Traffic Anomaly Detection and Video Summarization Using Spatio-Temporal Rough Fuzzy Granulation With Z-Numbers

Anima Pramanik, Sankar Kumar Pal, *Life Fellow, IEEE*, Jhareswar Maiti, *Member, IEEE*, and Pabitra Mitra, *Senior Member, IEEE*

***Abstract*— Existing traffic video summarization algorithms are capable of detecting one-class (i.e., collision) anomaly and cannot handle uncertainty issues arising between two-class anomalies, such as collision and near-miss. To address the issues, a new video summarization algorithm, namely Z-numbers-based spatio-temporal rough fuzzy granulation (Z-STRFG) is developed. In Z-STRFG, various spatio-temporal features are computed over the video frames and used for obtaining the approximate anomaly-prone regions in terms of granules. In these regions, uncertainty (i.e., fuzziness) may arise among three scenarios, namely collision, near-miss, and normal traffic. Therefore, two types of rough fuzzy granules (RFGs) along with their roughness scores are computed to distinguish the aforesaid three scenarios. For each RFG, Z-number is computed based on the membership value of its roughness score to ensure a higher degree of reliability in the detection of anomaly class. Aforesaid characteristics of Z-STRFG improve its speed and accuracy for traffic anomaly detection. The efficacy of Z-STRFG has been demonstrated over 130 real-time traffic videos containing collisions, near-misses, and normal traffics. The superiority of Z-STRFG over some state-of-the-art is also proved through extensive experiments.**

***Index Terms*— Traffic anomaly detection, video summarization, spatio-temporal features, rough fuzzy granules, Z-numbers.**

I. Introduction

TRANSPORTATION system plays a significant role in the movement of people and goods from one place to another place. However, it poises a serious concern due to the occurrences of collisions. Global road crash statistics reports show 1.3 million people die annually due to collisions [1]. In addition to investigating the collision, it has been recognized from the statistics reported in [2] that the investigation of near-miss is also important in reducing the probability of occurrences of collision, thereby improving the road safety. Nowadays, it is a common practice to use computer vision techniques for modeling and analyzing various traffic scenarios from surveillance videos. Cameras are installed on the road-side for monitoring, analyzing, and capturing the every minute details of traffic. As enormous data (e.g., video stream) is collected by the camera, it may be laborious and time consuming to scrutinize and identify the anomaly contents present within it. Therefore, it is essential to reduce the unwanted contents from the video-stream to make it more precise. This can be achieved by video summarization.

Video summarization is an economical way of representing the video clips consisting of relevant activities [3]. An optimal video summary should have two properties, such as appropriate representation of the desirable events and heterogeneity for ensuring the least redundancy. As it is useful for effective and quick browsing of the relevant activities from the video, it can extensively reduce the human efforts for finding the informative video contents. It has revealed demonstrated success in various domains, including remote investigation, medical diagnosis, and accident analysis [4]. Mostly, video summarization algorithms [5], [6] use the entire frame as a feature map for anomaly detection, thereby increasing the computational cost. Some improved algorithms, such as perceptual video summarization (PVS) [4] and video synopsis using stereo camera [7] can speed up the process of anomaly detection by considering the object-level features only. Aforesaid algorithms are capable of detecting collision only and cannot handle the uncertainty issue arising between collision and near-miss. Moreover, there remain a few challenges in real-time applications due to the occlusion issue. Deep convolutional neural networks (CNNs) [8] are useful for detecting objects in complex scenario and hence, object level features. Since the classification accuracy of deep CNN model is restricted to the classes of the training samples, some desirable regions i.e., foreground objects (FOs) with unknown classes may not be detected, thereby affecting the anomaly detection accuracy.

In order to address all the issues, an optimization algorithm is proposed based on the concepts of spatio-temporal features (STFs), rough fuzzy set (RFS), and Z-numbers for collision and near-miss detection and video summarization. Here, object

Manuscript received 23 November 2021; revised 5 May 2022 and 25 July 2022; accepted 4 August 2022. Date of publication 31 August 2022; date of current version 5 December 2022. The Associate Editor for this article was W. Lin. *(Corresponding author: Anima Pramanik.)*

Anima Pramanik and Jhareswar Maiti are with the Department of Industrial and Systems Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India (e-mail: apramanik17@gmail.com).

Sankar Kumar Pal is with the Center for Soft Computing Research, Indian Statistical Institute, Kolkata 700108, India.

Pabitra Mitra is with the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India.

This article has supplementary downloadable material available at https://doi.org/10.1109/TITS.2022.3198595, provided by the authors.

Digital Object Identifier 10.1109/TITS.2022.3198595

detection is done over the frames primarily to extract the foreground regions and to reduce the working area. This task is done in unsupervised way, as we aim to correct the localization of FOs. Then, the motion state of traffic flow (consisting of FOs) can be taken care of using STFs, and uncertainty issue can be handled using the concepts of RFS and Z-numbers. Initially, using salient STFs, approximate anomaly-prone regions are obtained over the video frames. The probability of occurrence of the event in these regions is higher than those belonging to any other regions. These regions are called granules, which are formed using the process, called granulation [8]. A granule is defined as a cluster of indiscernible contents drawn together. The formation of clusters within an image is called image segmentation [9], which is the primary task of various detections, such as object detection, anomaly detection, and intrusion.

Performing optimum image segmentation is effective in finding regions of interest (RoIs) for anomaly detection. Fuzzy geometry [10], histogram thresholding [9], and hierarchical syntactic approach [11] are the effective ways of image segmentation. In [8], it was demonstrated how effectively STFs-based granulation was more effective than only spatial or temporal features-based granulation for extracting the RoIs (i.e., segmented image) in terms of object localization. Therefore, in this study, STFs-based granules are formed over the video frames to extract the approximate anomaly-prone regions in terms of RoIs. These regions may belong to either collision, near-miss, or normal traffic. No definite classes can be assigned to these regions as there are an underlying fuzziness (i.e., uncertainty). Therefore, these granules are called fuzzy granules (FGs). The said uncertainty has been modeled by defining rough fuzzy granules (RFGs) based on the concept of RFS. By definition, a RFS [12] is the approximations of a fuzzy set in a crisp approximation space. Whereas, a fuzzy rough set [12] is the approximations of a crisp set in a fuzzy approximation space. Rough-fuzzy hybridization was first envisioned in [13] for developing efficient decision-making systems with real-life applications. This integration aims to combine judiciously the merits of fuzzy sets in handling uncertainties arising from overlapping regions or classes, and that of rough set in dealing with the uncertainties due to granularity in the domain of discourse. It, therefore, provides a stronger paradigm for handling uncertainties.

