2021 5th International Conference on Pattern Recogniti[1]on and Image Analysis (IPRIA)
Department of Electrical and Computer Engineering, University of Kashan, Iran, 28 & 29 April 2021

Prediction and Detection of car accidents in video by deep learning

Zahra Rezaei [1*], Hossein Ebrahimpour-Komleh [2*]

[1,2] *Department of Computer and Electrical Engineering University of Kashan, Kashan, Iran*
[1] *Statistics and Information Technology Research Institute, Judiciary Research Institute*
[*]z.rezaei2010@gmail.com
[*]ebrahimpour@kashanu.ac.ir

2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA) | 978-1-6654-2659-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IPRIA53572.2021.9483506

Abstract: Road accidents are considered as one of the leading causes of human fatality in recent decade. Using machine vision in automobile sensors and through warning even a second before accidents occur, a human disaster can be prevented. In order to predict accident at the stage, it is needed to learn temporal and spatial characteristics. In this paper, a primary real-time autonomous accident detection system has been proposed which is established on object detection deep learning in video prediction techniques. The events in video will be the accidents which are predicted several frames before the event. In this technique, 5000 accident images were labeled, and the model was classified using all the existing classes. Then, the special events were detected using object detection technique. The proposed model shows an accuracy of 92.98 and a speed of 200 images per second in video. The model can predict the event 0.92 second before accident. This model can be used in inter-city cameras with immediate analysis of meta data, and it has been piloted.

1. Introduction

Nowadays, driverless car is one of the most comprehensive technologies which has been piloted since 2009, and such cars have recently driven in crowded places and in real situation without any accidents. These cars have vision sensors which deliver images of the road and objects to a computer, and decide how to react based on the sensors; however, there are some challenges. First, the opposite cars are driven by human, and they are subject to any unpredictable behaviors of the drivers who may not observe traffic rules. Accordingly, the goal of the present paper is to predict event accidents to prevent road crashes using deep learning techniques. Second, labelling cost of each event in video is very high, and moreover, events which rarely occur are challenging in online video. Anomalies and events are usually context-dependent. For example, running in a restaurant can be an event while running in a park is considered a typical behaviour. Moreover, defining anomaly may be vague, and this term may have opaque definition. Someone may believe that walking on subway platform is a normal behaviour, while this may be considered an abnormal behavior from another person's point of view due to its uncertainty. These challenges make it difficult for automatic learning methods to identify video models which make abnormality in real world. The videos of accidents available in you tube was used to test the model.

According to the figures released by the World Health Organization (WHO) (2018) [1] [2]. The Sustainable Development proposed for 2030 agenda, has set a goal of reducing the number of deaths and injuries from road traffic crashes in half by 2020. If no sustained action is taken, it is predicted the road traffic crashes become the seventh main cause of death by 2030.Yearly, more than 1.25 million people lose their lives in road traffic crashes. Cost of keeping the injured, losing productivity, losing time to take care of the injured are the sources of the losses. Most countries spend 3% of their gross domestic product on road traffic crashes. Predicting accidents is a challenging issue as the events leading to accidents are diverse, and sometimes the accident happens out of a sudden which has not occurred earlier, and this is effective in prediction. When accidents are detected immediately, some lives may be saved, roads may be open more quickly, the time and resource waste may decrease and efficiency may increase. Damaged car detection is categorized in the class of object detection, and it has not been achieved yet.

Deriving temporal and spatial features is of great importance, and a technique which can learn features of various scenes is needed. 2000 videos were downloaded from you tube from which 2500 images of accidents and 2500 images of no-accident scenes were identified. Then, various conditions of

[1] 978-1-6654-2659-6/21/$31.00 ©2021 IEEE

accident scenes were analysed. The videos were truncated and from each video 20 seconds before and after the accident were classified. The images from the videos were used to train the model. Totally, 1800 videos were used for the training process and another 200 were used for the test process. Each accident clip included 24 frame per second, and the images were first classified into two classes of accident images and normal images. Then, the accident images were classified into car-car accident images and car-human accident images. All accident images were compared with images similar to the accident image in which no accident had occurred like Scissor door cars.

The proposed deep learning model can predict the accident 0.96 second in advance (recall: 0.9238 and Precision: 0.9509). In summary, the proposed method was tested in the following situations and it is useful in these cases: accident detection in crowded areas, accident prediction in crowded and less crowded scenes and accident prediction in auto-controlled cars. The rest of the present paper is classified in the following manner. Section II is devoted to review the existing literature on video accident detection, and section III describes the proposed detection technique. In section IV, the results obtained from employing the proposed technique along with data analysis are presented, and finally, a conclusion is presented.

1.1 Related Work

A number of studies have so far been carried out on accident and transportation systems. Most of them, however, have studied infrastructure development, improvements on physical infrastructures, etc [3]. It is worth noting that some studies have focused on some specific intelligent transportation system parts like the topics of traffic flow estimation, route optimization, accident prevention, event detection, etc. Such studies are mostly what this paper concern.

