Detection of Road Accidents Using Synthetically Generated Multi-Perspective Accident Videos

Thakare Kamalakar Vijay, *Graduate Student Member, IEEE*, Debi Prosad Dogra, *Member, IEEE*, Heeseung Choi, *Member, IEEE*, Gipyo Nam, and Ig-Jae Kim, *Member, IEEE*

***Abstract*— Road accidents are often caused by short abnormal events, including traffic violations, abrupt change in vehicular motion, driver fatigue, etc. Observing an accident event from the right camera perspective plays a crucial role while detecting accidents. However, it may not be possible to capture such abnormal events from a limited camera perspective. We present a deep learning framework to analyze the accident events recorded from multiple perspectives. First, we estimate feature similarity in videos recorded from multiple perspectives. We then divided the video samples into high and low feature similarity groups. Next, we extract spatio-temporal features from each group using two-branch DCNNs and fuse them using a rank-based weighted average pooling strategy followed by classification. We present a new road accident video dataset (MP-RAD), where each accident event is synthetically generated and captured from five independent camera perspectives using a computer gaming platform. Most of the existing road accident datasets use egocentric views or they are captured in fixed camera setups. However, our dataset is large and multi-perspective that can be used to validate ITS-related tasks such as accident detection, accident localization, traffic monitoring, etc. The dataset contains 400 accident events with a total of 2000 videos. We provide temporal annotations of all videos. The proposed framework and the dataset have been cross-validated with latest accident detection baselines trained on real-world road accident videos and vice-versa. The sub-optimal detection accuracy obtained using the baselines indicates that the proposed framework and the dataset can be useful for ITS related research. Code and dataset is available at: https://github.com/draxler1/MP-RAD-Dataset-ITS-**

***Index Terms*— Anomaly detection, road accident detection, multi-perspective input, feature similarity.**

I. Introduction

THE progress on designing of intelligent traffic monitoring systems and autonomous vehicles is happening in leaps and bounds. Despite spectacular progress in AI-driven research in the fields mentioned above, road safety and security are remaining key challenges. As the number of road accidents is increasing [1], [2], detection and localization of such events are now seriously being studied by the CV research community. Huang et al [3] have developed a two-stream convolution neural network to detect near-accident scenarios. Several recent works [4], [5], [6], [7], [8], [9] have mapped the accident detection problem into a video anomaly detection paradigm. Moreover, forecasting [1], [10] and early-collision detection [11] have grabbed the attention of researchers. However, due to the lack of adequate training samples, the challenges remain and need further investigation.

Traffic monitoring can be done either using fixed camera setups or cameras mounted on moving vehicles [1], [10], [12]. The advantage of a fixed camera setup over the first-person view setup can be two-fold: (i) fixed surveillance cameras are usually installed at a height to provide a wider viewpoint; (ii) in a fixed setup, we get better information about the traffic flow as the background remains unchanged. Moreover, existing traffic analysis video datasets [13], [14], [15], [16], [17] suffer from following issues: (i) only a limited number of accident videos are included as such events are rare; and (ii) egocentric view or single-camera setups do not provide wider viewing experiences.

In this paper, we have addressed the issues mentioned above by introducing a novel accident detection framework by exploiting feature similarity information shared between different views of an accident. Our proposed method takes advantage of the multi-perspective views. We then propose a new feature-similarity-based weighted average pooling to pool similar features across all viewpoints. To the best of our knowledge, this is the only method that uses synthetic accident videos for training and it produces SOTA results in accident detection domain. Moreover, we have introduced a new dataset, referred to as Multi Perspective-Road Accident Dataset (MP-RAD), consisting of accident videos generated synthetically using a well-known gaming platform. The dataset contains 400 accident events captured using five different camera perspectives resulting in a total of 2000 videos. Frame-level (temporal) annotations of the accident events are provided. This can be used to evaluate various traffic monitoring solutions including trajectory predictions [15], [18], collision anticipation [1], and temporal segmentation [19]. We show that the existing accident detection methods perform reasonably well on

Manuscript received 25 March 2022; revised 19 September 2022; accepted 1 November 2022. This work was supported in part by the Korea Institute of Science and Technology (KIST) through Institutional Program under Project 2E31582 and in part by NRF funded by the IIT Bhubaneswar under Project CP283 and Project 2018M3E3A1057288. The Associate Editor for this article was H. Huang. *(Corresponding author: Thakare Kamalakar Vijay.)*

Thakare Kamalakar Vijay and Debi Prosad Dogra are with the School of Electrical Sciences, Indian Institute of Technology Bhubaneswar, Bhubaneswar, Odisha 752050, India (e-mail: tkv15@iitbbs.ac.in; dpdogra@iitbbs.ac.in).

Heeseung Choi and Ig-Jae Kim are with the Artificial Intelligence and Robotics Institute, KIST, Daejeon 34141, South Korea, and also with the Yonsei-KIST Convergence Research Institute, Yonsei University, Seoul 03722, South Korea (e-mail: hschoi@kist.re.kr; drjay@kist.re.kr).

Gipyo Nam is with the Artificial Intelligence and Robotics Institute, KIST, Daejeon 34141, South Korea (e-mail: gpnam@kist.re.kr).

This article has supplementary downloadable material available at https://doi.org/10.1109/TITS.2022.3222769, provided by the authors.

Digital Object Identifier 10.1109/TITS.2022.3222769

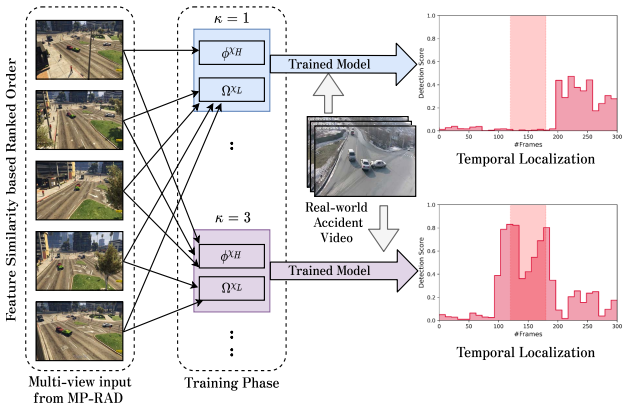

Fig. 1. Models trained using synthetically generated videos and testing on real-world accident videos. Shaded strips are temporal ground truths. It can be observed that $\kappa = 3$ (partitioning hyper-parameter) provides the best results as per the ground truths.

real-life accident videos even when these methods are solely trained using a synthetic dataset. An overview of the work can be found in Fig. 1.