RFGs, thus formed over the FGs, are classified as either collision, near-miss, or normal traffic through minimization of the roughness score. To further obtain the reliability of the detected anomaly class for an RFG, the concept of Z-numbers is used. Z number [14] consists of two tuples. First tuple is the probability of occurrence of an event. Whereas, the second tuple is the reliability of occurrence of the same event, and it is defined using the linguistic terms (LTs). A recent study [15] reveals how the concept of Z-numbers can be used in video processing to explain a scene with certainty in the form of natural language. In this study, Z-number for an RFG is computed using the membership value (MV) of its roughness score. Based on this MV, LTs are assigned to the second tuple of Z-number by the users. Therefore, it may contains a degree of uncertainty. Thus, it is required to measure the uncertainty in Z-number for the correct classification of an anomaly. Three rules are defined using the information of RFGs and Z-numbers for classifying the event as collision, near-miss, or normal traffic. All these tasks constitute an optimization algorithm, named as Z-numbers-based spatio-temporal rough fuzzy granulation (Z-STRFG). The number of FGs in a frame is less than the number of FOs present in the same frame. For any frame, only FGs are used in Z-STRFG for anomaly detection, thereby reducing the computational cost. Moreover, RFGs can handle uncertainty occurred in real-time scenarios, thereby enhancing the detection accuracy.

The effectiveness of Z-STRFG along with various comparisons has been demonstrated extensively over 130 traffic videos acquired from YouTube8M [4], Urban Tracker [4], and a plant in India. Plant video data is a new real-life traffic data, not available in on-line, and is named as Anomaly20. Based on the aforesaid discussion, innovation/contributions of the study are summarized as:

(i) A new video summarization algorithm, namely Z-STRFG is developed for detecting two types of traffic anomalies, viz, collision and near-miss, from surveillance videos using the formation of STFs and RFGs, and Z-numbers-based linguistic description.
(ii) Various STFs are defined over the video frames for representing the approximate anomaly-prone regions.
(iii) RFGs are formed over the said regions using the STFs in enabling efficient detection of collision and near-miss events through the minimization of roughness scores.
(iv) The use of Z-number for each RFG in ensuring the reliability of the detected anomaly is unique.
(v) A real-time traffic data, namely Anomaly20 is prepared based on traffic videos acquired from a plant in India.

The article proceeds as follows. Related works are reported in Section II. The basic theory of RFS is stated in Section III. Z-STRFG is illustrated in Section IV. Results are reported in Section V. Finally, we conclude this study in Section VI.

II. Related Works

A. Video Summarization

Video summarization algorithms are broadly categorized into two classes: static image-based summarization (containing optimal frames) [16] and dynamic video-based summarization (containing video clip) [4]. Summarization problem is highly dependent on the subjectivity. Local patch learning and blob sequence-based optimization is done in [17] for video summarization and anomaly detection. Ribbon carving approach-based optimization is presented in [18]. Clustering-based object tracking is done in [3] for summarizing the video. Sub-modular optimization is introduced in [19]. Here, unconstrained maximization of non-monotonic and non-negative sub-modular function is used as an optimization algorithm. Although aforesaid studies show the advancements in video summarization for simple traffic; however, they are restricted to the detection of collision.

B. Traffic Anomaly Detection

A detailed review on collision detection at road intersection is presented in [20]. Structural similarity-based approach is proposed in [21] for collision detection. This method fails to classify vehicles far from a surveillance camera. The symmetry property of the motion intersection field between two vehicles is used in [6]. The vehicle trajectory is analyzed in [22] for collision detection. A feature-fusion deep learning framework is developed in [23] for collision detection. Another important traffic anomaly is near-miss. Vehicle alignment, speed, and position are used for near-miss detection. A two stream network is developed in [24] for near-miss detection. The deep networks are restricted to the classes of the training samples. By definition, near-miss is the close proximity between two objects having no abrupt change in their trajectories. Whereas, for collision, an abrupt change in the motion flow is found. It creates uncertainty between collision and near-miss. Various well-known RFS-based granular computing methods, such as rough and kernel sets-based STFs [25] and fuzzy rough granules [26] are developed for handling the uncertainty issue. These methods are restricted to the one-class outliers and the use of numerical, categorical, or mixed data types. As per the author's knowledge, to date, no studies have been carried out to handle the said uncertainty issue.

Aforesaid issues of video summarization and traffic anomaly detection are addressed in this study, where a new video summarization algorithm, namely Z-STRFG is developed by incorporating the concepts of RFS and Z-numbers, as described in Section IV.

III. Preliminaries

In this section, the basic theory of RFS is presented. RFS is a generalization of the rough set, derived from the approximation of a fuzzy set defined over a crisp approximation space [12]. Let U and V be two finite non-empty universes of discourse and $R \in Q(U \times V)$ a binary relation from U to V. The ordered triple (U, V, R) is called a two universes approximation space. For any fuzzy set $Y \in F(V)$, the lower and upper approximations of Y, namely $\underline{R}(Y)$ and $\overline{R}(Y)$ concerning the approximation space (i.e., fuzzy sets) of U (for each $x \in U$), are defined as

$$\underline{R}(Y)(x) = \cap\{Y(y)|y \in F(x)\} \tag{1}$$

$$\overline{R}(Y)(x) = \cup\{Y(y)|y \in F(x)\} \tag{2}$$

where $F(x)$ represents the successor neighborhood of x for R and it is defined as

$$F : U \to Q(V), F(x) = \{y \in V : (x, y) \in R\} \tag{3}$$

where F represents the set-valued mapping from U to $Q(V)$. The ordered set-pair $(\underline{R}(Y), \overline{R}(Y))$ is referred to as a rough fuzzy set, and $\underline{R}(Y)$ and $\overline{R}(Y)$ are referred to as lower and upper rough fuzzy approximation operators, respectively.

IV. Proposed Z-STRFG

The proposed Z-STRFG for video summarization and traffic anomaly (i.e. collision and near-miss) detection is discussed in this section. An overview of Z-STRFG is shown in Fig. 1. It is seen from Fig. 1, initially, STFs are obtained from the consecutive P number of input frames, as explained in Section IV-A. These features are fed to the optimization algorithm to obtain FGs, which are further used to generate RFGs for anomaly detection, as explained in Section IV-B. To ensure the reliability of the detected anomaly class, Z-number is computed, as stated in Section IV-C.

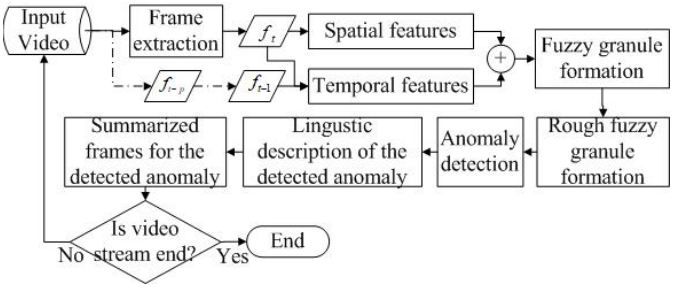


Fig. 1. Z-STRFG.

A. Salient Spatio-Temporal Features (STFs)

As said earlier, various STFs corresponding to each detected FO are defined and used for obtaining the FGs over the video frames. Initially, spatial features, namely distance, total area, and common area between two FOs, and positions of their centroids are computed in each frame of a video. Then, the change of the aforesaid spatial features is computed in the temporal domain through consecutive frames of the video for obtaining the STFs. Let, f_t be the current frame and $\{f_{t-i}\}_{i=1}^{P}$ be its previous $P(= 3)$ frames. First, FOs are detected over these frames using the Mixture of Gaussian, and fitted by bounding-boxes. Let a_i^t be the i^{th} FO in f_t, and $a_i^t(x, y)$ be the spatial co-ordinate of the centroid for a_i^t. For each detected FO, various salient features, namely speed saliency, neighbor-distance saliency, intersection over union (IoU) saliency, and area saliency are defined. Characteristics of these features are defined as follows.