Some studies have been proposed to predict events and to classify future events through analysing the current situation. As a case in the point, feature matching techniques were proposed by Ryoo [4] for recognition of unfinished activities. In addition, Hoai and la Torre[5] present a max-margin-based classifier to anticipate delicate facial expressions before they happen. In another study, Lan et al [6] presented a new technique to anticipate the peoples' future actions in unconstrained in-the-wild footages. Considering the fact that accident can be considered as a special event, our technique deals with event anticipation. Furthermore, as our dashcam videos are recorded by different moving cameras observing static (e.g., building, road signs/billboards, etc.) and moving vehicles (e.g., motorbikes, cars, etc.), they are very challenging. Consequently, a dynamic-spatial-attention mechanism for discovering delicate cues for anticipating accidents was proposed.

Anticipation has been suggested for tasks other than event prediction. For example, Ryoo [7] present a method to estimate human path in surrounding physical environment (e.g., road, pavement, etc.). The paper also revealed that the estimated path can be used to enhance the accuracy of object tracking. Yuen and Torralba [8] propose a technique to forecast motion from motionless images. Wang et al [9] present a modern variable model for deriving unknown human intentions Koppula and Saxena [10] worked on anticipating future activities from RGB-D data. This method was employed in real robotic systems to help humans in daily activities. Furthermore, Kopppula et al [11] ; Mainprice and Berenson [12] presented a model to predict human activities for increasing human-robot association. There are also many works for predicting drivers' goal in the intelligent vehicle community. Ryoo [4] Ma et al [13]; Ravindran et al [14] have used vehicle routes to anticipate the intention of lane change or turn manoeuvre. Manoeuvre anticipation through sensory-fusion from multiple cameras, GPS, and vehicle dynamics has been addressed in a few works.

The proposed method includes object detection deep learning with spatial and temporal attention mechanism to concentrate on object-specific cues at each frame dynamically in order to deal with the challenges in accident prediction. In general, all earlier methods focused on predicting specific-manoeuvres like lane change or turn. In contrast, our purpose is predicting different accidents observed in naturally captured YouTube videos.

2. Proposed Method

New tends in object detection include two main sections, namely, a feature extractor and a feature classifier similar to traditional object detectors. The deeper and wider convolutional architectures are used as the feature extractor at present. Detection of events like accidents is classified into the algorithm of object detection, in a way that our model, at first learns all normal and accident conditions in the training phase, and accident class in this stage is categorized at class1.

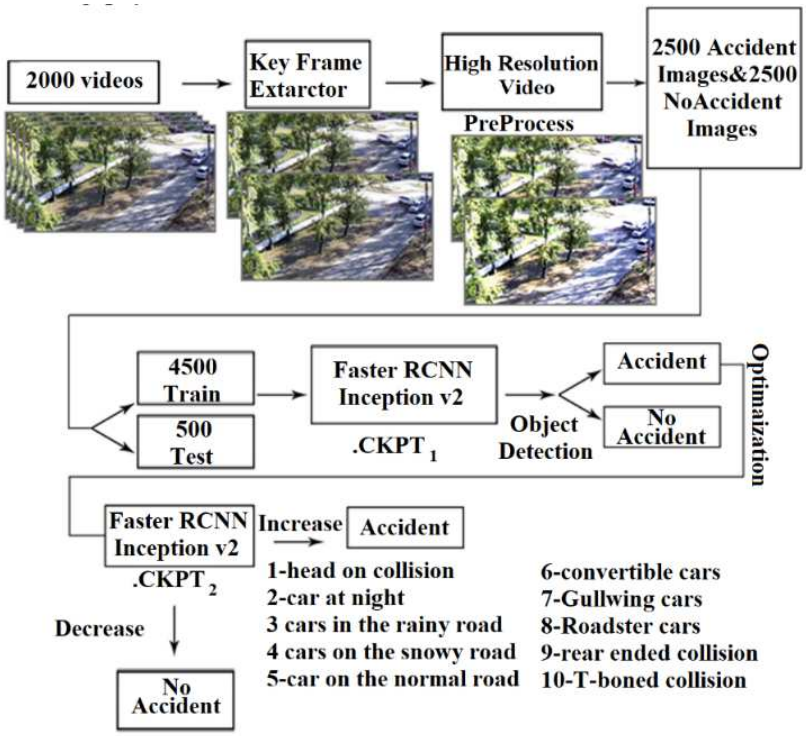


Fig. 1 The flow chart of the proposed method

In the proposed method 2000 short video clips from YouTube were used for training and testing, the videos were short and at most they were 20 seconds. 5000 images were taken from videos. Key Frame Extraction using Differential Evolution was employed to extract 5000 images. Average Entropy Difference and Average Euclidean Distance as the fitness function and a threshold level equation was conducted as frame selection. Then, in order to improve some frames which were not qualified for training, pre-processing was conducted on images and the quality of images was improved.

