2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)
September 12-14, 2018, Madrid, Spain

Deep Learning based Vulnerable Road User Detection and Collision Avoidance

Swadesh Kumar Maurya
School of Computer and Systems Sciences
Jawaharlal Nehru University
New Delhi, Delhi 110067
Email: swades89_scs@jnu.ac.in

Ayesha Choudhary
School of Computer and Systems Sciences
Jawaharlal Nehru University
New Delhi, Delhi 110067
Email: ayeshac@mail.jnu.ac.in

***Abstract*—In this paper, we propose a camera based novel, real-time framework for detection and tracking of vulnerable road users, such as pedestrians and cyclists. Our framework also gives a measure of the degree of vulnerability based on the direction of movement and distance from the vulnerable region. Pedestrians and cyclists are the most vulnerable road users and it is necessary to develop automated systems that can detect them and ensure their safety by alerting the driver. In our framework, we apply deep learning based method for 2D pose detection for detecting the pedestrians and cyclists in the view of the outside looking camera mounted on the dashboard of a vehicle. As the vehicle moves, the pedestrians and cyclists are detected and tracked across frames, their degree of vulnerability is measured and the driver is alerted in case of high vulnerability score. Experimental results show that the our framework is able to accurately detect vulnerable road users and measure their degree of vulnerability.**

***Index Terms*—road safety, intelligent vehicle, VRUs detection, driver assistance, collision avoidance, trajectory.**

I. Introduction

In this paper, we propose a novel, real-time framework for detection of vulnerable road users, such as pedestrians and cyclists, and prediction of their vulnerability of collision with the vehicle based on their direction of movement and distance from the vehicle. We assume that a camera is mounted on the dashboard of a vehicle that is looking outside and continuously observing the scene. As the frames are captured, they are processed for VRU detection and across time, the direction of movement and distance from the vehicle is computed to predict the possibility of collision with the pedestrian or cyclists.

Detection of vulnerable road users(VRUs), that is, pedestrians, cyclists, and motorcyclists, is an important task for their safety. According to the WHO report [1] globally, pedestrians constitute 22% of road injuries. Countries that have a middle income group, own 52% of the worlds vehicles, and these countries register an 80% of the worlds road traffic deaths. The fast rate of urbanization is causing rapid growth in urban areas and urban population, creating an ever increasing density of vehicles, pedestrians, cyclists and motorcyclists. Therefore, there is a growing need to develop real-time Advanced Driver Assistance Systems (ADAS) such that the driver can be alerted and accidents avoided based on prediction of the possibility of a VRU coming in front of the vehicle.

Therefore, there is a need to understand VRU movement, predict their intent and model their behavior. These are challenging problems because the VRUs may not be using specific lanes or may quickly change their direction of movement. Vehicles have various avenues to display the intent, such as braking light, turning indicators, etc., but pedestrians and cyclists do not have a commonly understood mode of sharing their intent, and so they may not be able to effectively communicate their intent. To address this problem and to ensure the safety of VRUs, we propose a novel method of detecting VRUs based on their movement and trajectory in the area under observation. We compute the degree of vulnerability and alert is displayed on the screen.

VRU detection and their prediction of vulnerability from a camera mounted on the vehicle is a challenging problem. The challenges occur due to illumination variation, environmental conditions such as rain,fog, etc., presence of buildings, trees and the shadows they cast, etc. However, it is also challenging because of the lack of depth information from a single camera. Moreover, the size of the VRU in the image may vary due to the variations in distance from the camera. Moreover, the VRU may be captured in different shapes due to the various movement patterns of the VRU. They may not be detected as VRUs if they are inseparable from the background. This could happen due to the VRU's clothing or carrying large objects with them. The Illumination may also play an important role because too much or low illumination may lead to poor performance of the assistance systems. Occlusion also is a big problem in the detection and tracking the trajectory of movement.

In our approach, the pedestrian or cyclist is detected in every frame using deep learning. Then, across frames, the pedestrian is tracked based on the current and previous detections. . The tracking point is a point on the detected body model which undergoes minimum variance. The record of the detected pedestrian is maintained in a sliding pattern so that it contains only the recent tracking points. The trajectory plotted for the movement of the pedestrian represents the most recent and relevant direction of movement. Based on the movement of direction and distance from the vehicle, the degree of vulnerability of the VRU is computed and the driver is alerted in case the vulnerability is high. Our approach is designed in such a way that no prior information is needed and is also independent from the environmental and traffic conditions.