A. Motivation and Contributions

Despite the potential benefits of surveillance cameras in traffic monitoring, the number of datasets focusing on accident detection is less due to the following reasons: (i) surveillance cameras are often controlled by the regulatory authorities, hence accessing traffic data is difficult; (ii) road accident events are rare in nature, thus one has to acquire a large volume of traffic data to capture a sizable number of accident events; (iii) publicly available datasets are inadequate due to limited samples, poor annotations, and lack of camera details. Therefore, we overcome the problems mentioned above by introducing a synthetically generated, large-scale, and multi-perspective road accident dataset. In summary, this paper makes the following contributions:

- We have proposed a new architecture for road accident detection that takes advantage of multi-view inputs. We have introduced feature similarity-based weighted average pooling that intelligently fuses the features obtained from the videos captured from similar camera perspectives.
- The paper also presents a synthetically generated multi-perspective road accident video dataset (MP-RAD). We have prepared the videos using a well-known gaming platform and temporally annotated all 2000 videos. The dataset can be useful for training, testing, and validating ITS related research works.
- We experimentally validate the proposed dataset, benchmark it against two major real-world road accident video datasets using the proposed architecture and the latest SOTA road accident detection methods.

The rest of the paper is organized as follows. In the next section, we present the related work. In Section III, we present the proposed methodology. Section V presents the datasets, experimental results, and discussions. In Section VI, we present the conclusion and future scopes.

II. Related Works

A. Accident Detection Methods

Several attempts [3], [4], [20], [21], [22], [23], [24], [25], [26], [27] have been made to detect accidents in videos. Zhang et al [28] and Meng et al [20] have used social media data to detect road accidents. Zu et al [26] have used Gaussian Mixture Model (GMM) for characterizing vehicular motion. Huang et al [3] use two-stream convolutional network architecture that can perform real-time accident detection. Singh et al [24] estimate accident scores using de-noising auto-encoders. Some deep learning-based methods [4], [5], [7] treat road accidents as abnormal events. Sultani et al [4] have extracted temporal features using C3D [29] and applied Multiple Instance Learning (MIL) classifier to detect abnormalities. Following their approach, several works [5], [7], [30], [31] detect road accidents as anomalous events. However, existing methods suffer due to limited training samples. Moreover, these methods use restricted viewpoint (both in a single camera and egocentric setup), leading to poor performance. We show that using multi-view inputs with viewpoint partitioning, the novel weighted pooling strategy proposed in this work can be used to address the aforementioned limitations.

B. Datasets for Traffic Monitoring

MIT traffic dataset [13] provides 19 minutes of raw video recorded using a fixed camera setup. NGSIM dataset [14] contains videos captured using overhead intersection cameras. CVRR [15] is a traffic dataset consisting of simulated intersections and real-life highway videos for benchmarking trajectory-based activity detection tasks such as prediction, clustering, and abnormality detection. QMUL dataset [16] contains an hour-long traffic surveillance data collected at a busy city intersection. KIT [17] and urban Tracker [32] datasets consist of a few videos recorded using stationary camera setups in varying environmental conditions like fog and rain. Apart from traffic monitoring datasets mentioned above, there are a few datasets highlighting road accident events. UCF-Crime [4] dataset contains 13-real world anomalies, including accidents, explosions, stealing, etc. It contains 150 road accident videos, 127 for training and 23 for testing. Car Accident Detection and Prediction (CADP) [1] dataset contains visuals of 1416 road accident events recorded using fixed camera setups. XD-violance [33] dataset contains 444 accident videos collected from the movies, sports, CCTV cameras, etc. Dashcam Accident Dataset (DAD) [11], BDD100k [34] and A3D [10] are a few datasets containing accident videos recorded in egocentric views. However, no dataset is available containing synthetically generated multi-view road accident events. The proposed MP-RAD dataset offers this to the CV research community.

III. Proposed Methodology

As shown in Fig. 2, the proposed accident detection framework consists of three major modules. First, we partition viewpoints into two subsets according to their feature

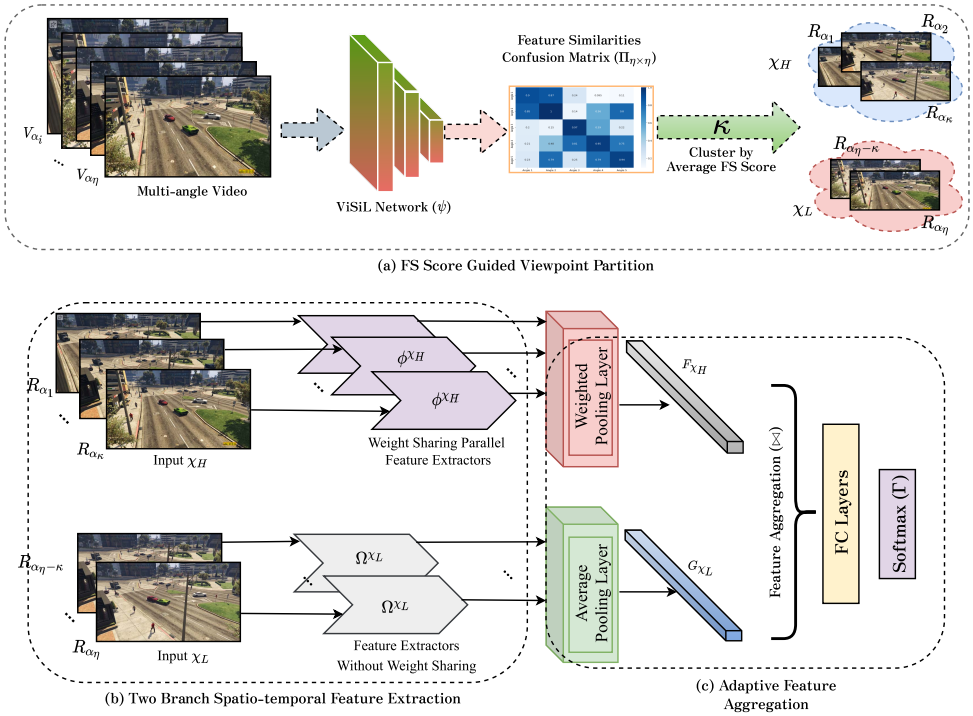

Fig. 2. Illustration of the proposed framework. (a) The accident event videos are divided into two subsets, χ_H and χ_L, according to average feature similarity (FS) score. (b) A two-branch spatio-temporal feature extractor is used to extract viewpoint features. (c) Finally, the features obtained from these two branches are aggregated using the global pooling layers before classification.

similarity score. Second, our proposed two-branch DCNN extracts spatio-temporal features from each set. Third, we fuse the obtained features using novel feature similarity-based weighted average pooling followed by classification. Assume an accident event (ξ) is being observed from a set of independent positions denoted by $[\alpha_1, \alpha_2, \ldots, \alpha_\eta]$. Let the collection of videos representing the event ξ be denoted by $\chi = [V_{\alpha_i}]^T$, where V_{α_i} represents the video recorded from the i^{th} position such that $1 \leq i \leq \eta$. We define an objective function ψ to calculate the feature similarity and a model hyper-parameter κ to partition the elements of χ into two sets χ_H and χ_L, such that $\chi = \chi_H \cup \chi_L$. In the next section, we explain how ψ is implemented. For each event ξ, a confusion matrix is constituted to assign a relative ranking of videos based on the feature similarity score. In the next stage, a two-branch DCNN has been designed to extract spatio-temporal features from the χ_H and χ_L sets. The final stage intelligently aggregates the features using an ranked-based, weight-adaptive pooling strategy before being fed to a trainable classifier module.

