

CRASH: Crash Recognition and Anticipation System Harnessing with Context-Aware and Temporal Focus Attentions (Appendix)

Anonymous Authors

CCS CONCEPTS

• Applied computing → Physical sciences and engineering.

KEYWORDS

Traffic Accident Anticipation; Autonomous Driving; Spatial-Temporal Analysis; Fast Fourier Transform; Dynamic Visual Fusion

1 DATASETS

We evaluate the efficacy of our model using three esteemed datasets: Dashcam Accident Dataset (DAD) [2], Car Crash Dataset (CCD) [1], and AnAn Accident Detection (A3D) [3] datasets, which provide a unique perspective on traffic accidents in various scenes.

DAD. The dataset comprises 620 dashcam videos from six major cities, such as Taiwan. Each video lasts 5 seconds at 20 fps and captures various accident types in congested urban settings. From these videos, 1750 segments are extracted and split into a 70% training set and a 30% testing set, with accidents predefined to occur in the final 10 frames of the footage.

A3D. The dataset comprises 1500 videos captured by car cameras in different East Asian cities, demonstrating diverse weather and lighting conditions. Each video is 5 seconds long, recorded at 20 fps, with 80% allocated for training and 20% for testing. Positive segments contain accidents that occur at the 80th frame.

CCD. The dataset consists of 4500 dashcam videos captured in diverse real-world driving conditions, such as day and night, clear and rainy weather. Each video is 5 seconds long and recorded at 10 fps. The dataset is split into a training set of 80% (3600 videos) and a testing set of 20% (900 videos), with accidents randomly placed in the last 2 seconds for the positive videos.

2 IMPLEMENTATION DETAILS

Our proposed model is implemented using PyTorch and trained on an NVIDIA A40 (48GB) GPU over 80 epochs with a consistent batch size of 10. We use the Adam optimiser, initialising the learning rate at 1×10^{-4} uniformly across all datasets. The object detector is configured to detect up to 19 candidate objects, and the embedding dimension for VGG-16 is set to 4096, and the hidden state dimension of the GRU is fixed at 512. In addition, the ReduceLROnPlateau strategy is used to schedule the learning rate, which adjusts the

rate in response to the model's performance across epochs. Further specifics regarding the implementation and essential parameter settings of our model are provided as follows:

Object Detector and Feature Extractor. These two modules accept the input dimension of video V as (B, T, W, H) , where B denotes the batch size, T represents the number of frames contained in the input video, and W and H correspond to the pixel length of the video's width and height, respectively. In our implementation, B is fixed at 10, and T is determined based on the actual circumstances of different datasets, where for DAD and A3D it is 100, while for CCD it is 50. The output of the Object Detector—object vectors O_F —is a feature matrix of size (B, T, N, d) , where N represents the top- N dynamic objects detected by the Cascade R-CNN with the highest recognition scores within the video stream. These objects are embedded into D -dimensional vectors through VGG-16 and further dimensionally reduced to d dimensions by a Multilayer Perception (MLP) to decrease the computational load. In the experiments, N is set to 19, D is set to 4096, and d is set to 512, implying that up to 19 detected objects are considered for accident prediction. After extracting context features through VGG-16 and performing dimensionality reduction, the output context vector C_F from the Feature Extractor has dimensions $(B, T, 1, d)$. For convenience in computation, we also set the dimensionality of scene features to d . At time step t , we use the corresponding F_o and F_c for accident prediction, with their dimensions being (B, N, d) and $(B, 1, d)$, respectively.

Object-aware Module. At time step t , F_o and the weighted dual-layer hidden state $H_{t-1} = \{H_{t-1}^1, H_{t-1}^2\}$ serve as inputs, where the dimensions of H_{t-1} are $(B, 2, d)$, and both H_{t-1}^1 and H_{t-1}^2 have dimensions $(B, 1, d)$. During the Object Focus Attention process, H_{t-1}^1 and H_{t-1}^2 are transformed into Q_t^1 and Q_t^2 through matrix multiplication, while F_o is converted into K_t and V_t . Two learnable weight parameters, W_α and W_β , are initialized with random values between 0 and 1. Ultimately, the output \bar{F}_o of the Object-aware Module has the shape $(B, 1, d)$.

Context-aware Module. The input context vector F_c with dimensions $(B, 1, d)$ is fed into the Context-aware Module. Initially, a one-dimensional convolution is employed to expand its channels, and the features within each channel are reshaped into 2-dimensional vectors. Subsequently, the Fast Fourier Transform (FFT), implemented via `torch.fft.rfft2`, is applied to transition the features into the frequency domain, resulting in the spectral features S_c . These features have dimensions (B, C, h, w) , where C represents the number of channels, set to 3, and w and h correspond to the width and height of the feature map, respectively, with the relationship $d = w \times h$. Finally, through Context-aware Attention and the Inverse Fast Fourier Transform (IFFT), S_c is transformed back into the physical space, yielding the final output \bar{F}_c with dimensions $(B, 1, d)$.

