right)^2, \tag{4}$$

$$L_{3DBPN} = 1^{3\ D10U<0.7} \left[L_{\text{loc}} + L_{\text{size}} + L_\theta \right]. \tag{5}$$

The candidate generation boxes with low score are filtered by the nonmaximal suppression algorithm to improve the 3D bounding boxes with high accuracy. The 3D detection algorithm of the vehicle provides the position prediction of the vehicle in the monitoring view, which is convenient for the next step of 3D tracking.

2.3. Vehicle Tracking. The relationship between frames is established in the current detection results, and the 2-3D extended Kalman filter (EKF) is fused to predict the motion and state of the vehicles on the road; thus, three-dimensional attitude tracking of the vehicle is realized [9]. After the 3D frame size and angle information of the vehicle detected in the initial frame are output, the initial posture of the vehicle is determined. In two consecutive frames, the extended Kalman filter is used to estimate the tracking state of the target in the next frame by 3D reprojection. The observation results are combined with 3D object detection and 2D detection to form all the observation models; the MRF model is used to select the appropriate observation results from a large number of generated hypothetical observations as the initial detection frame for the next frame. Combined with projection and backprojection operations, the locus points are applied to the center of the 3D bounding box. The recursive neural network LSTM is used to correlate and match the position of vehicles on the road across frames. The velocity is obtained from the LSTM model and is used to estimate the 3D pose to update the 3D position. According to the average size of the training data, the size of the 3D bounding box can be updated periodically with the distance of the visual field.

2.4. Accident Detection. Based on the motion attitude of the vehicle, this section establishes the traffic accident evaluation function A, which is composed of the changes of vehicle

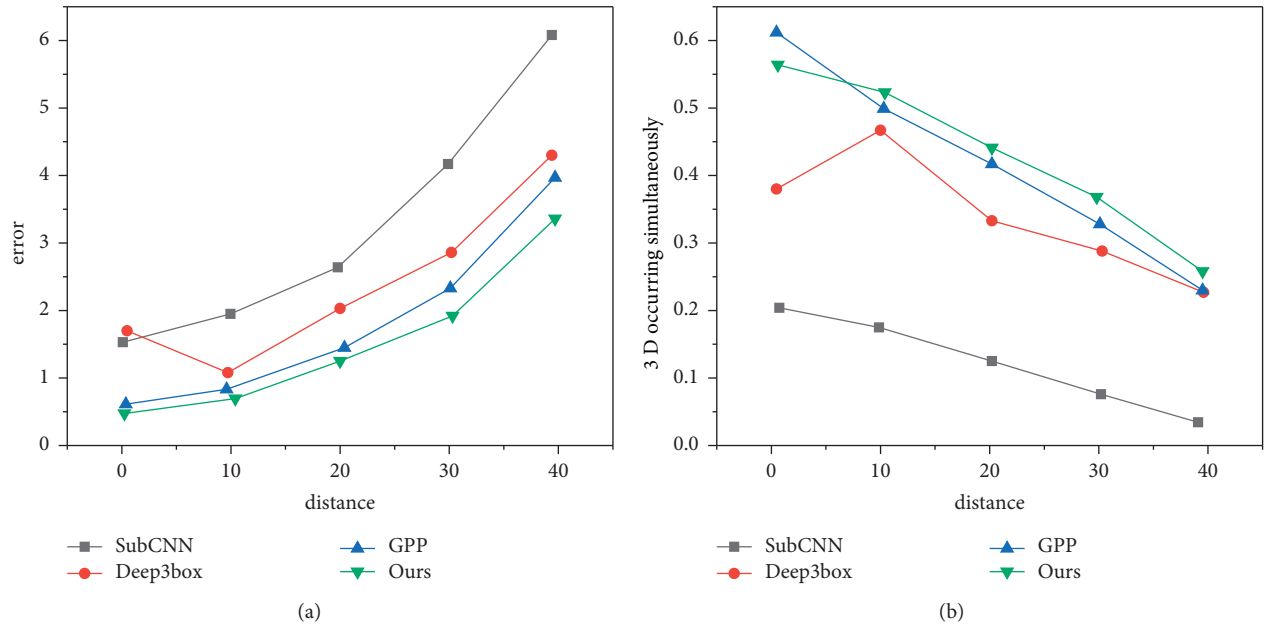

Figure 2: Comparison of 3D detection experiments on KITTI data set. (a) Experimental comparison of center point distance equalization error. (b) Experimental comparison of 3D cross-parallel ratio.