One of the classical problems in computer vision is super-resolution (SR), which intends to recover high-resolution images (or videos) from a low-resolution one. The pre-processing part was used only for challenging frames and 90% of the images did not need resolution improvement, then an expert classified the images into two classes of accident and non-accident images, and prepared 5000 images to be created by the model, and chose 4500 images to be used for training. Faster RCNN model was used as Meta-architecture. The findings reveal that employing fewer proposals Faster R-CNN by Ren et al [15] can enhance the speed significantly, and can compare it with SSD by Liu et al [16] and RFCN by Dai et al [17] without a big change of accuracy.

In the Faster R-CNN setting, detection happens in two stages (Figure 2). In the first stage, called the region proposal network (RPN), images are processed by a feature extractor (e.g., VGG-16), and features at some selected intermediate level (e.g., "conv5") are used to predict box classifier. The loss function for this first stage takes the form of Equation 1 using a grid of anchors tiled in spatial, scale and aspect ratio.

The best matching ground truth box b (if one exists) was first found for each anchor a. When such a match can be detected, we call it a "positive anchor", and assign it (1) a class label $y_a \in \{1,...,k\}$ and (2) a vector encoding of box b with respect to anchor a (called the box encoding $\Phi(b_a:a)$). If

no match is found, a is called a "negative anchor" and the class label was set to be ya = 0. If for the anchor a box encoding floc(τ;a,θ)and corresponding class fcls(τ;a,θ) are anticipated where τ is the image and θ the model parameters, then the loss for a is measured as a weighted sum of a location-based loss and a classification loss:

$$L(a, \tau, \theta) = \alpha.1[a \text{ is positive}].l_{loc}(\Phi(ba; a)\text{-}f_{loc}(\tau ; a; \theta))+\beta.l_{cls}(y_a, f_{cls}(\tau ; a; \theta))$$

where α; β are weights balancing localization and classification losses. In order to train the model, Equation 1 is averaged over anchors and is minimized with respect to parameters θ.

In the second stage, these (typically 300) box proposals are used to crop features from the same intermediate feature map which are subsequently fed to the remainder of the feature extractor (e.g., "fc6" followed by "fc7") in order to predict a class and class-specific box refinement for each proposal. The loss function for this second stage box classifier also takes the form of Equation 1 using the proposals generated from the RPN as anchors. Notably, one does not crop proposals directly from the image and re-run crops through the feature extractor, which would be duplicated computation. However, there is part of the computation that must be run once per region, and thus the running temporal depends on the number of regions proposed by the RPN.

After completion of the training model, the accident images, due to high similarity to non-accident images were

entered to the second model. In this model labelling was done based on accident from front, accident in rain, accident in snow, accident in fog, etc. Therefore, after the model was made in the second stage, due to customization, the false positive of overlapping images which were considered as accident images, was removed.

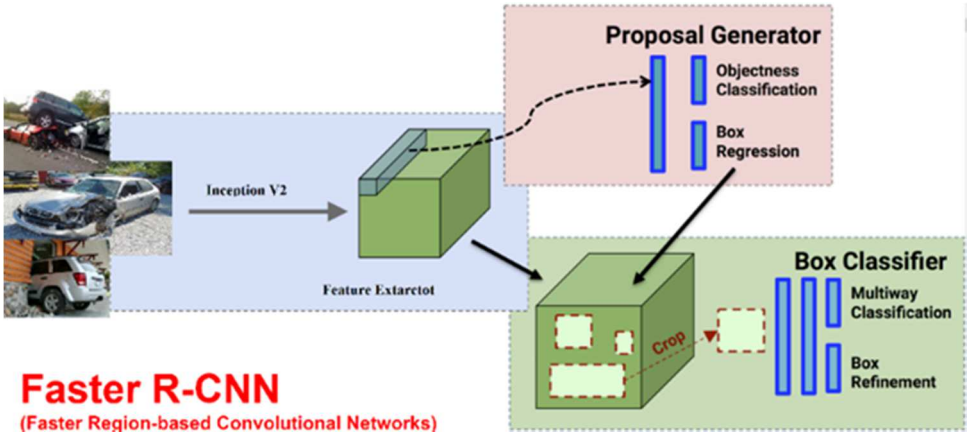


Fig. 2 The general architectural design for Faster R-CC [15]

3. Details of the proposed method

4. 2.1.1 Data collection

The data were extracted in the following situations: rainy, snowy, foggy weather, day, night, accident from front, accident from side, accident with passer-by, accident of two cars, and accident with street sign. The normal image of car was contrasted with accident image of cars. In addition, images of special cars like gullwing door cars, roadster car and convertible cars, due to lack of similarity were contrasted with normal and accident classes. In general, there are two classes with the aforementioned formats.

2.1.2. Extract key frames to model training

Video surveillance to identify key frames by humans is difficult and tedious. The use of intelligent methods in basic video analysis is effective in order to extract frames. Key frame extraction is an issue in video processing . These methods is based on four general algorithms:

1- Shot Edge Analysis [18] 2- Computational Mean Algorithm [19] 3- Histogram Mean Algorithm 4- Video Content Matching [20] . After several experiments, Video Content Matching method was used as a keyframe extractor .