978-1-5386-3543-8/18/$31.00 ©2018 IEEE

The paper is structured as follows: In Section II the related work is discussed. We discuss our proposed work in Section III and the experimental results in Section IV. Finally, we conclude in V.

II. Related Work

In recent years, there has been an increase in the research in the area of creating safety for vulnerable road users. Quintero et al. [3], perform path prediction of the pedestrian using a probabilistic approach by detecting different body parts and joints using stereo vision and Gaussian process dynamic model. In [4], they use the same 3D pose model and propose path prediction using the action classification. Goldhammer et al. [5] apply machine learning techniques to predict the future movement of VRUs based on their current trajectory at the intersection in an urban area. They use video-based motion classification and evaluates the physical state of the motion for starting and stopping.

Themann et al. [6], proposed a co-operative collision avoidance system which communicates the positions of the VRUs from vehicle to vehicle via an infrastructure level communication. They find that within 50 km/h velocity of the vehicle, the prediction gives a standard deviation of maximum 55 cm in the VRU's position. Bertozzi et al. [7] track the pedestrians using edge density and symmetry map on top of the Kalman filtering technique. In [8], [9], [10] various libraries are discussed to cover the critical aspects of VRU safety design. They focus on sensing and control systems and evaluate them on different traffic scenarios. Their model uses information from radar systems to estimate collision, and image based object model for classification to decide for autonomous braking to prevent collision with VRUs. In [11] they use Faster Region-based CNN(Faster-RCNN) architecture for simultaneously detecting and localizing any instances of cyclists in depth images using the down-sampled 3D LiDAR data.

Pedestrian detection from a moving vehicle is a challenging problem, and detecting the intent of the pedestrian based on its current position and movement pattern is highly dependent on the accurate detection and tracking of the VRU. In our proposed work, we perform VRU detection using the deep learning based system for 2D human pose detection as given in [2]. This method accurately detects multiple people in each frame. We track these detected VRUs across frames by clustering on their position and size. Using the past history of the track, we find the direction of movement of the detected VRUs and based on the proposed vulnerability measure, our proposed method decides which pedestrians and cyclists are vulnerable and which are safe.

III. Proposed Work

In our proposed framework, we assume that a camera is placed on the dashboard of a vehicle and continuously observing the outside environment. As the frames are captured, our system processes it to detect the pedestrians and cyclists using deep learning method in [2]. These VRUs are then tracked using clustering on their location and size. The tracks across time give the direction of movement of the pedestrians which along with the the distance of the VRU from the vehicle helps the system decide whether the pedestrian or cyclist is vulnerable or not.

A. VRU Detection

In our framework, we detect the VRUs using 2D human pose as given in [2]. This method robustly and accurately detects multiple pedestrians and cyclists in the frame at different scales. It uses a non-parametric approach of Part Affinity Fields(PAF) that associate the body part of a person in a multi-person image. PAF is the 2D vector that represents the position and orientation of the parts. The confidence map is generated from the annotated key-points, and in each confidence map, there is a single peak for each part present. If multiple people exist, then there will be a peak for each part for each person.

The confidence map uses non-maximum suppression to get different parts of the candidate without repetition. The optimal correspondence of parts to a k-dimensional matching is difficult, but with the greedy approach, the association of the part with the correct candidate gives better matches. This obtains a spanning tree with a minimum number of edges to get the human skeleton instead of using the complete graph and further divides the problem to the bipartite matching problem. In this way, the limb pair is matched independently with each part of the candidate and gives a skeleton of the expected person present in the image. The skeleton model has 18 key body points which is used to position the person's exact 2D pose. As each part is associated with a score value or confidence value, we also use the number of parts detected and average score value of parts to prune the valid persons in the image. This works very well in reducing invalid detection. The invalid detection occur mainly because of presence of noise or objects such as trees and traffic lights, which have a similar structure as a pedestrian. However, we further filter out invalid detections based on the number of connections between body parts and the height of the detected object as described in Section III-B. Figures 1 and 2 show the detection of pedestrians in our experiments.