1) Speed Saliency (SS): SS refers to the weighted average speed change of an object over frames. Let $SS(a_i^t)$ be the speed saliency for a_i^t. The $SS(a_i^t)$ is defined as

$$SS(a_i^t) = \frac{\sum_{j=1}^{P} |a_i^t(x, y) - a_i^{t-j}(x, y)| \cdot \exp^{-j^2}}{\sum_{j=1}^{P} j \cdot \exp^{-j^2}} \tag{4}$$

where $|a_i^t(x, y) - a_i^{t-j}(x, y)|$ defines the position change of a_i^t from f_{t-j} to f_t. $\exp^{-j^2}$ represents a non-negative weight associated with time and is sensitive to immediate frame.

2) Neighbor-Distance Saliency (NDS): NDS measures the minimum change in nearness for a FO in any frame with respect to other FOs present in the same frame. Let n be the number of FOs present in f_t, and $dis_{i,j}^t$ be the distance between centroids of a_i^t and a_j^t. Let $NDS(a_i^t)$ be the neighbor-distance saliency of a_i^t, defined as

$$NDS(a_i^t) = \wedge \bigcup_{j \in n}^{j \neq i} \frac{dis_{i,j}^t \cdot \exp^{-\sqrt{dis_{i,j}^t}}}{(w_i^t + w_j^t) \cdot \sum_j \exp^{-\sqrt{dis_{i,j}^t}}} \tag{5}$$

where $\wedge$ represents the minimum operation and w_i^t represents the width of the bounding-box fitted over a_i^t. The term $\exp^{-\sqrt{dis_{i,j}^t}}$ is a non-negative weight associated with a_i^t, and signifies that it is locally sensitive to its immediate neighbor. The term $\frac{1}{(w_i^t+w_j^t)}$ is associated with a_i^t, and signifies that it nullifies the effect of non-neighbor FOs having larger area and far away from a_i^t.

3) IoU Saliency (IoUS): IoUS measures the weighted average common area between a FO in f_t and all other FOs from f_{t-1}. Let $dis_{i,j}^{t,t-1}$, $(b_i^t \cap b_j^{t-1})$, and $(b_i^t \cup b_j^{t-1})$ be the distance, common area, and total area, respectively, between a_i^t and a_j^{t-1}. Then $IoUS$ for a_i^t is defined as

$$IoUS(a_i^t) = \frac{1}{CS(a_i^t)} \sum_{j=1}^{m} \frac{\frac{b_i^t \cap b_j^{t-1}}{b_i^t \cup b_j^{t-1}} \cdot \exp^{-\sqrt{dis_{i,j}^{t,t-1}}}}{\sum_{j=1}^{m} \exp^{-\sqrt{dis_{i,j}^{t,t-1}}}} \tag{6}$$

where m represents the number of FOs present in f_{t-1}. The term $CS(a_i^t)$ measures the number of centroids presents in b_i^t, and it is sensitive to the close proximity. The intuition behind this measurement is to find the weighted average overlapping area for each FO present in f_t.

4) Area Saliency (AS): AS measures the weighted total area between a FO in f_t and its immediate neighbors from f_{t-1}. Let $AS(a_i^t)$ be the area saliency of a_i^t, defined as

$$AS(a_i^t) = \wedge \bigcup_{j \in m_1, m_1 \leq m}^{a_i^t \in b_i^t \cap b_j^{t-1}} (b_i^t \cup b_j^{t-1}) \cdot \exp^{-dis_{i,j}^{t,t-1}} \tag{7}$$

The intuition behind this measurement is to find the minimum weighted total area between a FO in f_t and its immediate neighbors from f_{t-1}.

Here, SS, $IoUS$, and AS are the STFs, and NDS is the spatial feature. Aforesaid features are used in the optimization algorithm for traffic anomaly detection and video summarization, as explained in the next section.

B. Optimization Algorithm

The defined features, such as NDS, SS, IoUS, and AS are fed to the optimization algorithm for generating the approximate anomaly-prone regions in terms of FGs, as there is an underlying uncertainty among collision, near-miss, and normal traffic. The said uncertainty has been modeled by defining RFGs, as it is reputed in taking care of the uncertainty issues arising from the overlapping region(s). The granules are introduced based on the problem definition used in the optimization algorithm. Optimal video summarization is a maximization problem, where the perceptual quality of the summarized video should increases while minimizing the redundant frames. Let I and I_s be the input and summarized videos, respectively, and D be the perceptual quality of I_s. Then, maximization problem is defined as

$$\begin{aligned} &\max\ D \quad s.t.\ \ I_s \in I \\ &I_s = \left\{ \forall_l f_l \in I | \forall_{g,h} (FG_g^l \cap FG_h^l \neq 0) \right\}^{g \neq h} \end{aligned} \tag{8}$$

where $\left\{FG_1^l, FG_2^l, \ldots, FG_G^l\right\}$ defines the set of $G(g, h \in G)$ number of FGs formed over f_l, and $FG_g^l \cap FG_h^l$ defines the common area between g^{th} and h^{th} FGs. More this common area increases the perceptual quality (D) of f_l, and hence, it is added into the I_s.

As said earlier, RFGs are defined over FGs for handling uncertainty issue present in it. Frame ($\in I$) containing the lower approximation set (i.e., set of members certainty classified to the anomaly) of RFGs (i.e., $\left\{\underline{RFG}\right\}$) with minimum roughness score ($\approx$0), is considered as the optimal frame for anomaly detection. The fuzziness of FGs is best characterized by its membership function (MF), and is handled using RFGs. The formation of FGs over each frame is defined in Section IV-B.1. FGs are assigned with MVs, as defined in Section IV-B.2. Optimal frame selection using RFGs is stated in Section IV-B.3.

1) Formation of FGs: Three FG sets, namely $\{FG_1\}$, $\{FG_2\}$, and $\{FG_3\}$ are obtained over the video frames using the values of various STFs. Let $\left\{A^t\right\} = \left\{a_i^t\right\}_{i \in n}$ be the set of n number of FOs present in f_t. Formation of $\left\{FG_1^t\right\}$, $\left\{FG_2^t\right\}$, and $\left\{FG_3^t\right\}$ over f_t is defined as follows.

(i) $\left\{FG_1^t\right\}$ set extraction: $\left\{FG_1^t\right\}$ set consisting of n_1 FOs, is obtained over f_t using the values of NDS for $(P+1)$ frames, defined as

$$\left\{FG_1^t\right\} = \left\{\forall_i a_i^t | NDS(a_i^t) < T_1^t\right\}_{i \in n_1}^{n_1 \leq n} \tag{9}$$

where $T_1^t = \vee \left\{ \wedge \left\{ NDS(a_i^{t-j})_{i \in n}^{j \in P} \right\} \right\}$. Here $\vee$ represents the maximum operation. The $\left\{FG_1^t\right\}$ set is used to determine the maximum nearness between FOs and their neighbors.

