The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Traffic Risk Estimation from On-vehicle Video by Region-based Spatio-temporal DNN trained using Comparative Loss

Kwong Cheong Ng Yuki Murata Masayasu Atsumi

Dept. of Information Systems Sci., Graduate School of Eng., Soka University

Abstract: We propose a method to estimate the traffic risk during road navigation based on the region-based spatio-temporal deep neural network (DNN) trained by the comparative loss function. In this method, moving object regions are extracted using the object detector YOLO and their features are clipped out from the middle layer of the detector. Then, these feature sequence is used to estimate the traffic risk by the spatio-temporal DNN followed by the risk estimation network. Experiments were conducted using the KITTI and Dashcam Accident dataset images and we have shown that it is possible to estimate a dangerous traffic situation using the proposed risk estimation network.

1. Introduction

In the research of advanced driver-assistance systems (ADAS) and autonomous driving, numerous studies have been conducted to estimate the risk of traffic situation from images of driver's viewpoints [1, 2, 7]. In this paper, we propose a traffic risk estimation method based on the region-based spatio-temporal deep neural network (DNN) trained using the comparative loss function. In this method, first, moving object regions are detected using the object detector YOLO [4, 5] from each frame of a video and their features are extracted from a middle layer of the detector. Then, these object feature sequence is input into the spatio-temporal pattern encoding network which consists of a convolutional neural network (CNN) and a long short-term memory (LSTM) and the risk is estimated by the risk estimation network trained by the comparative loss function. The accuracy of the risk estimation is evaluated using a dataset of on-vehicle cameras.

2. Related Work

2.1 Traffic Risk Estimation

Several studies have been proposed in the research of estimating upcoming traffic incidents. Chan *et al.* [1] proposed a dynamic-spatial-attention Recurrent Neural Network (RNN) that could anticipates accidents before they occur. Suzuki *et al.* [2] extended the work of Chan by introducing an Adaptive Loss for Early Anticipation (AdaLEA) method which allows a model to learn earlier incident anticipation as training progresses. In addition, a quasi-recurrent neural network (QRNN) is adopted in the base model to enable stable output from temporal convolution on sequential data such as videos. However, these methods are specialized in anticipating traffic accidents, while our proposed method takes into account learning risk estimation of any traffic situation including accidents, risky incidents, congestion and so on through the comparative loss function. On the other hand, Mark *et al.* [6] suggested a Deep Predictive Model that can learn its own, task-specific filters which improves prediction performance. The model uses a Bayesian convolutional long short-term memory (ConvLSTM) method to process spatio-temporal visual data, proprioceptive data and steering commands to identify potential impending collisions. In addition, Ernest Cheung *et al* [8] proposed another approach called Trajectory to Driver Behavior Mapping (TDBM) that accounts the driving behavior of neighboring vehicles to perform risk assessment.

Contact: Kwong Cheong Ng, e18m5252@soka-u.jp

2.2 Traffic Datasets for Risk Estimation

Several datasets are available for studies of the traffic risk estimation. In [1], the Dashcam video dataset is proposed to evaluate the proposed method in which 678 dashcam videos captured across six cities in Taiwan are used. This dataset is also used in this paper to estimate risk for traffic accidents. To focus on near-miss traffic incidents, Suzuki *et al* [7] introduced a Large Scale Near-Miss Traffic Incident database (NIDB) that comprises over 6.2K videos and 1.3M frames, many of which are incident scenes and are classified into seven classes, including low/high risk for bicycles, pedestrians, and vehicles, as well as a background class. However, as the NIDB dataset is not open-source, it is not used in this paper. Mark *et al.* [6] uses a robotic simulation platform proposed in [9] to simulate experiment data, whereby the training dataset is generated using a series of dynamic street scenes involving two vehicles in a sparsed simulated environment. Due to the dataset is not generated from a real environment, the approach is not used in this paper.

3. Risk Estimation Method

3.1 Overview

As shown in Figure 1, the proposed network consists of the object detector YOLOv2, a moving objects' spatial pattern encoding CNN, a moving objects' spatio-temporal pattern encoding LSTM, and a risk estimation network. First, moving object regions are detected using the YOLO object detector from on-vehicle camera images and their features are extracted from its middle layer. Then, by inputting the feature sequence of these moving object regions into the spatio-temporal pattern encoding network which consists of a CNN and a LSTM, spatio-temporal feature of moving objects are extracted, and the traffic risk is estimated based on the risk estimation network. The risk estimation network is trained by the comparative loss function