2.1.3 Challenges of this method

High volume data like videos cannot be fed into a classifier directly: they contain much surplus data, and lead to more complicated calculations. Therefore, video data must be shown in a way so as to be able to be processed. This is challenging due to the following reasons:

• Changes in act models can be diverse in a class (for example, various road situation, the weather condition, etc.,) or classification of the event type (e.g. normal or accident). Variety in a class is due to style and appearance. The variety of data in a class is due to style and appearance. In order to register changes in car movement, car-human accident, there should be different objects (car, motorcycle, animal) in the scene.

• Environmental changes and noise: Real world images have a lot of noise, and they may be different due to changes in light and dynamic of background. The features must be able to control the environmental problems to work in high noise situations.

Considering hard ware costs and processing time of video, at first, pre-processing of images was done inside object detection module, then feature learning happened in the module stage, then the most check-pointed one in 5000 images was put in video code for momentum detection and anticipation. In order to show the video output, close key frames to the accident scene were chosen. Being able to anticipate abnormalities in real time is of great value. If the abnormality is detection and anticipation of accidents, even a few seconds before the event, it can reduce financial and economic loss considerably. Accordingly, attempts have been made for automatic substitution of accident detection with manual detection. Although so important, exact detection of abnormalities can be challenging. Processing of such systems usually includes several pre-processing, detection and description of features. Sequence model or anomaly identification based on specific criterion or threshold in deep learning method, occur inside the architecture of the model; therefore, the accuracy and speed of anomaly detection will be higher. Much research has been conducted in the area of accurate detection of anomaly in videos with acceptable wrong caution rate; however, this issue is challenging because of great changes in the environment and motion of human and objects, and temporal-spatial complexities due to vast dimensions of video data.

2. Discussion

2.1. Analysis of the method

According to the definition, when the probability of a class is more than 50%, the sample is allocated to the class: class I (the green box shows accident), class II (the blue box shows normal image.

The following images are the extracted frames of three videos of you tube in which accident is anticipated 18 frames before the accident. The first video (crowded images and two scenes one resulting in accident and the other one a normal one) the first anticipation occurred 27 frames before accident. As shown, in images 1 to 4, the images of cars which were covered by other cars were identified as accidents, while in the second stage of training in which the images were entered in the second model, false positive images were identified, and they were changed from accident images to non-accident images.