(ii) $\left\{FG_2^t\right\}$ set extraction: $\left\{FG_2^t\right\}$ set consisting of n_2 FOs, is obtained over f_t using the values of SS for $(P+1)$ frames, defined as

$$\left\{FG_2^t\right\} = \left\{\forall_i a_i^t | SS(a_i^t) < T_2^t\right\}_{i \in n_2}^{n_2 \leq n} \tag{10}$$

where $T_2^t = \vee \left\{ \wedge \left\{ SS(a_i^{t-j})_{i \in n}^{j \in P} \right\} \right\}$. The $\left\{FG_2^t\right\}$ set is used to determine the maximum tolerable speed for each FO.

(iii) $\left\{FG_3^t\right\}$ set extraction: $\left\{FG_3^t\right\}$ set consisting of n_3 FOs, is obtained over f_t using the values of $IoUS$ and AS, defined as

$$\left\{FG_3^t\right\} = \left\{\forall_i a_i^t | IoUS(a_i^t) > T_3^t \& AS(a_i^t) < T_4^t\right\}_{i \in n_3}^{n_3 \leq n} \tag{11}$$

where T_3^t and T_4^t are defined as:

$$T_3^t = \begin{cases} \vee \left\{IoUS(a_i^t)\right\}_{i \in n}, & \text{if } dis_i^{t,t-1} \leq m_1 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

$$T_4^t = \begin{cases} \wedge \left\{AS(a_i^t)\right\}_{i \in n}, & \text{if } dis_i^{t,t-1} \leq m_2 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where, m_1 and m_2 are two distance thresholds, and depends on the inherent information of the input video. The $\left\{FG_3^t\right\}$ set is used to determine the maximum common area between each FO and its neighbors. Following Eqs. 9, 10, and 11, it can be concluded that

Affirmation 1: A rough estimation of u-FGs ($\left\{FG_1^t\right\}$, $\left\{FG_2^t\right\}$, $\left\{FG_3^t\right\}$) over f_t is a set of $u(\wedge(n_1, n_2, n_3))$ FOs which may belong to the anomalous event.

FGs are assigned with MVs to characterize the fuzziness present in it, stated as follows.

2) Assignment of MVs: Three FGs are assigned with MVs. Let μ_h, μ_m, and μ_l be the MVs of three subsets (high, medium, and low, respectively) for a set (F) [27]. The left trapezoidal (Ltrap), middle trapezoidal (Mtrap), and right trapezoidal (Rtrap) MFs are used to represent the three subsets. These MFs are formulated using the four thresholds, namely a, b, c, and d, as defined in Eqs. S.1, S.2, and S.3 in the supplementary material. As FGs are formed using the knowledge of STFs, the MVs are assigned to the defined features for characterizing the fuzziness. To observe the influence of a feature on the performance, three sub-systems are created by only changing the values of the feature in each system. From an exhaustive study, it was found that, for an anomalous event corresponding to a FO, the values of SS, NDS, and AS should be high and $IoUS$-value should be low. For a_i^t, three MVs, namely $\mu_h^{SS}(F)(a_i^t)$, $\mu_m^{SS}(F)(a_i^t)$, and $\mu_l^{SS}(F)(a_i^t)$ are assigned to the three sub-systems corresponding to SS, and are measured using Eqs. S.1, S.2, and S.3, respectively. In the same way, for each a_i^t, MVs are assigned to the three sub-systems corresponding to NDS, and AS. Whereas, for $IoUS$-feature corresponding to a_i^t, three MVs, namely $\mu_h^{IoUS}(F)(a_i^t)$, $\mu_m^{IoUS}(F)(a_i^t)$, and $\mu_l^{IoUS}(F)(a_i^t)$ are assigned to the three sub-systems and are measured using Eqs. S.3, S.2, and S.1, respectively. Let $\mu_{in}^{()}(F)(a_i^t)$ and $\mu_{un}^{()}(F)(a_i^t)$ be the MVs of intersection and union sets of the three sub-systems for a_i^t corresponding to the feature (say,()), and are measured using Eq. S.4 (refer to the supplementary material). Based on the MVs of defined STFs, MVs are assigned to the three FGs, as defined in the following section.

(i) MVs of $\{FG_1^t\}$: The optimization algorithm takes $\{\forall_i NDS(a_i^t)\}$ as inputs and generates two output variables $\mu_{in}^{\{FG_1^t\}}(F)$ and $\mu_{un}^{\{FG_1^t\}}(F)$ for $\{FG_1^t\}$, defined as

$$\begin{aligned}\mu_{in}^{\{FG_1^t\}}(F) &= \forall_i \mu_{in}^{NDS}(F)(a_i^t) | a_i^t \in \{FG_1^t\}\\ \mu_{un}^{\{FG_1^t\}}(F) &= \forall_i \mu_{un}^{NDS}(F)(a_i^t) | a_i^t \in \{FG_1^t\}\end{aligned} \tag{14}$$

(ii) MVs of $\{FG_2^t\}$: The optimization algorithm takes $\{\forall_i SS(a_i^t)\}$ as inputs and generates two output variables $\mu_{in}^{\{FG_2^t\}}(F)$ and $\mu_{un}^{\{FG_2^t\}}(F)$ for $\{FG_2^t\}$, defined as

$$\begin{aligned}\mu_{in}^{\{FG_2^t\}}(F) &= \forall_i \mu_{in}^{SS}(F)(a_i^t) | a_i^t \in \{FG_2^t\}\\ \mu_{un}^{\{FG_2^t\}}(F) &= \forall_i \mu_{un}^{SS}(F)(a_i^t) | a_i^t \in \{FG_2^t\}\end{aligned} \tag{15}$$

where $\mu_{in}^{SS}(F)(a_i^t)$ and $\mu_{un}^{SS}(F)(a_i^t)$ represent the MVs of intersection and union sets for a_i^t corresponding to SS.

(iii) MVs of $\{FG_3^t\}$: Optimization algorithm takes $\{\forall_i IoUS(a_i^t)\}$ and $\{\forall_i AS(a_i^t)\}$ as inputs and generates two output variables $\mu_{in}^{\{FG_3^t\}}(F)$ and $\mu_{un}^{\{FG_3^t\}}(F)$ for $\{FG_3^t\}$, defined as

$$\begin{aligned}\mu_{in}^{\{FG_3^t\}}(F) &= \wedge \left\{\forall_i \mu_{in}^{IoUS}(F)(a_i^t), \mu_{in}^{AS}(F)(a_i^t)\right\}\\ \mu_{un}^{\{FG_3^t\}}(F) &= \vee \left\{\forall_i \mu_{un}^{IoUS}(F)(a_i^t), \mu_{un}^{AS}(F)(a_i^t)\right\}\end{aligned} \tag{16}$$

where $(\mu_{in}^{IoUS}(F)(a_i^t),\ \mu_{in}^{AS}(F)(a_i^t))$ and $(\mu_{un}^{IoUS}(F)(a_i^t),\ \mu_{un}^{AS}(F)(a_i^t))$ represent the MVs of intersection and union sets for a_i^t corresponding to $IoUS$ and AS, respectively.