trajectory, 3D bounding frame size, and rotation angle under the surveillance video. Vehicle track is the most intuitive expression of vehicle movement behavior in traffic video surveillance. Trajectory data were used to extract different types of vehicle tracks for prior training, and trajectory point P was classified by K nearest neighbor algorithm in $1-n$ frames. Given the weight parameter w_t, the trajectory function $Traj_{obj}$ and trajectory evaluation function D_t of the 3D bounding box of the vehicle are shown in formulas (6) and (7):

$$\text{Traj} = \{p_1, p_2, \ldots, p_n\} = \{(x_1, y_1, t_1), (x_2, y_2, t_2), \ldots (x_n, y_n, t_n)\}, \tag{6}$$

$$D_t = w_t \cdot \text{Traj}_{\text{obj}}. \tag{7}$$

Vehicle collisions also bring about rapid changes in the corresponding 3D bounding frame. An appropriate range of changes is set for each vehicle to compare the 3D bounding frame in normal driving. The prior average dimensions of length, width, and height of the bounding frame are set as $\overline{s}$, and the variation range of reasonable size s_i is set as shown in the following formula:

$$0.8\overline{s} \le s_i \le 1.2\overline{s}. \tag{8}$$

The dimension variance $O_{3D}(W, H, L)$ of the vehicle 3D bounding frame is calculated; the weight parameter w_s is set, and the vehicle bounding box size evaluation function D_s can be obtained as shown in formulas (9) and (10).

$$O_{3D}(W, H, L) = \frac{\sum_{i=1}^{n} (\overline{s} - s_i)^2}{n}, \tag{9}$$

$$O_{3D} = w_s \cdot O_{3D}(W, H, L). \tag{10}$$

The variances of the three rotation angles $O_{3D}(\alpha, \beta, y)$ of the vehicle enclosure were calculated accordingly; set the weight parameter $w\theta$ to obtain the vehicle rotation angle evaluation function D_θ, as shown in formulas (11) and (12).

$$O_{3D}(\alpha, \beta, y) = \frac{\sum_{i=1}^{n} (\overline{\theta} - \theta_i)^2}{n}, \tag{11}$$

$$O_\theta = w_\theta \cdot O_{3D}(W, H, L). \tag{12}$$

A vehicle accident detected in a continuous frame is determined by the threshold parameter A_t, and the detection result function is set as follows: there was a traffic accident, $A > A_t$; no traffic accident occurred, otherwise [10].

3. Results and Analysis

3.1. Training Process. This experiment mainly tests the performance of vehicle detection and tracking and accident detection under the view of intersection monitoring. The data set used for the experimental test is from KITTI data set and intersection monitoring data set, covering different weather environment and traffic flow, from which 100 video sequences are extracted and annotated. Three-fifths of the total traffic video frames were used for training, and the rest were used as test frames. Each batch of 1000 images was trained by the stochastic gradient descent method, and their average loss function values were calculated to update the parameter values. The learning rate was set to 0.001 in the first 30,000 iterations and 0.0001 in the last 10,000 iterations.

3.2. Experimental Results. The experiment is compared with several advanced 3D attitude detectors. The vehicles were detected using the KITTI dataset official evaluation method

Table 1: Experimental comparison of tracking effect.

Algorithm	MOTA (%)	FP (%)	FN (%)
CEM	78.4	1.1	3.5
3D CMK	88.6	0.9	1.9
Methods	89.3	0.9	1.2

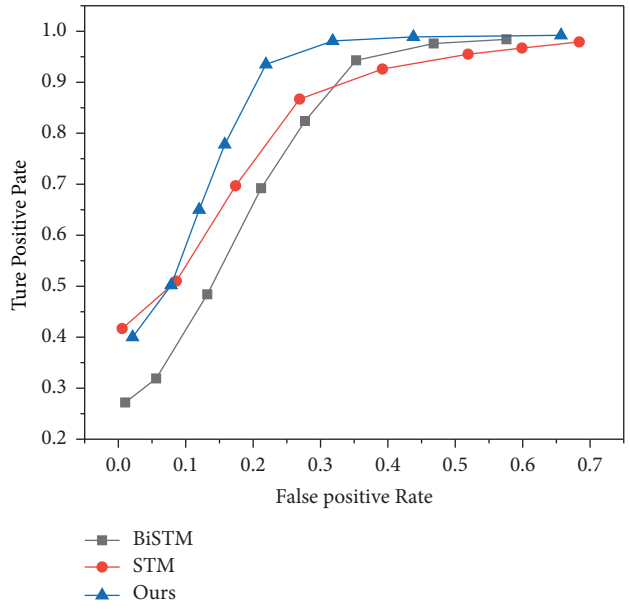

Figure 3: ROC curve of accident detection.

at three different levels of occlusion. In order to evaluate the accuracy of 3D object detection, three evaluation criteria are formulated according to the actual situation [11]. Firstly, the error rate between the real distance and the estimated distance between the center point of the object's 3D bounding frame and the camera is calculated. Secondly, the change trend of the intersection ratio between the detected 3D bounding frame and the real 3D bounding frame is compared with the increase in the real distance between the object and the camera. Figures 2(a) and 2(b) show the experimental comparison of the method with other methods under the first two evaluation criteria for 3D detection of the KITTI dataset. The final evaluation criterion is vehicle orientation comparison, which compares the mean directional similarity (AOS), mean accuracy (AP), and directional score (OS) by detecting three data sets with different degrees of occlusion. Compared with other three 3D detection methods, the comprehensive accuracy of the method is higher than that of other methods.