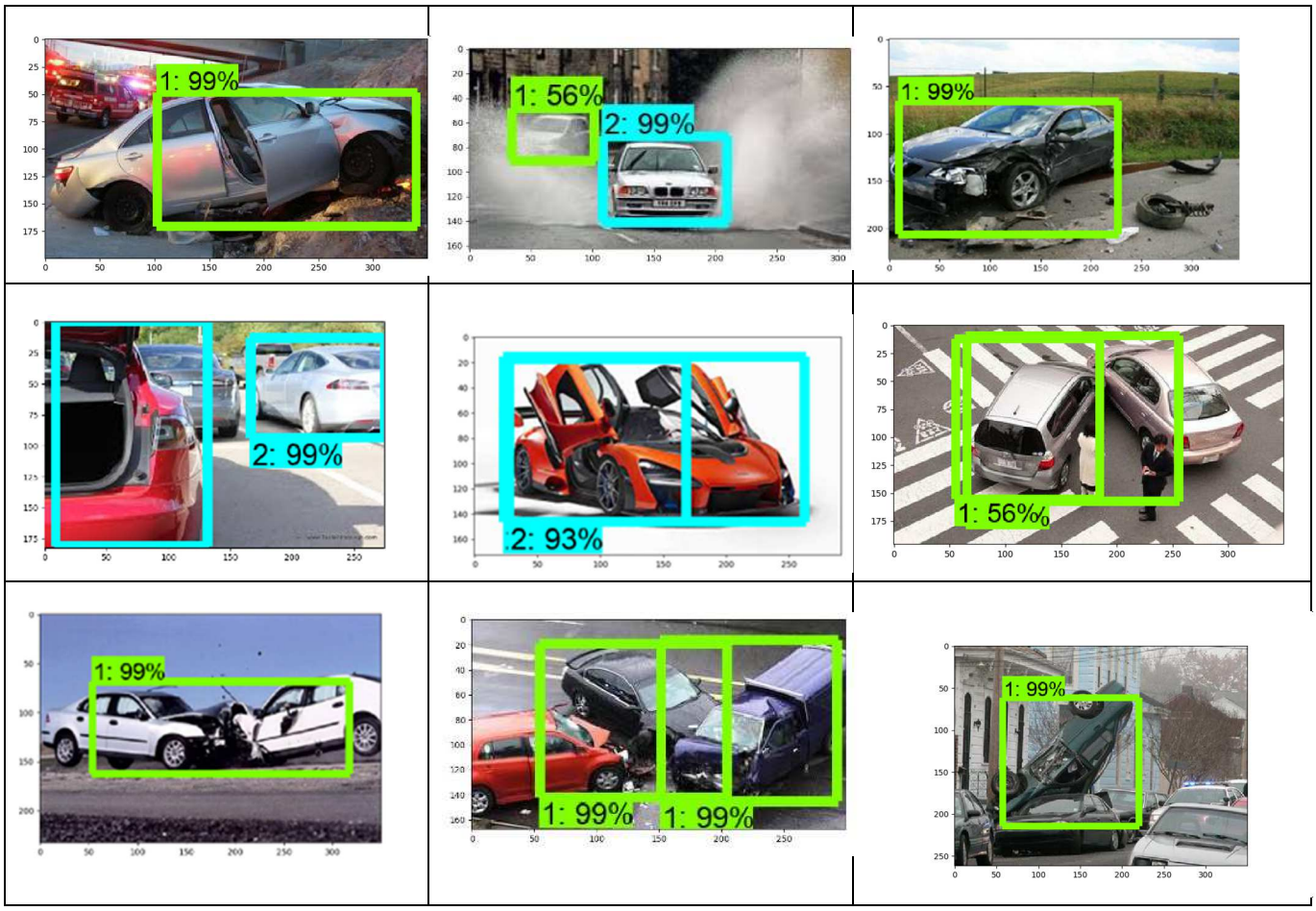
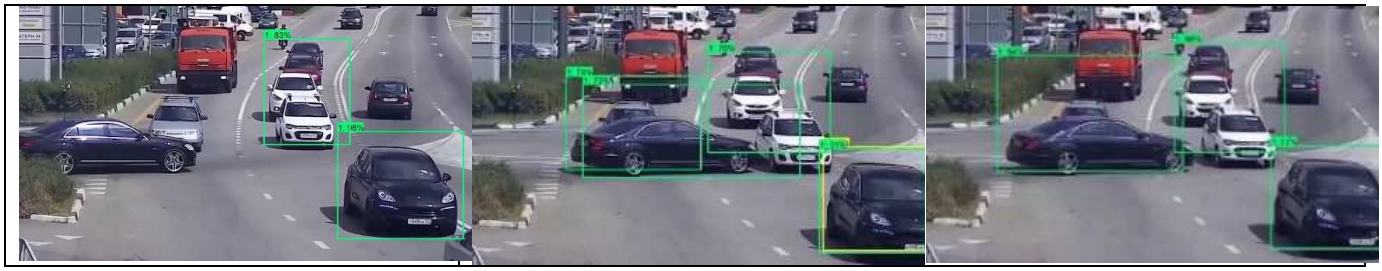


Fig. 3 The sample of extracted images from YouTube videos used for testing the method



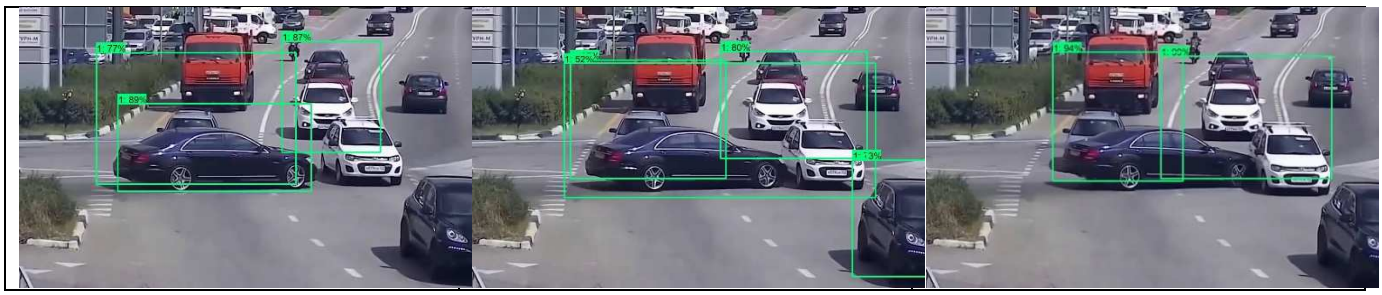


Fig.4 Crowded images and two images one resulting in accident and the other one normal, the first anticipation 72 frames before accident