As $u-FG$ event set represents the approximate anomaly-prone regions, this set with MVs are used for the generation of RFGs. Three rules are defined using the lower (i.e., set of members must belong to the anomaly) and upper (i.e., set of probable members belong to the anomaly) approximation sets of RFGs to obtain the optimal frames containing either collision, near-miss, or normal traffic, stated as follows.

3) Optimal Frame Selection Using RFGs: Optimal frames contain either collision or near-miss. To differentiate collision scenario from near-miss, two RFGs, namely RFG_{co} and RFG_{nm} are defined. Given a video (I) with T frames having a set of attributes $\mu_{in}^{\{FG_1\}}(F)$, $\mu_{un}^{\{FG_1\}}(F)$, $\mu_{in}^{\{FG_2\}}(F)$, $\mu_{un}^{\{FG_2\}}(F)$, $\mu_{in}^{\{FG_3\}}(F)$, and $\mu_{un}^{\{FG_3\}}(F)$, where F is the probable event set. As said earlier, I_s is the summarized video having s number of optimal frames, and D is its perceptual quality. I_s describes the anomaly type based on the roughness score of RFG_{co} and RFG_{nm} sets. Roughness score of a set is computed using the cardinality of its lower and upper approximations. Let υ be the number of RFGs obtained in f_t. D-lower approximation of υ-RFG is the set of FGs that can be certainty classified as members of the F based on the knowledge in D. D-upper approximation of υ-RFG is the set of possible members of F. D-lower and D-upper approximation sets for RFG_{co} and RFG_{nm} over f_t are $(\underline{RFG_{co}}^t, \underline{RFG_{nm}}^t)$ and $(\overline{RFG_{co}}^t, \overline{RFG_{nm}}^t)$, respectively. This is defined in the following section.

Lemma 1: For $\upsilon - RFG_{co}$, $\{\underline{RFG_{co}}^t\}$ set and its MVs, $\mu_{\{\underline{co^t}\}}(F)$ over f_t are defined as

$$\begin{aligned}\mu_{\{\underline{co^t}\}}(F) &= \wedge \left\{\forall_{k,l}^{k\neq l} \wedge \mu_{in}^{\{FG_k^t\}}(F), \mu_{in}^{\{FG_l^t\}}(F)\right\}\\ \{\underline{RFG_{co}}^t\} &= \left\{\forall_i a_i^t | \mu_{\{\underline{co^t}\}}(F)(a_i^t) \neq 0\right\}\end{aligned} \tag{17}$$

Lemma 2: For $\upsilon - RFG_{co}$, $\{\overline{RFG_{co}}^t\}$ set and its MVs, $\mu_{\{\overline{co^t}\}}(F)$ over f_t are defined as

$$\begin{aligned}\mu_{\{\overline{co^t}\}}(F) &= \wedge \left\{\forall_{k,l}^{k\neq l} \vee \mu_{in}^{\{FG_k^t\}}(F), \mu_{in}^{\{FG_l^t\}}(F)\right\}\\ \{\overline{RFG_{co}}^t\} &= \left\{\forall_i a_i^t | \mu_{\{\overline{co^t}\}}(F)(a_i^t) \neq 0\right\}\end{aligned} \tag{18}$$

Lemma 3: For $\upsilon - RFG_{nm}$, $\{\underline{RFG_{nm}}^t\}$ set and its MVs, $\mu_{\{\underline{nm^t}\}}(F)$ over f_t are defined as

$$\begin{aligned}\mu_{\{\underline{nm^t}\}}(F) &= \vee \left\{\forall_{k,l}^{k\neq l} \wedge \mu_{in}^{\{FG_k^t\}}(F), \mu_{in}^{\{FG_l^t\}}(F)\right\}\\ \{\underline{RFG_{nm}}^t\} &= \left\{\forall_i a_i^t | \mu_{\{\underline{nm^t}\}}(F)(a_i^t) \neq 0\right\}\end{aligned} \tag{19}$$

Lemma 4: For $\upsilon - RFG_{nm}$, $\{\overline{RFG_{nm}}^t\}$ set and its MVs, $\mu_{\{\overline{nm^t}\}}(F)$ over f_t are defined as

$$\begin{aligned}\mu_{\{\overline{nm^t}\}}(F) &= \vee \left\{\forall_{k,l}^{k\neq l} \vee \mu_{in}^{\{FG_k^t\}}(F), \mu_{in}^{\{FG_l^t\}}(F)\right\}\\ \{\overline{RFG_{nm}}^t\} &= \left\{\forall_i a_i^t | \mu_{\{\overline{nm^t}\}}(F)(a_i^t) \neq 0\right\}\end{aligned} \tag{20}$$

Roughness scores (R^t) for collision ($\upsilon - RFG_{co}$) and near-miss ($\upsilon - RFG_{nm}$) sets over f_t are defined as

$$R^t_{co} = 1 - \left\{ |\underline{RFG_{co}}^t| / |\overline{RFG_{co}}^t| \right\}$$
$$R^t_{nm} = 1 - \left\{ |\underline{RFG_{nm}}^t| / |\overline{RFG_{nm}}^t| \right\} \quad (21)$$

where $|()|$ represents the cardinality of a set (say, ()). The MV of R^t for a set (say, ()) is defined as

$$\mu_{R^t_{()}}(F) = \vee \left\{ \mu_{\underline{\{()^t\}}}(F), \mu_{\overline{\{()^t\}}}(F) \right\} \quad (22)$$

In all frames of I, roughness scores for both RFG_{co} and RFG_{nm} sets are computed, wherever applicable. For any frame f_t, if $R^t_{co} = R^t_{nm}$, or $|\underline{RFG_{co}}^t| = |\underline{RFG_{nm}}^t| = 0$, then its content is classified as normal scenario. Whereas, in f_t, if $R^t_{co} < R^t_{nm}$, then, its content is classified as collision, otherwise, near-miss. Thus, f_t is included in the I_s. In this way, uncertainty among three said traffic scenarios is handled. Further, to obtain the reliability of the detected anomaly class described by the I_s, the Z-number is computed for $\langle \{RFG_{co}\}, \{RFG_{nm}\} \rangle$, as defined in the next section.