The deep learning based 2D pose detection of the pedestrian and cyclist using [2] works well but it has not been trained specifically for road scene datasets, where various objects similar to structure of pedestrians also appear. Therefore, the 2D pose estimation to detect road users can be further improved by fine-tuning with transfer learning on road scene datasets, where the road users exist in different sizes, locations and perspectives.

B. Region division and Filtering

Once the skeleton is detected in the image, there exists a lot of cases of false detection. So we apply filtering by dividing the image into three regions vertically R_1, R_2, R_3, as shown in Figures 3 and 4. The region R_1 shows the uppermost region of the image which is farthest from the vehicle and therefore, the size of the pedestrians and cyclists are very small. They may go undetected in but since they are very far, their vulnerability

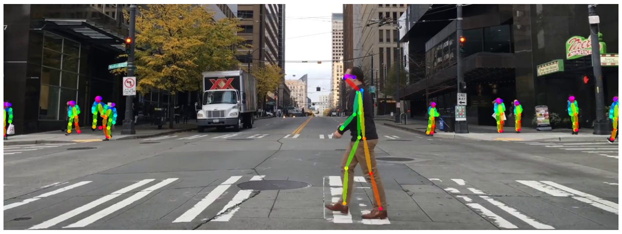

Fig. 1: An example of a detected pedestrian. *Images are best viewed in color.*

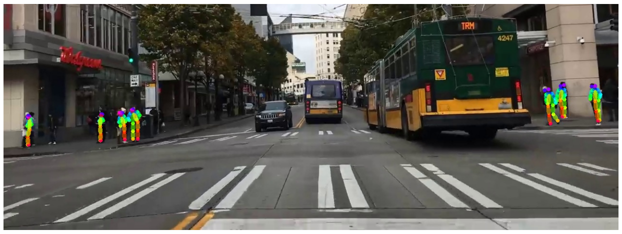

Fig. 2: An example of a detected pedestrians when there are many people together. *Images are best viewed in color.*

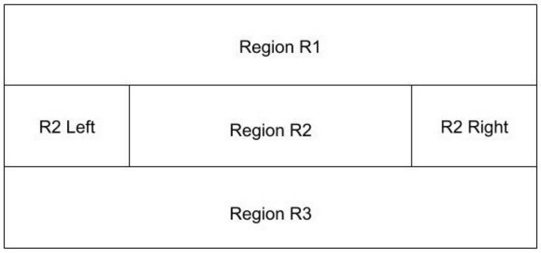

Fig. 3: Labels of the different regions in which we divide the image.

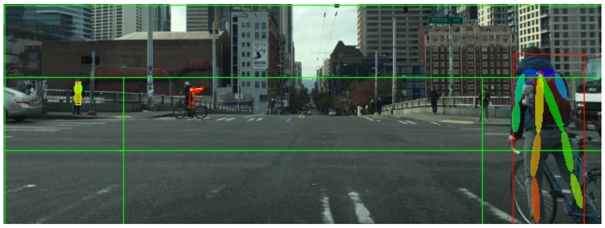

Fig. 4: Division of the image in the different regions. The topmost R_1 is farthest from the vehicle, while the middle of R_2 and R_3 are the vulnerable regions. Dividing the images into these regions, is necessary since we are working with a single camera and there is no depth information. Size of the vulnerable road user is dependent on the distance from the vehicle. These regions allow us to find the degree of vulnerability accurately without information of distance from the vehicle.

is very low and therefore, missed detections do not effect in this region. The region R_2 is the middle region of the scene and not too far from the vehicle. The people and cyclists in this region are important for as the vehicle moves, these people may become highly vulnerable, based on their pattern of movement. In this region the road scene can be further divided horizontally, the leftmost region shows the footpath on the left side and the rightmost shows the footpath on the right side. The middle region is the actual road area where the vulnerability is high and chances of collision is maximum. The third region R_3 is the nearest region of the road from the vehicle. Here, the people are very vulnerable, but when the people are so close to vehicles they are more cautious also. Moreover, they are large in size in the view and may cause occlusion, but they can be seen clearly by the driver.

We analyse each region R_1, R_2, R_3 and compute the average, minimum and maximum heights. For region R_2, the average range of height is evaluated and set for filtering the invalid detection of person skeletons. In this region, the middle portion shows the person located at a slightly far-off position from the vehicle so the range of height is usually medium. The region R_3 is the most vulnerable region in the whole image and height of the road users present here are comparatively larger because they are near to the vehicle, so there is a high degree of chance for collision, and a driver alert is a must for these road users.