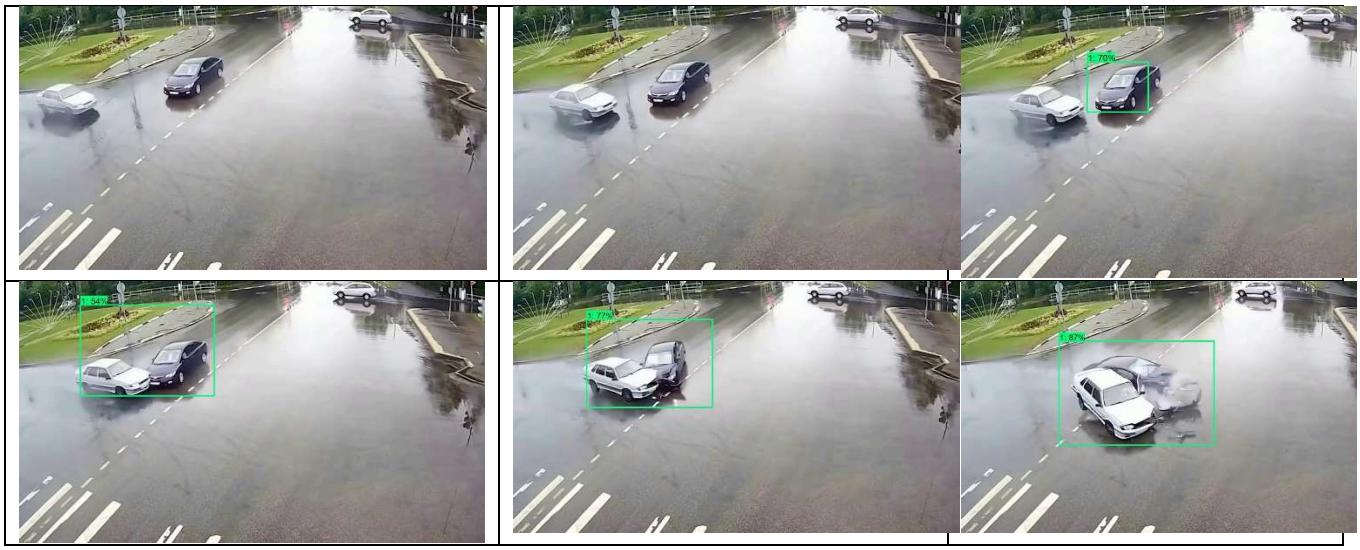


Fig. 5 Images in rainy weather and not-crowded situation, side to side image, the first anticipation 21 frames before accident