A. Feature Similarity-Guided Viewpoint Partitioning

The idea of partitioning α_η viewpoints into two groups is derived from the following intuition. Object features that are seen from similar camera perspectives are easier to group due to higher feature similarity. Pieces of evidence that are present in χ_H set provide samples with lesser angular variations. When we observe an object from closer angular perspectives, we get stronger evidence. However, independent features that are not covered in χ_H cannot be ignored as they provide a broader perspective of the scene. Visual evidences present in χ_L provide such features. Thus, we have partitioned the videos into χ_H and χ_L sets based on the feature similarity. We calculate feature similarity score (FS) for each viewpoint by using Eq. (1),

$$[F_{\alpha_i,\alpha_j}]^T = \psi\left(V_{\alpha_i}; V_{\alpha_j}\right), \quad 1 \leq j \leq \eta \tag{1}$$

where $[F_{\alpha_i,\alpha_j}]^T$ represents a η-dimensional feature similarity vector for angle α_i of the event and ψ is the function to estimate the feature similarity score between α_i and α_j. We have used the ViSiL network [35] to implement ψ. ViSiL utilizes fine-grained spatio-temporal correlation between a pair of videos, where intra-frame and inter-frame relations are preserved in better fashion. Moreover, the it uses *Tensor Dot* (TD) and *Champer Similarity* (CS) measure over the deep video features to calculate video-level similarity scores which provide accurate measures over its competent method LAMV [36]. We estimate η such feature similarity vectors, one for each angle/location. Thus, a confusion matrix $\Pi_{\eta\times\eta}$ is constituted. Each cell of Π represents a feature similarity score between $[0, 1]$ with the diagonal elements being 1. We now calculate the average feature similarity score for each angle (α_i) using Eq. (2).

$$\overline{F}_{\alpha_i} = \frac{1}{\eta}\sum_{j=1}^{\eta} F_{\alpha_i,\alpha_j}, \quad \forall\alpha_i \tag{2}$$

Now, a ranking of the positions/angles is obtained based on $\overline{F}_{\alpha_i}$. Higher the average feature similarity value, better the rank which is given in Eq. (3), where $R(.)$ is the rank of an angle/position with $i \neq j$.

$$R(\alpha_i) < R(\alpha_j), \text{ if } \overline{F}_{\alpha_i} > \overline{F}_{\alpha_j} \tag{3}$$

We now introduce κ, a hyper-parameter to partition χ into χ_H and χ_L using Eq. (4),

$$[\chi_H, \chi_L] = \sigma(\Pi, \kappa) \tag{4}$$

where σ is partitioning function that partitions χ into two sets, e.g. χ_H and χ_L with $|\chi_H| = \kappa$ and $|\chi_L| = \eta - \kappa$, respectively and κ is a value between $[1{:}\eta]$. The set χ_H contains κ number of videos with higher relative feature similarity scores and χ_L contains the remaining $\eta - \kappa$ videos with lesser feature similarity. The intuition behind such partitioning is to extract strong evidences from χ_H rather than from every η perspectives.

B. Spatio-Temporal Feature Extraction

Individual video frames represent spatial information. Consecutive frames of a video can represent the temporal information [37], [38] that is important for event detection [19] and classification [39]. To incorporate spatio-temporal features in the accident detection framework, we have extracted features using two separate branches denoted by the models (ϕ^{χ_H}) and (Ω^{χ_L}), respectively. The first branch is dedicated to extract features for highly correlated viewpoints (χ_H) and the second one extracts features independently from χ_L. The first branch (ϕ^{χ_H}) consists of a series of κ number of identical 3D-CNNs that share weights in a similar way to T-C3D [38]. The idea behind using a fixed network topology with unchanged weights is to extract similar set of spatio-temporal features from the videos in χ_H. The second branch (Ω^{χ_L}) consists of a series of $(\eta - \kappa)$ number of 3D-CNNs without any weight sharing. We extract the spatio-temporal features using Eq. (5)-Eq. (6) from two different sets of videos,

$$f_{\alpha_i}^{\chi_H} = A\left(\phi^{\chi_H}; W^H\right)_{\alpha_i \in \{1 \ldots \kappa\}} \tag{5}$$

$$g_{\alpha_j}^{\chi_L} = B\left(\Omega^{\chi_L}; W^L\right)_{\alpha_j \in \{(\eta-\kappa) \ldots \kappa\}} \tag{6}$$

where A represents the feature extractor function applied on the video ($\in \chi_H$) recorded from the i^{th} angle to generate feature-maps $f_{\alpha_i}^{\chi_H}$. Similarly, B represents the feature extractor function applied on the video ($\in \chi_L$) recorded from the j^{th} angle to generate feature-maps $g_{\alpha_j}^{\chi_L}$. Here, W^H and W^L are weights of the ϕ^{χ_H} and Ω^{χ_L} such that $1 \leq i \leq \kappa$, and $\eta - \kappa \leq j \leq \eta$. The obtained feature maps are then fed to pooling layers followed by feature aggregation.

C. Pooling Strategy and Feature Aggregation

In this subsection, we provide a detailed description of the proposed pooling layers and the feature aggregation function. The pooling strategies adopted in respective branches are depicted in Fig. 3.

1) Weighted Average Pooling for χ_H: The input set χ_H is processed in parallel using ϕ^{χ_H}. Here, we have introduced a *ranked-based weighted average pooling* scheme that pools the information based on the feature similarity score. The goal of this pooling is to produce a set of linear weights to perform an element-wise weighted average fusion on the outputs of each χ_H extractor. The pooling function is presented using

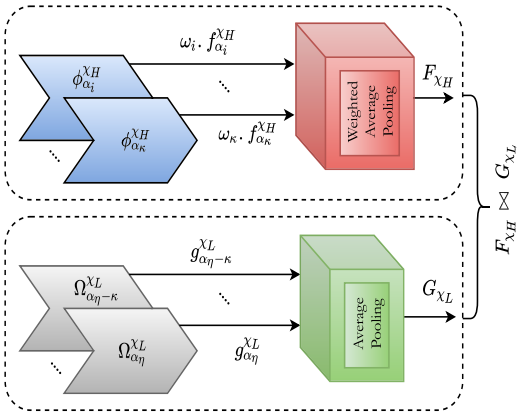

Fig. 3. Two pooling strategies used across different branches of the feature extraction modules. The top part of the figure presents the proposed weighted pooling strategy used in the ϕ^{χ_H} branch. Well-known average pooling is applied to Ω^{χ_L} branch as depicted in the figure.

Eq. (7), where ω_i is the weight for the i^{th} input such that $\sum_{i=1}^{\kappa} \omega_i = 1$ and $1 \leq i \leq \kappa$.

$$F_{\chi_H} = \frac{1}{\kappa} \sum_{i=1}^{\kappa} \omega_i . f_{\alpha_i}^{\chi_H} \tag{7}$$

The value of ω has been varied depending on the degree of feature similarity obtained using Eq. (3). This implies, higher the average feature similarity score, higher the value of ω. We have introduced this weighted average pooling based on the assumption that an event when looked from nearby locations (lesser change in the perspective of viewpoints), the evidence gets stronger.