We perform filtering at this stage to remove the detected pose skeleton that are invalid. In each frame, using region based approach where we analyze the height distribution, body part count and the associated part score, the invalid detections are pruned. Moreover, those detections that do not result in fitting of a skeleton are also treated as invalid and filtered out. The region based approach gives the flexibility to filter the invalid detection based on the criteria of the specific region. This pruning process removes most of the invalid skeleton detection from the previous stage of 2D pose estimation.

C. *Tracking detected VRUs*

In our framework, the road users get detected from the 2D pose estimation model. Then, in the second phase, filtering is performed to remove the invalid detections by applying the height constraint for each region as well as removing the points on which a skeleton is not fitted. Next we perform clustering on these detected skeletons size and position to track them across the frames on-the-fly. If the person is again detected in the next frame, then on the basis of the previous position and height, the track of the closest person is mapped with the newly detected person. We consider the neck point in our experiment for the purpose of tracking, because in the analysis of the skeleton model we found that, as compared to the other skeleton key points, the neck point is detected with a good score. Also, the movement or variation in position of the neck is very low when compared to the other points in the skeleton.

During this mapping process, if a person does not get detected within a specified number of frames, the record for this person gets removed assuming that the person is not present in the frame or occlusion has occurred. The person is considered as a newly detected person if we cannot find a person who is closest to him from a tracking point perspective in that specific

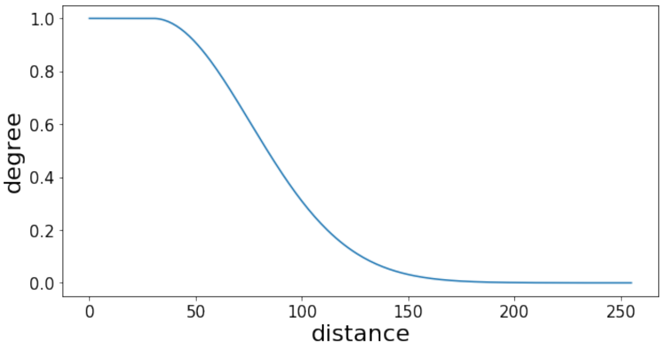

Fig. 5: Distance vs Degree of Vulnerability

area. The tracking history is maintained for a specific period and forgotten once the cluster for that object is not detected for a fixed time interval. These trajectory points are used to find the best fitting line to find the direction of movement of the VRU. Then, depending on the direction of movement, we compute the vulnerability score to decide the degree of vulnerability of the VRU. In region R_1, the vulnerability is taken as 0, and in the region R_3, the vulnerability is 1. In the middle of region R_2, the degree of vulnerability is also 1 since it is the road in front, however, in the left and right regions of R_2 the degree of vulnerability of the person is calculated based on the distance of the person from the left and right boundaries of the middle region, respectively.

D. Degree of vulnerability

The degree of vulnerability for the region R_2 is calculated using the following expression :

$$deg(x) = \begin{cases} 1, & \text{if } x < x_1 \\ e^{(-c(\frac{x-x_1}{x_2-x_1})^2)}, & \text{if } x_1 \leq x \leq x_2 \\ 0, & \text{if} x > x_2 \end{cases}$$

In this expression, x represents the distance from the boundary of the middle region, x_1, x_2 are the thresholds for the distances, where x_1 represents the threshold below which, the degree of vulnerability remains one and x_2 is the distance threshold above which, the degree becomes zero. Here, c is the normalizing constant. Figure 5 shows the plot of distance from the middle region R_2 boundary versus degree of vulnerability of the road user.

IV. Results

In this section, we discuss the experimental results of our approach on the road scenes captured by the camera mounted on a vehicle. The videos for are experiments are taken from youtube, and are well suited for our problem. In our experiments, the road user is detected using deep learning based method [2] by the 2D pose estimation showing the human body parts in different colors using ellipses fitted on every detected part. The detections on each frame can result in some invalid detections(shown in Figure 6), which are removed during the filtration step (shown in Figure 7).