C. Z-Numbers for Ensuring the Class of the Detected Event

For a RFG, to further ensure about the detected anomaly class, Z-number is computed using the MV of its roughness score. Z-number has two tuples (H, E), where H represents the real-valued uncertain variable and E defines the reliability of H. As the traffic anomalies are classified as collision or near-miss corresponding to a RFG through the minimization of its roughness score, it is used for the computation of Z-number in order to assess the reliability of the detected anomaly class. How is the concept of Z-numbers used in this study, is shown using an example: $X = \langle Collision \rangle$: name of the event, and Y: collision must happen in the frame f_t, if $\mu_{R^t_{co}}$ is (AI, > 0.9). This evidents that the decision criteria i.e., the MV of roughness score corresponding to a RFG_{co} must be greater than 0.9 for a collision scenario is 'Absolutely Important' (AI). Then, H: Context = $\left\langle \mu_{R^t_{co}} = (AI, > 0.9) \right\rangle$, and E: Relevance of H given X within the context of $Y = \langle Certainty \rangle$. This means reliability of the decision is certain. Based on the values of H, five LTs, namely 'Not Likely' (NL), 'Less Likely' (LL), 'Likely' (LY), 'Most Likely' (ML), and 'Certainty' (CY) are defined for E. As the $\underline{RFG}$ is used to extract the crisp knowledge from the video, the MV of $\underline{RFG}$ (i.e., $\mu_{\underline{\{()^t\}}}$) is used to define the MV of E (i.e., $\mu^t_{rel}(E_{()})$). As the $\overline{\text{LTs}}$ for E are user defined, this term (E) may contains uncertainty. Therefore, a quantitative indicator is defined to measure the uncertainty in Z-number to check if the LTs of E are correctly defined or not. Uncertainty in Z-number (say, $Z^{uc(t)}_{()}$) for an event set (say, ()) in f_t is defined as

$$Z^{uc(t)}_{()} = \frac{1}{2} \left\{ \mu_{R^t_{()}} + \mu^t_{rel}(E_{()}) \right\} \quad (23)$$

where $\mu_{R^t_{()}}$ and $\mu^t_{rel}(E_{()})(= \mu_{\underline{\{()^t\}}})$ are the MVs of roughness and reliability, respectively, for an event set (say, ()) from f_t. Three rules are defined using RFG and Z-number for traffic anomaly classification in each f_t, as follows.

Rule 1: If $\mu_{R^t_{co}} = (AI, > 0.9)$, $R^t_{co} < R^t_{nm}$, and $Z^{uc(t)}_{co} < Z^{uc(t)}_{nm}$, then the event is classified as 'Collision'.

Rule 2: If $\mu_{R^t_{nm}} = (AI, > 0.9)$, $R^t_{co} > R^t_{nm}$, and $Z^{uc(t)}_{co} > Z^{uc(t)}_{nm}$, then the event is classified as 'Near-Miss'.

Rule 3: If Rule 1 and Rule 2 are not satisfied, then the event is classified as 'Normal Scenario'.

The $\mu_{R^t_{()}}$ is found to be greater than 0.9 for obtaining maximum accuracy using the validation 21 videos. Therefore, in the aforesaid rules, 0.9 is considered as the threshold for anomaly classification. The results of Z-number for ensuring the correctly defined LTs and classified anomalies are explained in Section V-C.2. The pseudo-code of Z-STRFG is shown in Section S.II in the supplementary material. Experimental results are discussed in the next section.

V. Experiments

Our study has three broad sub-objectives: (i) to perform video summarization using v-RFGs over the real-time videos, (ii) to demonstrate the utility of Z-numbers over v-RFGs for real-time anomaly classification, and (iii) to demonstrate the superiority of Z-STRFG over some state-of-the-art video summarization and deep learning-based traffic anomaly detection algorithms. The details of used datasets, experimental set up, and experimental results are discussed in Sections V-A, V-B, and V-C, respectively.

A. Datasets

To assess the effectiveness of Z-STRFG for video summarization and traffic anomaly detection, a total of 130 real-time traffic videos have been used. The videos contain all possible nature of the three types of scenarios, namely collision, near-miss, and normal traffic. These videos are contributed by YouTube8M [4] (YT8M), Urban Tracker [4] (UT), and Anomaly20 (An20) datasets. As many researchers have already used YT8M and UT videos in their studies [4], [5], [6], [16], [22] and proved their usability by providing detailed experimental results, we have used these videos in our study. YT8M contains 2727 videos on various traffic events. Of them, 53 videos containing collision and normal traffic are selected. YT8M has no annotated videos; therefore, it is used for testing purposes only to get the generalization ability of the Z-STRFG. UT contains 4 videos of normal traffic. Z-STRFG is used on the UT videos to check whether it successfully differentiates normal traffic from the collision and near-miss. While these two datasets (i.e., YT8M and UT) are found suitable for experimental purposes, this is not true for the near-miss detection. There are no near-miss videos present in these two datasets. Therefore, we have developed a dataset and named it An20. This dataset is collected from a plant in India. It contains a total of 30, 29, and 14 videos, which are related to near-miss, collision, and normal traffic, respectively. They are used for both the validation and testing of the developed Z-STRFG.

B. Experimental Setup and Parameter Analysis

In experimental setup, algorithms are coded in Ubuntu 14.04, with Python 3.7 using an Intel $i5$ processor clocked at

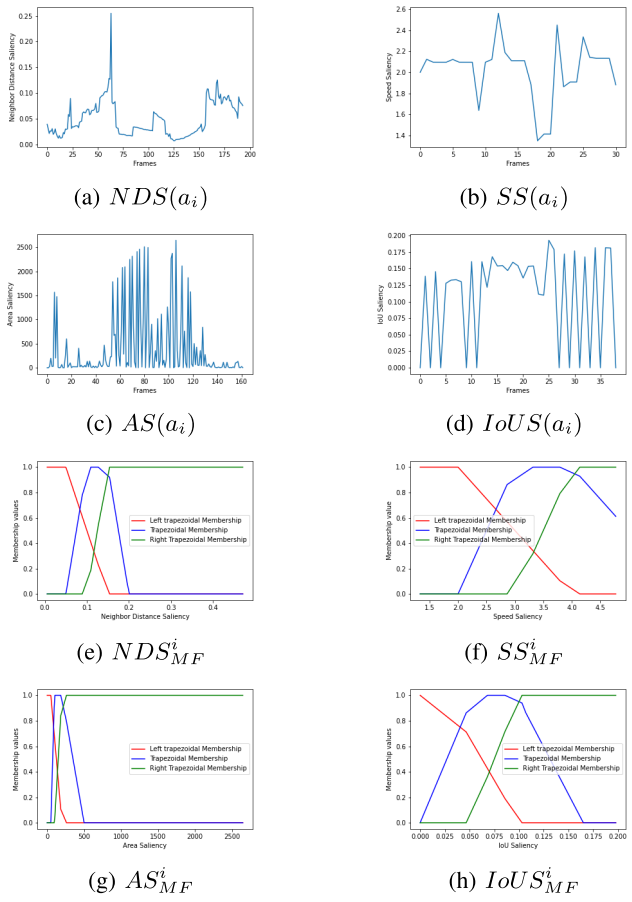
3.30 GHz and 8.00 GB of memory. The libraries used are cv2, Numpy, and Scikit. As the formation of FGs over video frames is based on four thresholds (i.e., T_1, T_2, T_3, and T_4), parameter analysis is done to obtain the optimum FGs. Thresholds T_1, T_2, and T_3 depend on the hyper-parameter m_1, and threshold T_4 depends on m_2. The optimal values of m_1 and m_2 are obtained using the grid search algorithm over the 21 validation videos. From the experiment, we have found that mean absolute error (MAE) is minimum for the combination of $m_1 = 2$ and $m_2 = 5$. Therefore, $m_1 = 2$ and $m_2 = 5$ are used in this study. A detailed information about thresholds selection is given in Section S.III in the supplementary material. According to the abnormality nature of the video content, STF-values are updated automatically. Hence, based on the values of STFs, FGs and RFGs are formed and used for self-learning of the thresholds using the roughness minimization of the RFGs. In this way, the developed Z-STRFG algorithm holds its generalization capability. Hence, the chance of over-fitting of Z-STRFG is negligible. A detailed analysis has been given in Section S.IV in the supplementary material to check whether Z-STRFG algorithm is over-fitted or not.