In 3D target tracking, scale change, direction change, and occlusion will affect tracking performance. The video sequences from 50 traffic surveillance angles were tracked and compared with other tracking methods. The average accuracy (MOTA), false positive (FP), and false negative (FN) of multitarget tracking were used to evaluate the robustness of the model. As shown in Table 1, the 3D target tracking trajectory in this paper is clear, and the vehicle attitude of 3D bounding box regression is relatively accurate [12].

For the evaluation method of traffic accident detection performance, the ROC curve was drawn by using the two evaluation indexes of true positive rate (TPR) and false positive rate (FPR). As shown in Figure 3, compared with BiSTM and STM, the proposed algorithm can identify traffic accidents on the road in a variety of environments with higher accuracy.

4. Conclusions

A vehicle accident detection method based on 3D object detection to generate 3D semantic bounding box is proposed; the method extracts the vehicle posture from the convolutional neural network and predicts the 3D bounding box of the vehicle. 3D object detection and 2D to 3D extended Kalman filter were combined to recover the 6 degrees of 3D vehicle attitude and tracking track in the video sequence, and the 3D semantic bounding frame was used to establish the evaluation index of accident detection. The experimental results show that each batch of 1000 images was trained by the stochastic gradient descent method, and their average loss function values were calculated to update the parameter values. The learning rate was set to 0.001 in the first 30,000 iterations and 0.0001 in the last 10,000 iterations. The MOTA of the CEM algorithm is 78.4%, FP is 1.1%, and FN is 3.5%, and the MOTA of the 3-DCMK algorithm is 88.6%, FP is 0.9%, and FN is 1.9%. The MOTA of this method is 89.3%, FP is 0.9%, and FN is 1.2%, and the 3D target semantic detection of vehicle accident video is stable.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] Y. Zhu, "Research on human vehicle accident reconstruction based on aeb system," *Open Access Library Journal*, vol. 05, no. 10, pp. 1–12, 2018.

[2] H. Woo, Y. Ji, H. Kono et al., "Lane-change detection based on vehicle-trajectory prediction," *IEEE Robotics & Automation Letters*, vol. 2, no. 2, pp. 1109–1116, 2017.

[3] Y. Chung and I. Chang, "How accurate is accident data in road safety research? an application of vehicle black box data regarding pedestrian-to-taxi accidents in korea," *Accident Analysis & Prevention*, vol. 84, pp. 1–8, 2015.

[4] Z. Q. Wei, X. W. Wang, D. N. Jia, and H. Liu, "An information collection and transmission strategy of vehicle state-aware system based on obd technology and android mobile terminals," *Applied Mechanics & Materials*, vol. 719-720, pp. 573–579, 2015.

[5] H. Wang, S. He, J. Yu, L. Wang, and T. Liu, "Research and implementation of vehicle target detection and information recognition technology based on ni myrio," *Sensors*, vol. 20, no. 6, p. 1765, 2020.

[6] B. Ma, H. G. Xu, Y. Chen, and M. Y. Lin, "Parametric research of tire mark intensity in accident scene based on vehicle-road system," *Zhongguo Gonglu Xuebao/China Journal of Highway and Transport*, vol. 31, no. 4, pp. 250–261, 2018.

[7] Z. Lv, Y. Zhou, Y. Li, and J. Mo, "Research on highway vehicle detection algorithm based on video image," *International Journal of Information and Communication Technology*, vol. 17, no. 1, p. 22, 2020.

[8] K. Al-Shara, "Automatic vehicle accident detection based on gsm system," *Iraqi Journal for Computers and Informatics*, vol. 43, no. 2, pp. 9–13, 2017.

[9] H. U. Lin, "Research on the effect of seat height on cyclists' dynamics response in vehicle-bicycle accident," *Journal of Mechanical Engineering*, vol. 54, no. 21, p. 81, 2018.

[10] H.-C. Son, D.-S. Kim, and S.-Y. Kim, "Vehicle-level traffic accident detection on vehicle-mounted camera based on cascade bi-lstm," *Journal of Advanced Information Technology and Convergence*, vol. 10, no. 2, pp. 167–175, 2020.

[11] M. Chen, L. Jin, Y. Jiang, L. Gao, F. Wang, and X. Xie, "Study on leading vehicle detection at night based on multisensor and image enhancement method," *Mathematical Problems in Engineering*, vol. 2016, Article ID 5810910, 13 pages, 2016.

[12] S. J. Davis and B. K. Barton, "Effects of secondary tasks on auditory detection and crossing thresholds in relation to approaching vehicle noises," *Accident Analysis & Prevention*, vol. 98, pp. 287–294, 2017.