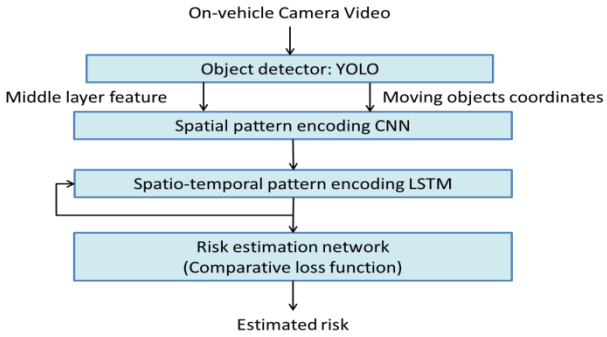

Figure 1: Overview of proposed network

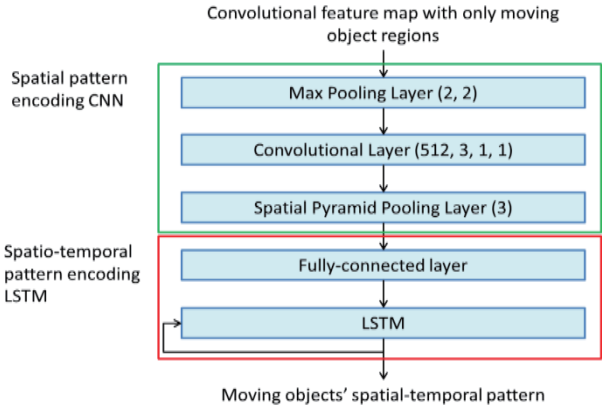

Figure 2: Spatio-temporal pattern encoding network of moving objects

using a set of pairs of images with different relative risk levels.

3.2 Object Feature Extraction

To estimate the risk level on an on-vehicle camera images, it is necessary to detect objects which are involved in traffic risk. Although it is necessary to attend various objects during driving, the important attention targets are moving objects. Therefore, in this research, a pre-trained YOLO using the COCO dataset is further trained using the KITTI dataset [3] and the Dashcam Accident dataset [1], which are on-vehicle camera datasets, to detect moving objects. The target objects to be detected are classified into the following seven categories – car, truck, person, tram, bicycle, motorbike and bus. These categories were unified between the two datasets.

After detecting the moving objects, their features are extracted using the coordinates of them from middle convolutional layers of the YOLO. The 21st layer is used as the middle layer for YOLOv2.

3.3 Moving Objects' Spatio-temporal Feature

The moving objects' spatio-temporal feature is extracted by applying the spatio-temporal pattern encoding network to a composite feature map in which only the convolutional features of moving object regions are clipped from a feature map of the middle layer of YOLO (Figure 1). The network consists of the spatial pattern encoding CNN followed by a spatial pyramid pooling layer and the spatio-temporal pattern encoding LSTM. Figure 2 shows the detailed configuration of the network. In the figure, the (2, 2) of the max pooling layer represents the kernel size and the stride. The (512, 3, 1, 1) of the convolutional layer represents the number of output channels, the kernel size, and the padding width. The (3) in the spatial pyramid pooling layer is the number of pyramid levels.

3.4 Risk Estimation and Comparative Loss Function

In general, it is difficult to evaluate traffic risk by objective numerical values. On the other hand, it is easier to evaluate which is dangerous between a situation where an accident is likely to occur and a situation where no accident occurs or which is dangerous between a congested situation and a non-congested situation. Therefore, we introduce the comparative loss function that learns a correct risk estimation function through relative comparison of the risk level between pairs of two spatio-temporal patterns of moving objects. In training, each estimated risk is tuned according to a supervised signal that gives which is more dangerous. Let two camera images be labelled as v_1, v_2, their relative risk be labelled as $R(v_1, v_2)$, and r_1, r_2 be estimated risks by a sigmoid activation function of the risk estimation network (a perception) respectively. The $R(v_1, v_2)$ is a supervised signal which takes a value of 1 when v_1 is more dangerous than v_2, -1 when v_1 is safer than v_2, and 0 when both are comparatively equal. Then, the comparative loss function is defined by the following equation:

$$L(r_1, r_2, R(v_1, v_2)) =$$

$$\begin{cases} \max(r_2 - r_1 + m, 0) & \dots R(v_1, v_2) = 1 \\ |r_2 - r_1| & \dots R(v_1, v_2) = 0 \\ \max(r_1 - r_2 + m, 0) & \dots R(v_1, v_2) = -1 \end{cases} \quad (1)$$

Here, m is a parameter that provides a margin of relative comparison.

In risk estimation, the degree of risk is computed by the risk estimation network from the moving objects' spatio-temporal pattern in each frame interval of a video.