C. Experimental Results and Discussions

Experiments along with comparisons are conducted to evaluate the effectiveness of the Z-STRFG for video summarization and traffic anomaly detection, in line with three objectives, as discussed in Sections V-C.1, V-C.2, and V-C.3.

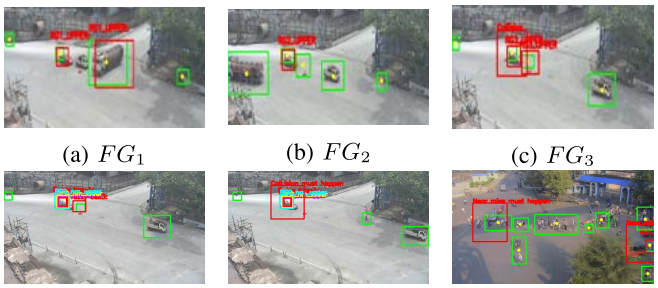
1) The Effectiveness of υ-RFG for Obtaining the Optimal Video Summary: A good video summary should contains three stages (i.e., pre-anomaly, anomaly, and post-anomaly) of an event. These stages are analyzed using the information of υ-RFGs, which are obtained over FGs. $I(\forall_t f_t \in I)$ is the input video over which FGs and RFGs are obtained to extract the summarized video (I_s) from I. FGs are obtained based on the STFs thresholding. The graphs of various STFs using the optimal values of $m_1 = 2$ and $m_2 = 5$ for a FO (a_i, say), are shown in Figs. 2(a) to 2(d). From these figures, it is evident that for an anomalous situation, STFs, including NDS, SS, and AS for a FO (a_i) should be low and IoUS should be high. The graphs of MVs for all the said STFs corresponding to the collision and near-miss events are shown in Figs. 2(e) to 2(h). From these figures, it is evident that the Ltrap, Mtrap, and Rtrap MFs are fitted over the MVs for all STFs. From Fig. 2(f), it is seen that thresholds $a = 2$, $b = 2.8$, $c = 3.9$, and $d = 4.1$ satisfy the Ltrap, Mtrap, and Rtrap MFs corresponding to the MVs of SS. In the same way, thresholds selection for membership assignment corresponding to NDS, IoUS, and AS is done based on the Figs. 2(e), 2(g), and 2(h), respectively. If NDS, SS, AS, and $IoUS$ for a FO follow the decision criterion, as shown in Eqs. 9, 10, and 11, then FG_1, FG_2, and FG_3 are formed over the a_i. The formation of FG_1, FG_2, and FG_3 over different frames (i.e., 45^{th}, 62^{nd}, and 68^{th}) corresponding to a_is are shown in Figs. 3(a), 3(b), and 3(c), respectively.

As said earlier, FGs are formed using STFs and there are an underlying fuzziness, which are characterized by MVs. υ-RFG set is obtained over FGs using the MVs of their



(a) $NDS(a_i)$ (b) $SS(a_i)$ (c) $AS(a_i)$ (d) $IoUS(a_i)$ (e) NDS^i_{MF} (f) SS^i_{MF} (g) AS^i_{MF} (h) $IoUS^i_{MF}$

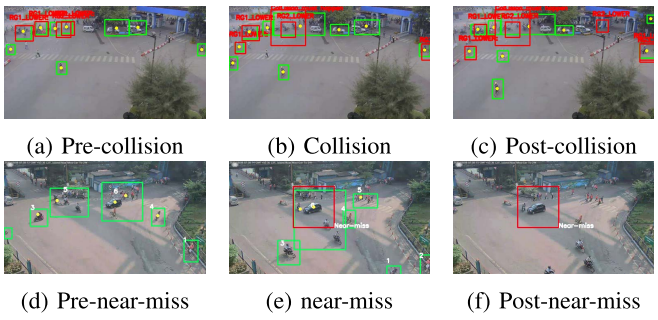
Fig. 2. STFs and corresponding MVs for a_i.



(a) FG_1 (b) FG_2 (c) FG_3 (d) υ-RFG sets (e) Collision (f) Near-miss

Fig. 3. Formation of the υ-FG, υ-RFG, and event sets.

union and intersection sets which are related to the MVs of STFs. From Figs. 2(e) to 2(h), it is evident that, for all STFs, intersect value between Ltrap, Mtrap, and Rtrap MVs is non-zero. Therefore, for a_i, FG_1, FG_2, and FG_3 have non-zero MVs. These MVs are used to extract the set of RFGs. Two RFG sets, $\upsilon - RFG_{co}$ and $\upsilon - RFG_{nm}$ are obtained over a_i for different frames of I. Both $\upsilon - RFG_{co}$ and $\upsilon - RFG_{nm}$ have lower approximation sets (i.e., $\underline{\{RFG_{co}\}}$ and $\underline{\{RFG_{nm}\}}$) and upper approximation sets (i.e., $\overline{\{RFG_{co}\}}$ and $\overline{\{RFG_{nm}\}}$). The formation of these sets over a video frame is depicted in Fig. 3(d). From this figure, it is seen that lower and upper approximation sets for both RFG_{co} and



(a) Pre-collision (b) Collision (c) Post-collision

(d) Pre-near-miss (e) near-miss (f) Post-near-miss

Fig. 4. Z-STRFG output.

TABLE I
TRANSFORMATION RULES OF LT FOR STFS

LT	MV_{SS}	MV_{AS}	MV_{NDS}	MV_{IoUS}
High	(2,3.9)	(20,300)	(0.0,0.5)	(0.06, 0.11)
Medium	(2, 2.8, 3.9,4.1)	(20, 100, 300, 500)	(0.5, 1, 1.7, 2)	(0.03, 0.06, 0.11, 0.16)
Low	(2.8, 3.9)	(80, 400)	(0.9, 1.7)	(0,0.11)

RFG_{nm} are obtained over a_i. It is evident from this figure that $\underline{\{RFG_{co}\}} = \overline{\{RFG_{co}\}}$, therefore, $R^t_{co} = 0$. On the contrary, in the same frame, $\underline{\{RFG_{nm}\}} < \overline{\{RFG_{nm}\}}$, and hence $R^t_{nm} \neq 0$. Therefore, in this frame, $R^t_{co} < R^t_{nm}$. This evidents that collision might have happened in this frame. Resultant frame where collision is occurred is shown in Fig. 3(e). In the same way near-miss is also detected with the condition $R^t_{nm} < R^t_{co}$. Resultant frame where near-miss is occurred is shown in Fig. 3(f). From Fig. 3(f), it is seen that more than two FOs (corresponding to the object dense scenario) are partially occluded, and hence, detected using a single bounding-box, as object localization is done in unsupervised way. Various STFs along with rules for traffic flow are defined to form the RFGs in discriminating between collision, near-miss, and normal traffic. Hence, the performance of Z-STRFG has not been much affected by the functionality of object detection performance, even in an object dense scenario.