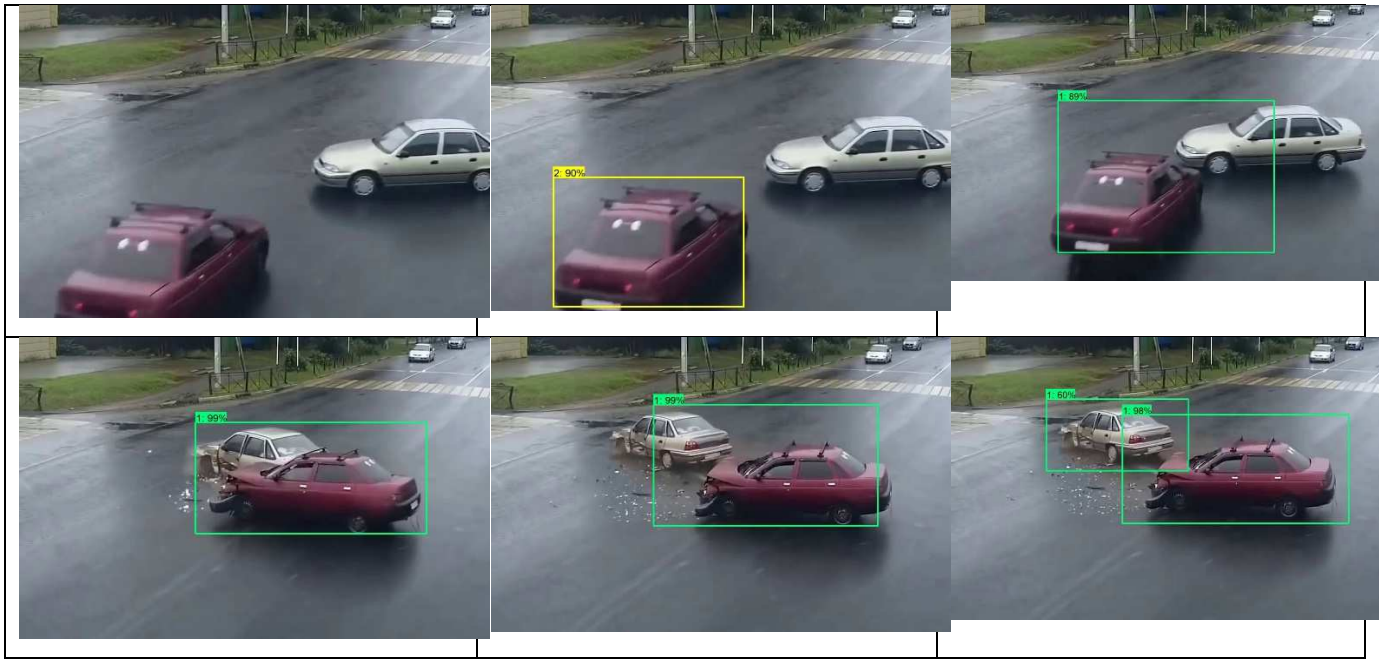


Fig.6 Accident from front, the first anticipation 18 frames before accident

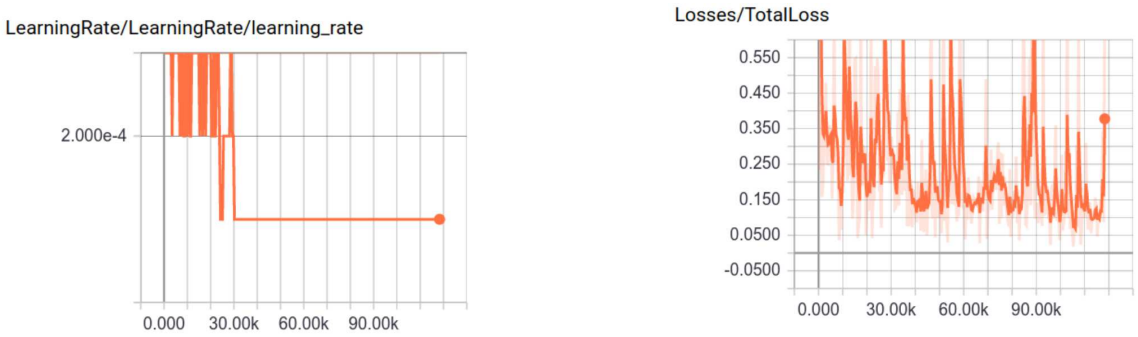


Fig.7 The output of tensor board in 120 thousand epochs of training Investigating of learning Rate, Total Loss

In this method, two modules are used to increase localization, which reduces false positives. The first module is binary classification and the second module is multi-class classification, which includes types of accident from different directions, in different weather conditions, special vehicles and ... If the probability of the image label is on the border, i.e. about 50 to 70% re-enters the second module, this module assigns the sample to its correct class by increasing or decreasing the probability. Finally, two similar modules with different training and checkpoints with different parameters increase the likelihood of choosing the right class.

Another point in the classification that improves the prediction of the proposed method by a few hundred percent is the use of the Pseudo labeling technique, in which images that were correctly identified in the test data from both classes were added to the training data at each stage. It was done in a phased manner and caused a better learning pattern of the accident images than non-accident images.

In general, challenging situations which constituted about 25% of accident images is as follows mAP (mean Average Precision) for Object which is shown in Table 1.

Table 1 Output results before training the second module

mAp (No accident images)	mAp(accident images)	Challenging scenes
27%	71%	Cars in rainy scenes
42%	63%	Cars in Snow scenes
21%	61%	Cars in Foggy scenes
65%	83%	Cars in Crowded scenes

MAp output the proposed method was reported in difficult diagnostic situations, i.e. stormy and snowy rainy scenes, because in clear and transparent scenes, the detection accuracy is high, as it is known that in stormy conditions, many normal scenes lead to accidents due to lack of vision. The data subset to the model makes it possible to better distinguish the model and distinguish between crash and non-crash situations

The output of the proposed method has been improved after the addition of more subcategories in the second module. As it is known, the accuracy of the model has reached 94% and the kappa criterion has reached 88%.

In order to reduce false positive, the challenging images of the model were contrasted with proportionate, non-accident images, and they were tested again. The output was improved with an average of 15% in each class. To put it simply, in order to deal with the error resulted from overlapping of cars which were in a row, the second module has been arranged.

3.2 Evaluation Metrics

Precision (positive predictive value) and recall (sensitivity): are appropriate fraction of retrieved related samples from all and relevant instances respectively. Application of these metrics is depending on understanding and measuring of relevance.

Accuracy: The accuracy criterion is the accuracy of the x-group classification against all items where the x-tag for investigating records is suggested by means of

classification. This criterion indicates how much classification output is trustable.

F-measure: This criterion is a combination of call metrics and accuracy and it is used in cases where it is impossible to consider special importance to each of the two criteria.

Kappa: This criterion is often used to test the reliability of the viewer and to compare the accuracy of the system in terms of how much generated output is coincidental.

Table 2 Evaluation metrics formula

$$Precision = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}$$
$$Accuracy = \frac{TP+TN}{TP+TN+PF+PN}$$
$$F\text{-}Score = \frac{Precision*Recall*2}{Precision+Recall}$$
$$Kappa = \frac{Pr(a)-Pr(e)}{1-Pr(e)}$$

Table 3 General results of the proposed model

Measures				
Accuracy	Precision	Recall	F-Score	Kappa
0.9427	0.9509	0.9238	0.9473	0.8845

The details of Faster R-CNN inception V2 model:

```
Meta-architecture: Faster R-CNN
Number of classes: 2
Preprocess images:
image_resizer {
      min_dimension: 600
      max_dimension: 1024
     }
Feature Extractor: Inception V2
{
first_stage_features_stride: 16
grid_anchor_generator {
      scales: [0.25, 0.5, 1.0, 2.0]
      aspect_ratios: [0.5, 1.0, 2.0]
      height_stride: 16
      width_stride: 16
first_stage_max_proposals: 300
crop_size: 14
}
post_processing{
max_detections_per_class: 100
      max_total_detections: 300
Training phase:
Batch Size=32
momentum_optimizer_value: 0.9
num_steps: 1500000( but stop 120000 because of
decrease and to be fix loss function }
andom_horizontal_flip
Learning schedule: Manually
Stepped{initial_learning_rate: 0.0002}
Classification Loss function: SoftmaxCrossEntropy
Configuration System:
GEFORCE GTX 1080 Ti, Core I7, RAM 16G DDR4
```

5. Conclusion

Considering the fact that car accidents cause the death of millions of people annually, investigating accident anticipation and detection methods, in addition to automobiles, can be useful in surveillance cameras to detect accidents, and to deviate other cars in order to prevent accidents.