2) Average Pooling for χ_L: We adopt the standard average pooling in order to fuse the outputs of Ω^{χ_L} feature extractors as given in Eq. (8).

$$G_{\chi_L} = \frac{1}{\eta - \kappa} \sum_{j=\eta-\kappa}^{\eta} g_{\alpha_j}^{\chi_L} \tag{8}$$

The basic assumption of average pooling is to utilize the feature information of all the remaining independent angles/views from χ_L.

3) Aggregation Function and Classification: The final classifier model is a series of fully-connected layers. The training process of this model begins by dividing each input video into a fixed number of temporal segments. The feature extractors ϕ^{χ_H} and Ω^{χ_L} extract the features of these temporal segments followed by respective pooling layers. The outputs of two separate pooling layers are then fused before feeding them to the classifier model, i.e., $\overline{FG} = (F_{\chi_H} \bowtie G_{\chi_L})$, where $\bowtie$ represents the concatenation-based fusion strategy. In such a strategy, the feature vectors from different branches are arranged one after another, horizontally or vertically. We have used vertical stacking during the concatenation. The fused output $\overline{FG}$ is then fed to the classifier that generates the probability of an accident event, i.e, $p = \Gamma(\Lambda(\overline{FG}))$, where p is the Softmax probability for the accident class, Λ is the output neuron of the classifier model and Γ is the widely used Softmax function. Since our model predicts accident event only, we have incorporated a binary-cross entropy loss function

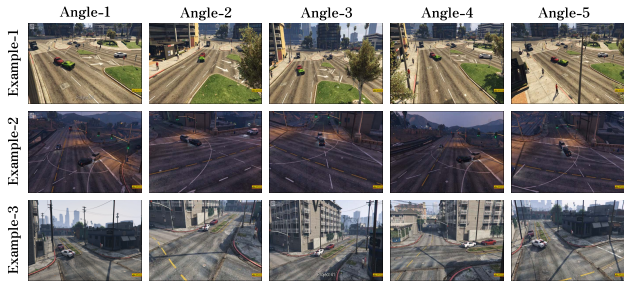

Fig. 4. A few sample videos of the MP-RAD dataset. Each row presents an event from five different camera perspectives.

as given in Eq. (9), where L is final loss, Z is the total number of temporal segments in a video, Y_i values represent ground truths, and p_i is the Softmax probability for accident class of the ith segment.

$$L = -\frac{1}{Z}\sum_{i=1}^{Z}[Y_i \cdot \log(p_i) + (1 - Y_i) \cdot \log(1 - p_i)] \quad (9)$$

The training process of our classifier is similar to the process commonly used in supervised methods such as fully connected networks in ANN, Random Forest (RF) etc. After training, the model is used to test the real-world/synthetic accident video. The last layer of the trained-classifier model generates the Softmax probability of the accident class for every temporal segment.

IV. Proposed MP-RAD Dataset

In this section, we present a detailed discussion about the proposed MP-RAD dataset with statistics. Though the existing datasets provide wider application-specific incentives, a few areas still need attention. For examples, lack of available training samples, availability of full annotations, and fixed camera setups are a few of them. To overcome these issues, we have prepared a new dataset, referred to as Multi-Perspective Road Accident Detection (MP-RAD) dataset. It contains 400 unique road accidents events synthetically generated using a gaming platform Rockstar's GTA-V. Each event is captured from five independent camera angles resulting a total 2000 video samples.

A. Video Collection

Capturing videos of real-life road accident events can be challenging and time consuming. Synthetically generated videos can be good substitutes when real-life data collection is challenging. Thus, we have recorded the accident events by varying parameters such as camera angles, type of vehicles involved, speed of vehicles, and lighting and weather conditions. Our data collection process follows a two-step approach: (i) A player drives a vehicle and commits an accident intentionally with nearby vehicles or pedestrians. (ii) Using varying camera settings, we have captured the event from five viewing perspectives. The intrinsic and extrinsic parameters of the recording setups (e.g. height, angle, focus, etc.) are random for each event. We mimic accidents in the simulated environment by observing real-world accident events available in datasets such as UCF-Crime [4] and CADP [1]. A few samples of the proposed MP-RAD dataset are presented in Fig. 4.

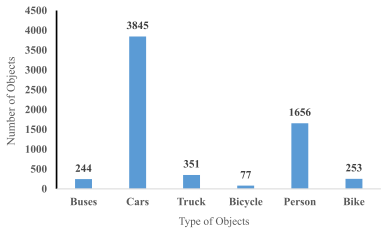

Fig. 5. Distribution of objects in the MP-RAD dataset.

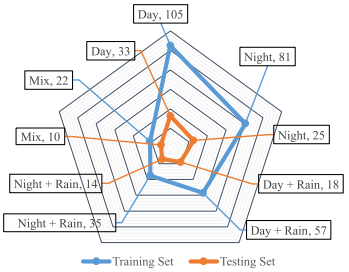

Fig. 6. Distribution of videos according to varying environmental conditions in the training and test sets.

TABLE I
Details of Popular Datasets Used for Accident Detection

Dataset	# of videos	Annotations	Source
UCF-Crime [4]	150	No	CCTV
CADP [1]	1416	Temporal + Spatial	CCTV
XD-Violence [33]	444	Temporal	Movie + Dashcam
Street Accident[40]	620	Temporal	Dashcam
A3D [10]	1500	Temporal + Spatial (Eye-gaze)	Dashcam
Herzig *et al.*[41]	803	Temporal	Dashcam
DoTA [42]	4677	Temporal + Spatial	Dashcam
Ours	**2000**	**Temporal**	**Gaming Platform**

B. Dataset Annotations

We have performed temporal annotations (frame-level) of the videos. We have marked the start and end frames of accident events. Similar to the works [1], [34], we have asked volunteers to annotate MP-RAD videos. The average frame number is taken as the final boundary of an accident event. A comparative study on the datasets is presented in Table I.

C. Dataset Statistics

The dataset contains objects including pedestrians, bikes, cars, trucks, bicycles, buses, and few other types of vehicles. The distribution of objects appeared in the dataset is presented in Fig. 5. Different weather conditions such as sunny day, clear night, day-raining, and night-raining have been considered. All videos have been recorded with 1920 $\times$ 1080 resolution at 30 FPS. The average length of an accident videos is 6.11 seconds. Since the primary focus of the dataset is to record accident events, we have not included normal behaviour or other anomaly categories as opposed to UCF-Crime [4] and XD-Violence [33] datasets. Out of 400 events, we have randomly selected 300 events for training and remaining 100 events for testing. The distribution of videos according to varying environmental conditions is presented in Fig. 6.

V. Results and Discussions

A. Datasets and Evaluation Metric

We have carried out several experiments to validate the MP-RAD dataset and the proposed method. We have selected UCF-Crime accident category [4] and CADP [1] datasets as these videos were recorded using fixed camera setup. Other datasets mentioned in Table I have not been included in our experiments primarily due to egocentric/Dashcam recording.