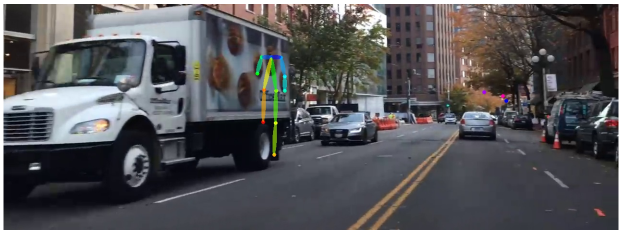

Fig. 6: Instance of an invalid detection. *Images are best viewed in color.*

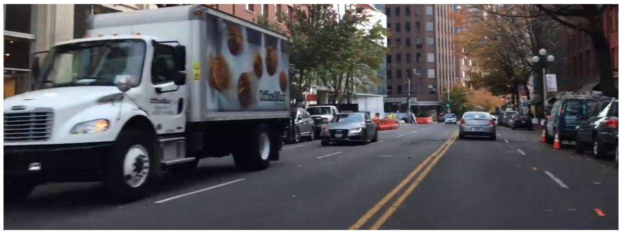

Fig. 7: The same scene in Figure 6, after filtering. *Images are best viewed in color.*

We perform experiments on two different videos taken from the youtube [12], [13]. We show the detected road users in bounded boxes of different colors to show their state of vulnerability. There are three states of vulnerability for the road users, red color box indicating highly vulnerable users, with degree of vulnerabiltiy equal to 1, the yellow color bounding boxes indicate medium vulnerability, while the green colored bounding box indicates that the road user is safe and has degree of vulnerability equal to0. Figure 8, 9, 10 and 11 show our experimental results for video [12].

The Figures 8, 9, 10 and 11 are instances of the same video [12] across time. It can be seen that despite illumination variation and various dynamic occlusions in the scene, the pedestrians and cyclists are correctly detected and tracked and their vulnerability score is correctly computed by our proposed framework. Similarly, Figures 12, 13, 14, 15 and 16 show results of our proposed framework on voideo [13]

It can be seen from the result images that our framework

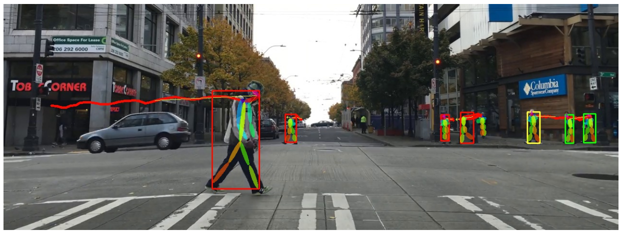

Fig. 8: It can be seen that even though there are a large number of people in the scene, degree of vulnerability of the pedestrians has been accurately assigned based on the region and direction of movement. Red box shows highly vulnerable, yellow box shows predictability of vulnerability in the future and green box shows that the person is not vulnerable.

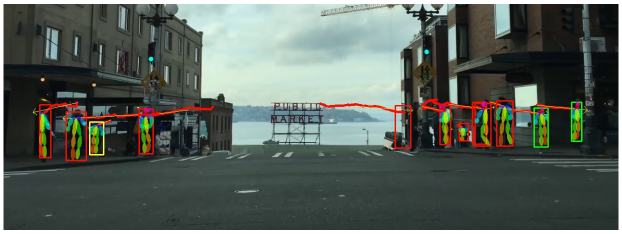

Fig. 9: Detection, tracking and degree of vulnerability of the pedestrians and cyclists. Red box shows highly vulnerable, yellow box shows predictability of vulnerability in the future and green box shows that the person is not vulnerable.

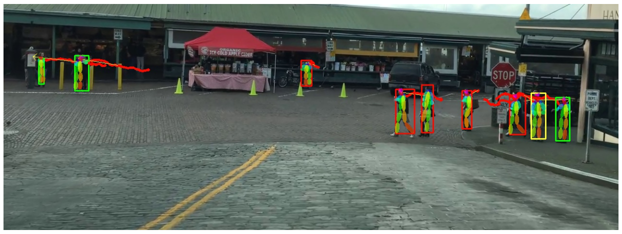

Fig. 10: Inspite of illumination variations as the vehicle moves in the scene, accurate detection, tracking and coputation of degree of vulnerability of the pedestrians and cyclists is carried out through our proposed framework. Red box shows highly vulnerable, yellow box shows predictability of vulnerability in the future and green box shows that the person is not vulnerable.