The proposed method helps in reducing false positive in overlapped images through adding a second module. At the end of the second module, as the accident threshold exceeds 78%, the system identifies the image as accident image, and through alarming to the driver, makes him aware of the danger of accident. The proposed method showed an accuracy of 94.27% in the test stage. Considering the fact that it has the anticipation of 18 frames before accident on average, it is a good result. Further lines of research can work on data base of formal language to show the event in accident.

Acknowledgements

The authors have no proprietary, financial, professional or other personal interest of any nature in any product, service or company. The authors alone are responsible for the content and writing of the paper. There is no conflict of interest in this paper.

6. References

[1] W. h. organization, *http://www.who.int/gho/road_ safety/mortality/en/,* 2018.

[2] H. P. K. N.Trong, "A Comprehensive Survey on Human Activity Prediction," *International Conference on Computational Science and Its Applications,* no. DOI: 10.1007/978-3-319-62392- 4_30, July 2017.

[3] K.CHEN, D.ZHANG,L. YAO,B.GUO,Z.YU, "Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges and Opportunities," *J. ACM, Vol. 37, No. 4, Article 111. Publication date: August 2018,* no. arXiv:2001.07416v2 [cs.HC] 22 Jan 2021.

[4] R. MS, "Human activity prediction: Early recognition of ongoing activities from streaming videos," Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 1036-1043, 2011.

[5] Hoai M, la Torre FD, " Max-margin early event detectors," International Journal of Computer Vision, vol. 107, no. 2, p. 191–202, 2014.

[6] Lan T, Chen TC, Savarese S, " A hierarchical representation for future action prediction," European Conference on Computer Vision, p. 689–704, 2014.

[7] R. M, " Human activity prediction: Early recognition of ongoing activities from streaming videos," ICCV, p. 201–214, 2011.

[8] T. A. Yuen J, "A data-driven approach for event prediction," European Conference on Computer Vision, p. 707–720, 2010.

[9] Wang Z, Deisenroth MP, Amor HB, Vogt D, Schölkopf B, Peters J, "Prob- abilistic modeling of human movements for intention inference," Proceedings of robotics: Science and systems, 2012.

[10] S. A. Koppula HS, "Anticipating human activities using object affordances for reactive robotic response," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 1, p. 14–29, 2016.

[11] Koppula HS, Jain A, Saxena A, "(2016) Anticipatory planning for human-robot teams," Experimental Robotics, p. 453–470, 2016.

[12] B. D. Mainprice J, "Human-robot collaborative manipulation planning using early prediction of human motion," Intelligent Robots and Systems (IROS),2013 IEEE/RSJ International Conference , p. 299–306 , 2013.

[13] Ma H, Lu N, Ge L, Li Q, You X, Li X, "Automatic road damage detection using high-resolution satellite images and road maps," Geoscience and remote sensing symposium (IGARSS), 2013 IEEE International, p. 3718– 3721, 2013.

[14] Ravindran V, Viswanathan L, Rangaswamy S, "A Novel Approach to Automatic Road-Accident Detection using Machine Vision Techniques", TERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE, 2016.

[15] Ren S, He K, Girshick R, Sun J, "Faster R-CNN: towards real-time ob- ject detection with region proposal networks," IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 6, p. 1137– 1149, 2017.

[16] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC, "Ssd: Single shot multibox detector," European conference on computer vision, p. 21–37, 2016.

[17] Dai J, Li Y, He K, Sun J, "R-fcn: Object detection via region-based fully convolutional networks," Advances in neural information processing systems, p. 379–387, 2016.

[18] Tian, Rong Pan，Yumin; Wang, Zhong“ ,Key-frame Extraction Based on Clustering ”,The 2010 IEEE International Conference on Progress in Informatics and Computing ,2010.

[19] Huang L, Ye C H“ ,The research and implementation of keyframe extracion methods in contentbased teaching video retrieval ”,Consumer Electronics, Communications and Networks (CECNet) 2012 2nd International Conference on. IEEE ,pp .2179-2182 ,2012 .

[20] Uma, BH Shekar ， KP“ ,Kirsch Directional Derivatives Based Shot Boundary Detection: An Efficient and Accurate Method ”,Procedia Computer Science ,VOL 58 ,pp. 565-571 , 2015.