1) UCF-Crime: Is a large-scale video anomaly dataset containing 13 real-world anomalies recorded using CCTV cameras. It contains 150 real-world road accident videos (127 for training and 23 for testing). Car Accident Detection and Prediction **(CADP)** dataset contains 1416 real-world accident videos collected from YouTube.

2) Evaluation Metric: Following previous works [4], [5], [7], [24], [31] on accident detection, we compute frame-level receiver operating characteristics (ROC) curve and area under the curve (AUC) as evaluation metrics. A larger AUC implies higher distinguishing ability of the method. Since we do not build the normality profile, we have not used FAR for evaluation.

B. Implementation Details

We have extracted C3D [29] and I3D [39] features. ViSiL network [35] has been used to implement ψ. All training videos are divided into non-overlapping segments of 10 frames. We have experimented by varying the segment size to 8, 10, and 16 frames, respectively and tested different fusion ($\bowtie$) such as concatenation, multiplication, etc. No significant performance variation has been observed. ϕ and $\mathbf{\Omega}$ have been kept fixed for all experiments. We extract features from the layer before *pool_5* of the C3D network and layer before *global_avg_pool* of the I3D network. We have carried out all experiments with the following hyper-parameters: $\eta = 5$, $1 \leq \kappa \leq \eta$. The final classifier model is a 3-layer MLP similar to the model proposed by Feng et al [30], where the number of units are 512, 32, and 1, respectively, regularized by dropout with probability of 0.6 between each layer. ReLU and Sigmoid functions are deployed after the first and last layers, respectively. We have trained the model with a learning rate of 0.01 for 350 iterations using Adagrad [43] optimizer.

C. Dataset Cross-Validation

Since MP-RAD is a synthetically generated dataset, a cross-validation is needed to show its relevance. We have carried out cross-validation in two ways: (i) training using MP-RAD and testing on real-world videos (ii) training using real-world videos and testing on MP-RAD videos. The results of first approach is presented in Table II.

The results reveal that our method outperforms other baselines when trained on MP-RAD and tested on real-world videos. Moreover, SOTA methods, when trained on MP-RAD dataset, perform reasonably well. This proves that even synthetically generated samples can be used to train models. The results of second approach is presented in Table III. The results reveal that all the baselines can detect synthesized accidents even when they are trained on real-world videos. Our method could not perform as good as MIST [30] due to unavailability of multi-view videos in UCF-Crime and CADP datasets.

TABLE II

Quantitative Comparisons With Existing Accident Detection Methods Explicitly Trained on MP-RAD and Tested on UCF-Crime and CADP Datasets

		AUC(%)	
Method	**Features**	**UCF-Crime**	**CADP**
Hasan *et al.* [44]	AE^{RGB}	50.01	50.26
Lu *et al.* [45]	Dictionary	50.46	50.49
Sadek *et al.* [25]	HFG	50.90	50.70
Yun *et al.* [46]	MIF	50.20	51.81
Sultani *et al.* [4]	$C3D^{RGB}$	51.21	51.76
	$I3D^{RGB}$	53.30	52.10
Singh *et al.* [24]	AE^{RGB}	52.32	53.33
Zhong *et al.* [5]	$C3D^{RGB}$	53.17	55.19
	TSN^{RGB}	55.10	57.50
	$TSN^{OpticalFlow}$	52.07	52.36
Tian *et al.* [31]	$I3D^{RGB}$	55.31	54.40
Peng Wu *et al.* [33]	$C3D^{RGB}$	52.20	51.20
Pang *et al.*[47]	$C3D^{RGB}$	57.72	52.33
	$I3D^{RGB}$	59.33	54.39
Feng *et al.*[30]	$C3D^{RGB}$	59.44	55.35
	$I3D^{RGB}$	59.96	57.65
Ours	$C3D^{RGB}$	**60.36**	**58.09**
	$I3D^{RGB}$	**61.60**	**59.81**

TABLE III

Quantitative Comparisons of Existing Accident Detection Methods Explicitly Trained on UCF-Crime and CADP Datasets and Tested on MP-RAD Dataset

		AUC (%)	
Method	**Features**	**UCF-Crime**	**CADP**
Hasan *et al.* [44]	AE^{RGB}	50.00	50.14
Lu *et al.* [45]	Dictionary	50.49	50.84
Sadek *et al.* [25]	HFG	51.33	51.60
Yun *et al.* [46]	MIF	51.90	52.32
Sultani *et al.* [4]	$C3D^{RGB}$	52.03	53.39
	$I3D^{RGB}$	54.17	55.60
Singh *et al.* [24]	AE^{RGB}	53.51	54.56
Zhong *et al.* [5]	$C3D^{RGB}$	54.44	55.64
	TSN^{RGB}	56.72	58.23
	$TSN^{OpticalFlow}$	55.79	52.60
Tian *et al.* [31]	$I3D^{RGB}$	52.13	55.22
Peng Wu *et al.* [33]	$C3D^{RGB}$	51.30	53.36
Pang *et al.* [47]	$C3D^{RGB}$	54.57	57.20
	$I3D^{RGB}$	54.48	57.78
Feng *et al.* [30]	$C3D^{RGB}$	**60.71**	**62.85**
	$I3D^{RGB}$	**64.20**	**68.12**
Ours	$C3D^{RGB}$	51.19*	52.19*
	$I3D^{RGB}$	52.33*	53.55*

D. Comparisons With Related Methods

In Table IV, we present AUC (%) performance of SOTA methods on UCF-Crime, CADP, and MP-RAD datasets. Results reveal that our baseline model with $\eta = 1$ (as UCF-Crime and CADP do not provide multi-view data) performs slightly poorer than the method proposed by Sultani et al [4]. The method proposed by Feng et al [30] significantly outperforms all baselines when trained and tested on UCF-Crime and CADP datasets. However, the proposed method outperforms all baselines including MIST [30] and MIL [4] when trained and tested on the MP-RAD dataset. This happens due to the exploitation of multi-view inputs by the proposed method.