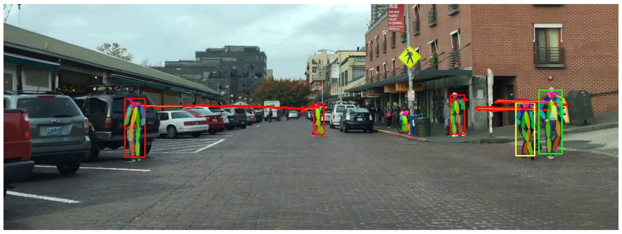

Fig. 11: Accurate detection, tracking and degree of vulnerability of the VRUs from online analysis of video [12]. Red box shows highly vulnerable, yellow box shows predictability of vulnerability in the future and green box shows that the person is not vulnerable.

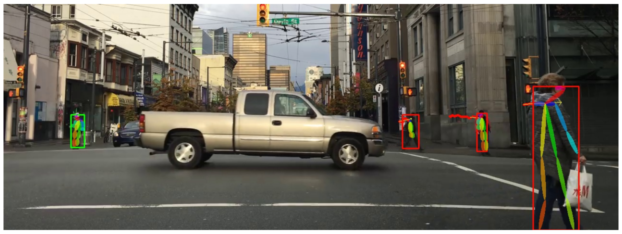

Fig. 12: A frame from video [13], showing correct detection of VRUs and their vulnerability score by color of the bounding box. Red box shows highly vulnerable, yellow box shows predictability of vulnerability in the future and green box shows that the person is not vulnerable.

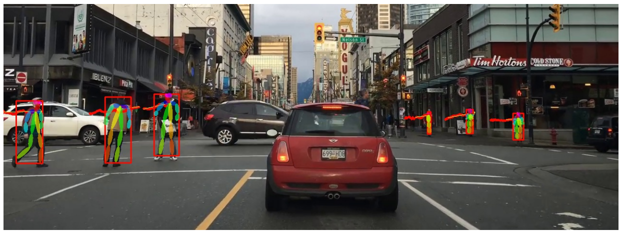

Fig. 13: Another frame from online analysis of video [13]. Red box shows highly vulnerable, yellow box shows predictability of vulnerability in the future and green box shows that the person is not vulnerable.

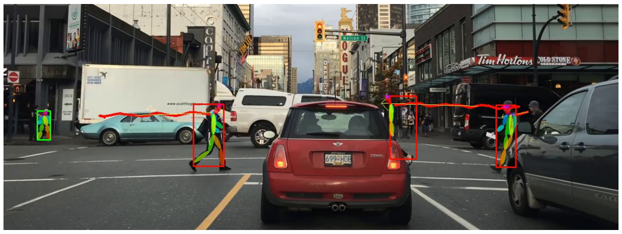

Fig. 14: In spite of occlusion and a large number of vehicles in the scene, the VRUs are correctly detected, tracked and their degree of vulnerability is correctly shown by the color of the bounding box. Red box shows highly vulnerable, yellow box shows predictability of vulnerability in the future and green box shows that the person is not vulnerable.

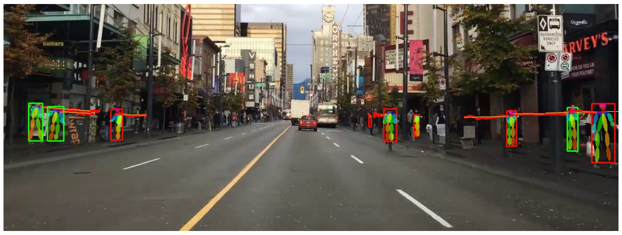

Fig. 15: Pedestrians in region R_1 do not get detected, however, they are very far from the vehicle and therefore, their degree of vulnerability is 0. The objects of interest that are closer to the vehicle are correctly detected and their vulnerability is correctly computed. Red box shows highly vulnerable, yellow box shows predictability of vulnerability in the future and green box shows that the person is not vulnerable.

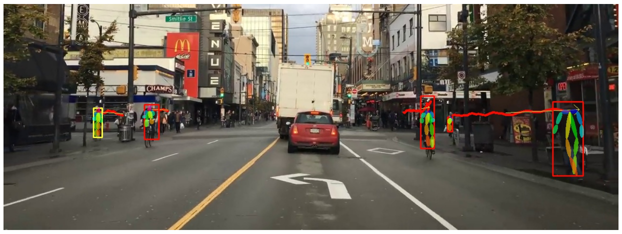

Fig. 16: An instance showing accurate detection, tracking and degree of vulnerability of the pedestrians and cyclists. Red box shows highly vulnerable, yellow box shows predictability of vulnerability in the future and green box shows that the person is not vulnerable.