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or internal use, or for the internal or personal use of specific clients, is granted by ACM for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM MM, 2024, Melbourne, Australia

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

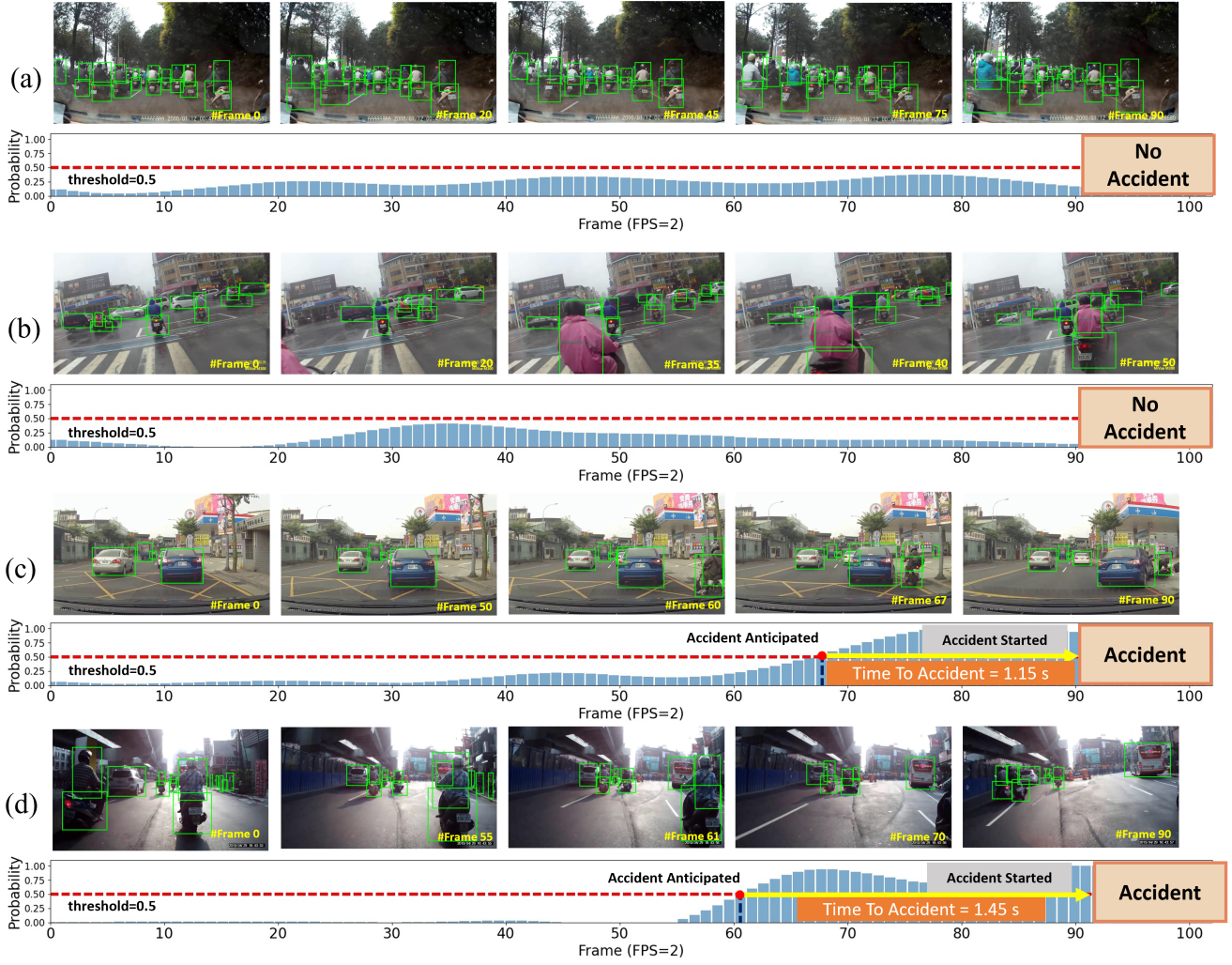


Figure 1: Qualitative Results of the CRASH on the DAD dataset: (a) and (b) depict scenarios without accidents, while (c) and (d) illustrate scenarios with accidents. The blue bars represent the model's output of the probability of an accident occurrence, with the threshold uniformly set at 0.5