Our method improves AUC by 12%, 13%, 14%, and 20% as compared to MIST [30], GCN [5], Ordinal Regression [47],

TABLE IV

QUANTITATIVE COMPARISONS WITH EXISTING ACCIDENT DETECTION METHODS. (*)-TESTED USING BASELINE MODEL AS OUR METHOD IS DESIGNED FOR MULTI-VIEW INPUTS AND UCF-CRIME AND CADP DATASETS CONTAIN SINGLE VIEW ACCIDENT VIDEOS. EACH AUC (%) COLUMN REPRESENTS THE TESTING PERFORMANCE OF DETECTION METHOD TRAINED AND TESTED ON SAME DATASET

		AUC (%)		
Method	**Features**	**UCF-Crime**	**CADP**	**MP-RAD**
Hasan *et al.* [44]	AE^{RGB}	50.02	52.11	52.80
Lu *et al.* [45]	Dictionary	51.49	51	53.81
Sadek *et al.* [25]	HFG	51.30	51.41	55.20
Yun *et al.* [46]	MIF	50.90	53.33	55.35
Sultani *et al.* [4]	$C3D^{RGB}$	52.08	55.16	59.46
	$I3D^{RGB}$	56.39	57.59	57.60
Singh *et al.* [24]	AE^{RGB}	57.22	64.81	60.37
Zhong *et al.* [5]	$C3D^{RGB}$	57.35	58.01	61.15
	TSN^{RGB}	62.59	65.88	64.30
	$TSN^{OpticalFlow}$	59.21	60.35	62.28
Tian *et al.* [31]	$I3D^{RGB}$	60.03	59.26	58.05
Peng Wu *et al.* [33]	$C3D^{RGB}$	51.43	54.57	55.70
Pang *et al.* [47]	$C3D^{RGB}$	63.29	64.72	60.09
	$I3D^{RGB}$	66.55	64.20	63.81
Feng *et al.* [30]	$C3D^{RGB}$	66.71	64.67	62.29
	$I3D^{RGB}$	**69.70**	**66.37**	65.13
Ours	$C3D^{RGB}$	51.30*	53.26*	**74.69**
	$I3D^{RGB}$	55.76*	56.08*	**77.25**

TABLE V

QUANTITATIVE COMPARISONS ON MP-RAD DATASET USING MODELS PRE-TRAINED ON REAL-WORLD DATASETS AND FINE-TUNED ON MP-RAD DATASET

		AUC(%)			
		UCF-Crime		CADP	
Method	Features	pretrained	fine-tuned	pretrained	fine-tuned
Sultani *et al.* [4]	$C3D^{RGB}$	52.03	53.59	53.39	55.12
	$I3D^{RGB}$	54.17	55.46	55.60	56.83
Pang *et al.* [47]	$C3D^{RGB}$	54.48	57.81	57.20	59.12
	$I3D^{RGB}$	54.57	58.13	57.78	60.51
Feng *et al.* [30]	$C3D^{RGB}$	60.21	62.81	62.85	64.20
	$I3D^{RGB}$	60.71	63.26	68.12	69.58

MIL [4], respectively. The results reveal that multiple-instance-pseudo-label-generator along with attention module presented in MIST [30] works well for fixed-view inputs. However, when the method is trained and tested on MP-RAD dataset, even the TSN^{RGB} stream of GCN [5] performs equally well as compared to MIST [30] and Ordinal Regression [47]. Fig. 7 shows some results obtained using the proposed method on three datasets when trained on MP-RAD dataset. It may be observed that the method works well on all cases. We now show the effect of fine tuning the pretrained models of a few baselines using MP-RAD dataset. The results presented in Table V suggest that when the models are fine-tuned using MP-RAD, the performance improves.

E. Qualitative Analysis of Results

Further to evaluate the performance of the proposed method on MP-RAD and real-world datasets, we have prepared visualization of the results. Fig. 7 depicts temporal localization of accidents in testing videos. We have shown three examples from each of the datasets. Results (a-c) are obtained from testing videos of the MP-RAD dataset. Detection results highly agree with ground truths while generating lower confidence scores for non-anomalous frames. Similarly, results (d-f) and (g-i) are generated on testing videos taken from the CADP and UCF-Crime datasets. Moreover, third column results (c, f, and i) present each dataset's poor detection performance. In (c), our method has detected accidents with a relatively low

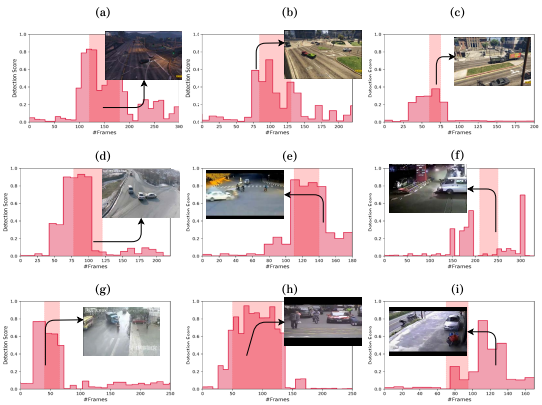

Fig. 7. Visualization of results of the proposed method. (a-c) Results using videos from the MP-RAD dataset, (d-f) present results using the CADP dataset, and (g-i) depict the results using UCF-Crime dataset, where X-axis denotes the frame number, Y-axis denotes the detection score. Highlighted (in red) portions are ground truths. The corresponding images show the anomalous frames in the input video.

Fig. 8. Class Activation Mapping (CAM) for the accident class performed to visualize the spatial attention of the model on testing videos from MP-RAD, CADP and UCF-Crime dataset. (better viewed in color).

score (0.4) because the impact of the collision was minimal. For the instance (h), there is no typical accident; however, one car lost control and collided with a nearby parked vehicle. Our method has detected other occurrences of accidents during this collision. For the instance (i), the actual accident occurred outside the video and post-accident scene captured at the corner of the windows. However, as soon as it became visible, our method has generated a higher score for the anomalous frames.

We now show the visualization of the spatial activation map using Grad-CAM [48] for attention analysis. As shown in Fig. 8, our model is able to produce hard attention to the most sensitive region of the video during accidents. This certainly helps to classify frames contains the accident event. This also verifies that, despite the multi-input scenario, our framework extracts similar feature across five angles and learns discriminating abnormal patterns.

F. Ablation Study

We have carried out ablation experiments using I3D feature to show the effectiveness of the model. Since we have three modules: (i) feature similarity-guided viewpoint partitioning (ii) feature extraction, and (iii) global pooling layers applied at the end of ϕ_H^χ and Ω_L^χ followed by the classifier, we have employed various changes in these modules to study their effectiveness. After removing the viewpoint partitioning strategy, ϕ_H^χ and Ω_L^χ are merged to the series of κ number of I3D networks. Moreover, we have replaced the proposed weighted pooling by well-known average pooling.

TABLE VI

ABLATION EXPERIMENTS ON MP-RAD DATASET. (VIEWPOINT PARTITIONING) = RANK-BASED ARRANGEMENT OF THE VIEWPOINTS FOR THE BETTER FEATURE EXTRACTION. (RB-POOLING) = RANKED-BASED WEIGHTED AVERAGE POOLING PERFORMED AFTER FEATURE EXTRACTION DONE BY THE ϕ_H^{χ}

κ	Viewpoint Partitioning	RB-Pooling	AUC (%)
2	-	-	55.75
	✓	-	56.81
	-	✓	56.08
	✓	✓	61.83
3	-	-	55.75
	✓	-	67.70
	-	✓	65.96
	✓	✓	**77.25**
4	-	-	55.75
	✓	-	70.65
	-	✓	68.21
	✓	✓	72.55

TABLE VII

EFFECT OF κ ON THE PERFORMANCE

		AUC (%)	
Features	κ	UCF-Crime	CADP
C3D/I3D	1	51.23/52.40	52.19/53.42
C3D/I3D	2	53.37/55.69	53.98/55.14
C3D/I3D	3	**60.36/61.60**	**58.09/59.81**
C3D/I3D	4	59.50/60.03	57.03/58.70
C3D/I3D	5	57.70/59.11	55.49/57.59

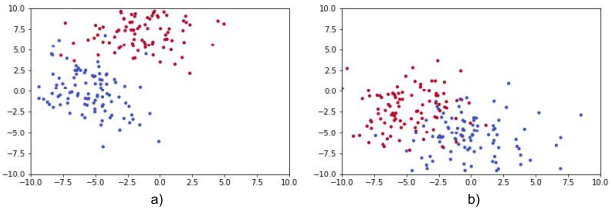

Fig. 9. (a) Feature space visualization of the proposed trained model with viewpoint partitioning and pooling versus; (b) Feature space visualization of baseline. Red data points are anomalous regions and the blue data points represent normal regions.