TABLE I: Confusion matrix showing the actual vulnerability of the VRUs vs the computed degree of vulnerability by our proposed framework.

		Predicted vulnerability	
		Vulnerable	Not Vulnerable
Actual	Vulnerable	242	32
	Not Vulnerable	30	101

works well inspite of crowded scenes and large illumination variations. The combined road user detection rate for both the videos is 91.66%. Most of the pedestrians and cyclists that do not get detected are from the region R_1 which is very far from the vehicle, so the objects of interest are very small in size. This is also not a major issue because their vulnerability score is 0.

Table I shows the confusion matrix for the vulnerability of VRUs based on the actual vulnerability and the computed degree of vulnerability.

V. Conclusion

We have proposed a novel, real-time framework for detection, tracking and measuring the degree of vulnerability of the pedestrians and cyclists on road. Our framework requires that an outside looking camera is mounted on the dashboard of the vehicle and as a frame is captured it is processed to discover VRUs in its view. In case, it detects pedestrians and cyclists in its view that are measured as highly vulnerable, it alerts the driver. Our framework is also capable of predicting the vulnerability of people that are currently far from the vehicle but may be soon very vulnerable because of their direction of movement. Our work aims to create a driver assistance system for increasing the safety of the vulnerable road users. We have used deep learning based method for accurate detection of VRUs and unsupervised learning for tracking them. Moreover, experimental results show that our framework is capable of correctly measuring the degree of vulnerability of the pedestrians and cyclists in the view of the camera.

References

[1] ”Pedestrian Safety: A Road Safety Manual for Decision-Makers and Practitioners.”, World Health Organization (2013).

[2] Cao, Z., Simon, T., Wei, S. E., and Sheikh, Y., Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, No. 2, p.7, 2017 .

[3] Quintero, R., Almeida, J., Llorca, D. F., and Sotelo, M. A., Pedestrian path prediction using body language traits. In Proceedings of IEEE Intelligent Vehicles Symposium , pp. 317–323, 2014 .

[4] Quintero, R., Parra, I., Llorca, D. F., and Sotelo, M. A., Pedestrian path prediction based on body language and action classification. In IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), pp. 679–684, 2014.

[5] Goldhammer, M., Kohler, S., Zernetsch, S., Doll, K., Sick, B., and Dietmayer, K., Intentions of Vulnerable Road Users-Detection and Forecasting by Means of Machine Learning. arXiv preprint arXiv:1803.03577, 2018 .

[6] Themann, P., Kotte, J., Raudszus, D., and Eckstein, L., Impact of positioning uncertainty of vulnerable road users on risk minimization in collision avoidance systems. In IEEE Intelligent Vehicles Symposium (IV), pp. 1201–1206, 2015.

[7] Bertozzi, M., Broggi, A., Fascioli, A., Tibaldi, A., Chapuis, R., and Chausse, F., Pedestrian localization and tracking system with Kalman filtering. In Intelligent Vehicles Symposium, 2004 IEEE pp. 584–589, 2004.

[8] Garate, V. R., Bours, R., and Kietlinski, K., Numerical modeling of ADA system for vulnerable road users protection based on radar and vision sensing. In IEEE Intelligent Vehicles Symposium (IV), pp. 1150–1155, 2012 .

[9] Premebida, C., Ludwig, O., and Nunes, U., Exploiting lidar-based features on pedestrian detection in urban scenarios. In 12^{th} International IEEE Conference on Intelligent Transportation Systems, pp. 1–6, 2009.

[10] Alemneh, E., Senouci, S. M., and Brunet, P, PV-Alert: A fog-based architecture for safeguarding vulnerable road users. In Global Information Infrastructure and Networking Symposium (GIIS), pp. 9–15, 2017.

[11] Saleh, K., Hossny, M., Hossny, A., and Nahavandi, S., Cyclist detection in LIDAR scans using faster R-CNN and synthetic depth images. In IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 1–6, 2017.

[12] https://www.youtube.com/watch?v=rjVD1-1mbuE

[13] https://www.youtube.com/watch?v=6tyFAtgy4JA