Multi-layer Fusion. Finally, the outputs from the Feature Extractor, Object-aware Module, and Context-aware Module are concatenated into a mixed feature F_m with the dimensions of $(B, 1, d + d + d)$. F_m is subsequently passed into a two-layer Gated Recurrent Unit (GRU) to obtain the current time step's hidden state $h_t = \{h_t^1, h_t^2\}$, with dimensions $(B, 2, d)$. By concatenating the first and second layers of all hidden states from the past M frames, we obtain $\bar{H}_t = \{\bar{H}_t^1, \bar{H}_t^2\}$, and both \bar{H}_t^1 and \bar{H}_t^2 have dimensions (B, M, d) . In the Temporal Focus Attention, the number of layers is set to K , with each layer having 2 blocks that perform multi-head attention operations on \bar{H}_t^1 and \bar{H}_t^2 , respectively. In our implementation, the number of layers K and the number of heads are both set to 8. The output of each layer is then weighted, aggregated, and concatenated to obtain the final output H_t with dimensions $(B, 2, d)$.

3 QUALITATIVE RESULTS

In our research, through a comparative analysis of the experimental results on the DAD and CCD datasets, we delve into the model's predictive performance on non-accident (negative) and accident (positive) videos, showcasing these through visualization techniques. Specifically, in Fig. 1 and Fig. 2, the first two images illustrate non-accident scenarios, while the latter two depict accident scenarios. For instance, in the scene presented in Fig. 1 (a), numerous electric scooter riders navigate through constantly changing distances among each other. Despite the complexity of the scene, the model's accident prediction score remains at a high level without ever reaching the preset threshold. Fig. 1 (b) captures a moment where an electric scooter suddenly enters the frame within 20 frames, rapidly closing the distance with the main subject. Consequently, the model's predicted score climbs, nearing the threshold, but as the electric



Figure 2: Qualitative Results of the CRASH on the CCD dataset: (a) and (b) depict scenarios without accidents, while (c) and (d) illustrate scenarios with accidents. The blue bars represent the model’s output of the probability of an accident occurrence, with the threshold uniformly set at 0.5

scooter distances itself, the score correspondingly decreases. In Fig. 2 (a-b), we specifically highlight the model’s predictive capability when encountering obscured views (such as wipers obstructing the view) and scenarios at night-time intersections, where the prediction scores are generally lower, with only minor fluctuations as it adjusts to new scenes. These observations fully demonstrate our model’s high sensitivity to accident prediction and its adaptability to varying traffic environments. By making careful evaluations, the model effectively reduces the risk of false positives, showcasing its accuracy and stability in predicting accidents in complex traffic scenarios.

To thoroughly test the model’s ability to predict accidents, we specifically select a series of complex scenarios for analysis, including nighttime (as shown in Fig. 2 (c)) and snowy conditions (as depicted in Fig. 2 (d)). Despite the presence of numerous interfering

factors, such as streetlights and car headlights, which could affect the clarity of the footage, our model is still able to accurately identify the participants in the accident using the object detector and make timely predictions. Notably, in most cases, the subjects involved in the accidents are present in the video from the beginning. However, we also consider special circumstances—as illustrated in Fig. 1(c), where the subject likely to be involved in an accident only appears 60 frames into the video, merely 30 frames (approximately 1.5 seconds) before the accident occurs. Our model is capable of analyzing and predicting the intention of a blue sedan to make a right turn and the trajectory of an electric scooter within a short time frame, issuing a prediction by the 67th frame. Furthermore, in Fig. 1(d), we demonstrate a relatively rare situation where a traffic accident is caused by the actions of the driver itself, rather than a direct collision with another road user. By analyzing the abnormal

trajectory and body movements of a motorcycle rider, our model was able to predict a potential risk 1.45 seconds before the accident occurred.

These experimental results showcase our model’s capacity to assess the risk of accidents in complex traffic environments, based on the dynamic behaviors of different road users. It not only accurately identifies the subjects involved in accidents but also provides early warnings, thereby proving the model’s predictive validity and interpretability.

REFERENCES

- [1] Wentao Bao, Qi Yu, and Yu Kong. 2020. Uncertainty-based Traffic Accident Anticipation with Spatio-Temporal Relational Learning. In *Proceedings of the 28th ACM International Conference on Multimedia (MM ’20)*. 407–410.
- [2] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. 2017. Anticipating Accidents in Dashcam Videos. In *Computer Vision – ACCV 2016*. Springer International Publishing, Cham, 136–153.
- [3] Yu Yao, Mingze Xu, Chiho Choi, David J Crandall, Ella M Atkins, and Behzad Dariush. 2019. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 9711–9717.