Table VI presents the ablation results of the MP-RAD dataset. Results presented in Table VII reveal that $\kappa = 3$ configured with viewpoint partitioning and weighted average pooling employed at the end of ϕ_H^{χ} performs the best.

However, for the same configurations with the absence of weighted average pooling, performance of the method degrades by 10%. A similar performance degradation can be observed when κ is varied. On the other hand, viewpoint partitioning based on feature similarity score helps the model to learn robust features as κ increases from 2 to 4. The results shown in Table VI illustrates that κ, viewpoint partitioning and weighted average pooling play significant role in detecting accidents. Moreover, we also visualize the feature space using t-SNE [49] of the trained model with the inclusion of weighted average pooling and viewpoint partitioning versus the baseline. Fig. 9 indicates that the trained model achieves more discriminating features representation than the baseline model.

G. Viewpoint Variation Experiments

To study the relationship between the camera viewpoint and the scene, we have explored two questions: (i) *Does angular variation affects the final decision?* (ii) *How the proposed method responds to a large angular variation vs. small angular variations?* To study this, 25 accident events have been captured using low as well as high camera angle variations. Table VIII presents the results of this study. It may be observed that a lower variation ensures higher feature similarity, resulting significant performance improvement. As the viewpoints are closer, the proposed partitioning and weighted average pooling work well. These two modules ensure that similar set of features get extracted when higher inter-viewpoint correlations exist. Additional results, dataset samples can be found in supplementary files.

TABLE VIII

THE PERFORMANCE OF THE PROPOSED METHOD IN TERMS OF AUC (%) ON LOW-VARIATION AND HIGH-VARIATION DATASET PARTITION

κ	Low Variation	High Variation
2	68.17	66.81
3	**79.10**	**74.12**
4	74.22	73.20
5	72.93	70.56

H. Discussions

The results reveal that our model when trained on synthetic data can detect real-world accidents. The AUC (%) comparisons and dataset cross-validation experiment results shown in Table II and Table IV support this claim. The visuals shown in Fig. 7 indicate that the model generates high detection scores for the frames containing an accident event and lower scores for normal frames. It has also achieved AUC of 77.25% on MP-RAD dataset by exploiting multi-perspective nature of the input videos. The same can be confirmed by visual results depicted in Fig. 8 and Fig. 9. Fig. 8 depicts spatial attention that is specifically highlighting the accident region in the video. Fig. 9 depicts higher discriminating capability of the model by plotting feature space using t-SNE. Extensive experiments have revealed that the proposed method (i) can detect real-world accident even if the model is purely trained on synthetic accident videos, (ii) achieves state-of-the-art performance by exploiting multi-perspective nature of the input videos as compared to the latest accident detection methods such as [3], [4], [5], and [24]. Moreover, our proposed dataset plays a key role in achieving better results. The multi-perspective nature of the inputs helps to understand the accident events from five different perspectives, resulting extraction of robust features that describe an accident event. The AUC (%) results shown in Table VII reveal that the multi-perspective inputs (e.g. with higher κ) certainly help to detect accidents in a better way. More importantly, all accident detection methods [3], [4], [5], [7], [24] including ours that are purely trained on synthetic dataset can be further fine-tuned on real-world data to produce more accurate detection results. Such a fine-tuned method can boost the recognition performance by significant margins. AUC (%) performance comparisons of the recent methods with pre-trained and fine-tuned model are shown in Table VII.

However, the work presented in this paper has a few limitations and scopes for improvement: (i) In the proposed

MP-RAD dataset, the camera positions are random and not recorded. Thus, the dataset cannot be used to understand the relation of the scene w.r.t. camera setups. (ii) The number of viewpoints are limited to five, which may not be sufficient to study the effectiveness of multi-view inputs. (iii) Our method relies on multiple view points of an event. Therefore, it is necessary to benchmark the method with multi-view datasets. (iv) We have not included FAR-based comparisons. This can be carried out in future.

VI. Conclusion

We have proposed a new accident detection method that exploits feature similarity across viewpoints. The viewpoint partitioning provides a stronger base for feature similarity exploitation. Moreover, the weighted average pooling ensures that higher feature similarity gets more importance. AUC (%) comparisons with baselines reveal that the proposed method performs better when multi-view inputs are used in training. Moreover, sub-optimal performance of the baselines on MP-RAD suggests that the dataset is equally challenging. We also introduce MP-RAD dataset that is large-scale, fully temporally annotated, and synthetically generated using a gaming platform. Our primary objective is to show that accident detection accuracy can be improved using multi-view synthetic dataset. This will certainly help the ITS research community to test and validate accident detection related research works.

References

[1] A. Shah, J. Baptiste Lamare, T. Nguyen Anh, and A. Hauptmann, "CADP: A novel dataset for CCTV traffic camera based accident analysis," 2018, *arXiv:1809.05782*.

[2] S. R. E. Datondji, Y. Dupuis, P. Subirats, and P. Vasseur, "A survey of vision-based traffic monitoring of road intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 10, pp. 2681–2698, Oct. 2016.

[3] X. Huang, P. He, A. Rangarajan, and S. Ranka, "Intelligent intersection: Two-stream convolutional networks for real-time near-accident detection in traffic video," *ACM Trans. Spatial Algorithms Syst.*, vol. 6, no. 2, pp. 1–28, Jan. 2020.

[4] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6479–6488.

[5] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1237–1246.

[6] J. X. Zhong, N. Li, W. Kong, T. Zhang, T. H. Li, and G. Li, "Step-by-step erasion, one-by-one collection: A weakly supervised temporal action detector," in *Proc. 26th ACM Int. Conf. Multimedia (ACMM)*, 2018, pp. 35–44.

[7] Y. Zhu and S. D. Newsam, "Motion-aware feature for improved video anomaly detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2019, pp. 270–282.

[8] G. Wang, X. Yuan, A. Zheng, H.-M. Hsu, and J.-N. Hwang, "Anomaly candidate identification and starting time estimation of vehicles from traffic videos," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 382–390.

[9] J. Zhang, L. Qing, and J. Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4030–4034.

[10] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 273–280.

[11] O. Kopuklu, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 91–100.

[12] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 900–918, Mar. 2019.

[13] X. Wang, X. Ma, and W. E. L. Grimson, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 539–555, Mar. 2009.