4. Dataset

In experiments, both the KITTI dataset [3] and Dashcam Accident dataset [1] were used for training the object detection network YOLO, and the Dashcam Accident dataset was used for risk estimation experiments. The Dashcam Accident dataset contains videos with accidents and videos without accidents, and only videos with accidents were used in experiments. Each video consists of 100 frames and accidents occur in the last 10 frames. Therefore, in risk prediction experiments, each video was divided into a non-accident part of the former 50 frames and an accident part of the latter 50 frames. The dataset used for the risk estimation is configured as shown in Table 1.

Table 1. Configuration of a dataset. Videos with accidents (positive) samples and videos without accidents (negative) samples. Numerical values in each cell represent the number of videos/ the number of frames

	Training	Test
positive	455/22750	164/8200
negative	455/22750	164/8200

5. Experiments

5.1 Outline of Experiments

The evaluation of risk estimation was performed for two cases: one was performed between videos with accidents and videos without accidents, and the other was performed between two videos without accidents. In the former, the comparative loss function specifies the relative comparison in which videos with accidents are more dangerous. Here, the relative comparison margin was set to 0.5. In the latter, the comparative loss function specifies relative comparison in which videos with more moving objects are more dangerous. Here in this case, the relative comparison margin was set to 0.1.

5.2 Results

The detection performance of moving objects using YOLOv2 was evaluated by the mean average precision (mAP) and the intersection over union (IOU). As a result, the mAP was 0.2902 and IOU was 0.5996. Since the result of detection performance was not so accurate, we evaluated the risk prediction performance by extracting moving object features from YOLOv2 using ground truth object region boxes.

Table 2 shows the result of risk estimation between videos with accidents and videos without accidents.

In table 2, BG_ZERO and BG_GN represent that the region other than moving objects is filled with zero (for BG_ZERO) and filled with Gaussian noise (for BG_GN) respectively. In addition, 'DO' represents that a dropout is applied to the output of the fully connected layer in Figure 2.

Table 3 shows the result of risk estimation between two videos without accidents but with the different numbers of moving objects.

Table 2. Accuracy of risk estimation between videos with accidents and videos without accidents

	Training (%)		Test (%)	
	w/ DO	w/o DO	w/ DO	w/o DO
BG_ZERO	99.771	97.706	69.565	62.733
BG_GN	97.706	94.839	70.807	72.050

Table 3. Accuracy of risk estimation between videos without accidents but with the different number of moving objects

	Training w/o DO (%)	Test w/o DO (%)
BG_ZERO	93.349	86.957

Based on the results of these experiments, we found that the proposed method achieved a better result when BG_GN was used. This means that the Gaussian noise is useful to achieve robust training for risk estimation. In addition, we found that the proposed method was able to estimate dangerous situation not only caused by accidents but also triggered by congestion.

6. Conclusion

In this paper, we have proposed a traffic risk estimation DNN which is trained by the comparative loss function. This network encodes a spatio-temporal pattern of moving objects based on YOLO and the spatio temporal network and estimates traffic risk using the risk estimation network. Then, in the experiments, we have shown that it is possible to estimate traffic risk by the proposed network. As a future work, we are going to improve the accuracy of preliminary risk prediction by extending the risk estimation network. To improve the efficiency of the feature extraction, we are going to study the latest object detector YOLOv3 in hope of replacing YOLOv2 for better performance. In addition, we hope to study the effect of detecting traffic signs in traffic risk estimation and add them as a new object detection category.

References

[1] Chan, F-H., et al., Anticipating Accidents in Dashcam Videos, ACCV, 2016

[2] Suzuki, T., et al., Anticipating Traffic Accidents with Adaptive Loss and Large-scale Incident DB, CVPR, 2018

[3] Geiger, A., et al, Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite, CVPR, 2012

[4] Redmon, Joseph., Divvala, Santosh., et al., You Only Look Once: Unified, Real-Time Object Detection, 2016

[5] Redmon, Joseph., Farhadi, Ali., YOLO9000: Better, Faster, Stronger, 2018

[6] Strickland, Mark., Fainekos, Georgios., et al., Deep Predictive Models for Collision Risk Assessment in Autonomous Driving, 2017

[7] Suzuki, T., et al, Drive Video Analysis for the Detection of Traffic Near-Miss Incidents, 2018

[8] Cheung, Ernest., Bera, Aniket., et al., Efficient and Safe Vehicle Navigation Based on Driver Behavior Classification, CVPR, 2018

[9] Rohmer, Eric., Singh, Surya P. N., et al., V-REP: a Versatile and Scalable Robot Simulation Framework, 2013