[14] *Traffic Analysis Tools: Next Generation Simulation—FHWA Operations*. Accessed: Nov. 17, 2021. [Online]. Available: https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm

[15] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2287–2301, Nov. 2011.

[16] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *Int. J. Comput. Vis.*, vol. 98, no. 3, pp. 303–323, Jul. 2012.

[17] *Institut Fuer Algorithmen und Kognitive Systeme*. Accessed: Nov. 17, 2021. [Online]. Available: http://i21www.ira.uka.de/image_sequences/

[18] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*, Jun. 2011, pp. 3169–3176.

[19] L. Wang et al., "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[20] H. Meng, X. Wang, and X. Wang, "Expressway crash prediction based on traffic big data," in *Proc. Int. Conf. Signal Process. Mach. Learn. (SPML)*, 2018, pp. 11–16, doi: 10.1145/3297067.3297093.

[21] P. Ahmadi, M. Tabandeh, and I. Gholampour, "Abnormal event detection and localisation in traffic videos based on group sparse topical coding," *IET Image Process.*, vol. 10, no. 3, pp. 235–246, Feb. 2016.

[22] G. Liang, "Automatic traffic accident detection based on the Internet of Things and support vector machine," *Int. J. Smart Home*, vol. 9, no. 4, pp. 97–106, Apr. 2015.

[23] J. Ren, Y. Chen, L. Xin, J. Shi, B. Li, and Y. Liu, "Detecting and positioning of traffic incidents via video-based analysis of traffic states in a road segment," *IET Intell. Transp. Syst.*, vol. 10, no. 6, pp. 428–437, Aug. 2016.

[24] D. Singh and C. K. Mohan, "Deep spatio-temporal representation for detection of road accidents using stacked autoencoder," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 879–887, Mar. 2018.

[25] S. Sadeky, A. Al-Hamadiy, B. Michaelisy, and U. Sayed, "Real-time automatic traffic accident recognition using HFG," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3348–3351.

[26] Z. Hui, X. Yaohua, M. Lu, and F. Jiansheng, "Vision-based real-time traffic accident detection," in *Proc. 11th World Congr. Intell. Control Autom.*, Jun. 2014, pp. 1035–1038.

[27] T.-N. Le, S. Ono, A. Sugimoto, and H. Kawasaki, "Attention R-CNN for accident detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 313–320.

[28] Z. Zhang, Q. He, J. Gao, and M. Ni, "A deep learning approach for detecting traffic accidents from social media data," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 580–596, Jan. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X1730356X

[29] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[30] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "MIST: Multiple instance self-training framework for video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14009–14018.

[31] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021.

[32] J.-P. Jodoin, G.-A. Bilodeau, and N. Saunier, "Urban tracker: Multiple object tracking in urban mixed traffic," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2014, pp. 885–892.

[33] P. Wu et al., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 322–339.

[34] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2633–2642.

[35] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and Y. Kompatsiaris, "ViSiL: Fine-grained spatio-temporal video similarity learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019.
[36] L. Baraldi, M. Douze, R. Cucchiara, and H. Jegou, "LAMV: Learning to align and match videos with kernelized temporal layers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7804–7813.
[37] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
[38] C. G. Kun Liu and W. Liu, "T-C3D: Temporal convolutional 3D network for real-time action recognition," in *Proc. Conf. Artif. Intell. (AAAI)*, 2018, pp. 7138–7145.
[39] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.
[40] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Proc. Asian Conf. Comput. Vis. (ACCV)*. Taipei, Taiwan: Springer, 2016, pp. 136–153.
[41] R. Herzig et al., "Spatio-temporal action graph networks," in *Proc. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Oct. 2019, pp. 2347–2356.
[42] Y. Yao, X. Wang, M. Xu, Z. Pu, E. Atkins, and D. Crandall, "When, where, and what? A new dataset for anomaly detection in driving videos," 2020, *arXiv:2004.03044*.
[43] R. Anil, V. Gupta, T. Koren, and Y. Singer, "Memory efficient adaptive optimization," in *Proc. Conf. Neural Inf. Process. Syst.*, 2019, pp. 1–10.
[44] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.
[45] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in Matlab," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720–2727.
[46] K. Yun, H. Jeong, K. M. Yi, S. W. Kim, and J. Y. Choi, "Motion interaction field for accident detection in traffic surveillance video," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 3062–3067.
[47] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12170–12179.
[48] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[49] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Thakare Kamalakar Vijay (Graduate Student Member, IEEE) received the B.E. degree in computer science and engineering from the Government College of Engineering and Research, Awasari, Pune, in 2016, and the M.Tech. degree in computer science and engineering from the National Institute of Technology, Nagpur, in 2018. He is currently a Research Scholar with the School of Electrical Sciences, Indian Institute of Technology Bhubaneswar, Bhubaneswar. His research interests include computer vision, video-surveillance, image processing, deep learning, and pattern recognition.

Debi Prosad Dogra (Member, IEEE) received the B.Tech. degree in computer science and engineering from HIT Haldia in 2001, the M.Tech. degree in computer science and engineering from IIT Kanpur in 2003, and the Ph.D. degree in computer science and engineering from IIT Kharagpur in 2012. He is currently an Assistant Professor with the School of Electrical Sciences, IIT Bhubaneswar, India. He has published more than 100 research papers in international journals and conferences in the areas of computer vision, image segmentation, and healthcare analysis.

Heeseung Choi (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2004, 2006, and 2011, respectively. He was a Research Member of the Biometrics Engineering Research Center (BERC), South Korea, from 2004 to 2011. He was a Post-Doctoral Fellow with the Department of Computer Science and Engineering, Michigan State University, from 2011 to 2012. He is currently a Senior Research Scientist at the Center for Artificial Intelligence (CAI), Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology (KIST). His research interests include computer vision, biometrics, image processing, deep learning, forensic science, and pattern recognition.

Gipyo Nam received the B.S. degree in digital media technology from Sangmyung University, Seoul, South Korea, in 2009, and the Ph.D. degree in electronics and electrical engineering from Dongguk University in 2014. He is currently a Senior Research Scientist at the Center for Artificial Intelligence (CAI), Artificial Intelligence and Robotics Institute, Korea Institute Science and Technology (KIST), Seoul. His research interests include pattern recognition, biometrics, and image processing.

Ig-Jae Kim (Member, IEEE) received the B.S. and M.S. degrees in EE from Yonsei University, Seoul, South Korea, in 1996 and 1998, respectively, and the Ph.D. degree in EECS from Seoul National University in 2009. He is currently the Director of the Artificial Intelligence and Robotics Institute, Korea Institute of Science and Technology (KIST), Seoul. He is also an Associate Professor at the Korea University of Science and Technology and a Guest Professor at Korea University. He had worked with the Massachusetts Institute of Technology (MIT) Media Laboratory as a Post-Doctoral Researcher (2009–2010). He has published over 100 fully-refereed papers in international journals and conferences, including *ACM Transaction on Graphics*, *Pattern Recognition*, CVPR, SIGGRAPH, and Eurographics. He is interested in pattern recognition, computer vision and graphics, deep learning, and computational photography.