It is also observed from these figures that frames containing either $\underline{RFG_{co}}$ or $\underline{RFG_{nm}}$ represent the anomaly location, and would include in the I_s. Therefore, we can interpret that I_s with υ-RFG set is a meaningful subset of I with the following properties: (i) it is the subset with lower cardinality than I but convey same information about an anomaly, and (ii) threshold(s) selection based on υ-RFG set is more prominent and speedy than threshold(s) selection based on the information of entire frame. As a result, video frames get quickly optimized using υ-RFG set. Thus, the effectiveness of υ-RFG set for obtaining the optimized video summary is proved. Further, Z-numbers are computed for υ-RFG set to ensure the reliability of the detected anomaly. Experimental results are stated in the next section.

2) The Usefulness of Z-Numbers Along With υ-RFG Set for Anomaly Classification: Z-number corresponding to the υ-RFG set is computed to obtain the reliability of the classified anomaly. RFGs are obtained over FGs, whose MVs are computed using the MVs of STFs. Therefore, the first tuple (H) of Z-number is defined by some LTs based on the MVs of STFs. Various LTs are assigned to each STF based on its MVs. The transformation rules of LTs for each STF are defined after checking the range of MVs corresponding to the same STF using 21 videos. Results are shown in Table I. Three LTs, namely 'High', 'Medium, and 'Low' are assigned to each STF. In Table I, MV_{SS}, MV_{AS}, MV_{NDS}, and MV_{IoUS} represent the range of MVs (i.e., values of thresholds a, b, c, and d) for SS, AS, NDS, and IoUS, respectively. From Fig. 3(f), it is evident that, thresholds $a = 2$, $b = 2.8$, $c = 3.9$, and $d = 4.1$ are selected for membership assignment corresponding to SS. These thresholds are used for defining LTs corresponding to SS as shown in the 2^{nd} column of Table I. It is seen that if MV_{SS} is varied from 2 to 3.9, then 'High' term is assigned to SS. It means this range of speed saliency is highly sensitive to the anomaly. Likewise, all other LTs and MVs are assigned to each STF. If either 'Medium' or 'Low' term is assigned to any STF, then it is either moderate or less sensitive to the anomaly.

TABLE II
Z-NUMBERS FOR R^t_{co}

Variable (X)	Constraint (H)	Reliability (E)	Z^{uc}_{co}
Collision	(AI,> 0.9)	CY	0.3
Collision	(AI, $\geq 0.7, \leq 0.9$)	ML	0.45
Collision	(AI, $\geq 0.4, \leq 0.7$)	LY	0.65
Collision	(AI, $\geq 0.2, \leq 0.6$)	NL	0.72
Collision	(Im,$\geq 0.7, \leq 0.9$)	LY	0.45
Collision	(Im, $\geq 0.4, \leq 0.7$)	ML	0.62
Near-miss	(Im, $\geq 0.4, \leq 0.7$)	LY	0.55
Near-miss	(Im, $\geq 0.2, \leq 0.6$)	NL	0.6
Near-miss	(LI, > 0.9)	LL	0.63
Near-miss	(LI, $\geq 0.7, \leq 0.9$)	NL	0.67

TABLE III
Z-NUMBERS FOR R^t_{nm}

Variable (X)	Constraint (H)	Reliability (E)	Z^{uc}_{nm}
Near-miss	(AI,> 0.9)	CY	0.35
Near-miss	(AI, $\geq 0.7, \leq 0.9$)	ML	0.48
Near-miss	(AI, $\geq 0.4, \leq 0.7$)	LY	0.55
Near-miss	(AI, $\geq 0.2, \leq 0.6$)	NL	0.6
Near-miss	(Im,$\geq 0.7, \leq 0.9$)	LY	0.45
Near-miss	(Im, $\geq 0.4, \leq 0.7$)	LL	0.53
Collision	(LI, > 0.9)	LL	0.72
Collision	(LI, $\geq 0.7, \leq 0.9$)	NL	0.69

As said earlier, E of Z-number represents the reliability of the decision. The ranges of values for LTs corresponding to E are 'NL' $= (0, 0.2)$, 'LL' $= (0.2, 0.4)$, 'LY' $= (0.4, 0.7)$, 'ML' $= (0.7, 0.9)$, and 'CY' $= (0.9, 1)$. Whereas, three LTs, namely 'Absolutely Important' (AI), 'Important' (Im), and 'Less Important' (LI) are defined for the tuple H. The transformation rules of LTs for H corresponding to collision and near miss events are shown in Tables II and III, respectively. As LTs are user-defined, there is an uncertainty in Z-number, and it is measured to ensure the reliability of the detected anomaly class. For two frames (refer to Figs. 3(e) and 3(f), respectively), uncertainty in Z-number for both R^t_{co} (i.e., $Z^{uc(t)}_{co}$) and R^t_{nm} (i.e., $Z^{uc(t)}_{nm}$) are measured using the information of LTs for both H and E, as reported in Tables II and III, respectively. From the Table II

TABLE IV
REDUCTION RATIO CONTAINING ANOMALY CONTENT

Frames	Ki	Akoz	Song	Yun	PVS	Z-STRFG
Fig. 4(a)	0.32	0.21	0.15	0.23	0.08	0.03
Fig. 4(b)	0.35	0.18	0.14	0.33	0.07	0.02
Fig. 4(c)	0.30	0.16	0.13	0.35	0.09	0.02

(first row), it is seen that the decision rule $\mu_{R^t_{co}} > 0.9$ is 'AI' for classifying an anomaly as 'Collision'. Moreover, 'CY' is assigned to E. It indicates that reliability of the decision is certain. Then, $Z^{uc(t)}_{co}$ for collision detection is 0.3. Similarly, from Table III (first row), it is seen that, $\mu_{R^t_{nm}} > 0.9$ is 'AI' for near-miss detection, and this decision is certainty reliable. Then $Z^{uc(t)}_{nm}$ for near-miss detection is 0.35. For f_t, if $Z^{uc(t)}_{co} < Z^{uc(t)}_{nm}$, $\mu_{R^t_{co}} = (AI, > 0.9)$, and $R^t_{co} < R^t_{nm}$, then this frame must contains collision scenario. From these tables, it is evident that by using Z-number with υ-RFG set, uncertainty between collision and near-miss is handled, and the anomalies are detected successfully, thereby improving the detection accuracy. The effectiveness of Z-STRFG over some state-of-the-art is depicted in the next section.

3) Comparative Studies: To assess the performance of Z-STRFG-based video summarization, five types of comparative studies based on: (i) anomaly classification, (ii) reduction ratio, (iii) detection rate, (iv) CPU runtime, and (v) statistical analysis are done, as discussed in the following section.

a) Anomaly classification: Experiments are performed at different circumstances, namely pre-anomaly, anomaly, post-anomaly, intersection, and non-intersection to prove the effectiveness of Z-STRFG over some state-of-the-art, such as Ki and Lee [16], Aköz and Karsligil [5], Yun *et al.* [6], Song *et al.* [22], and PVS [4]. From the study, it was found that PVS is better than others except Z-STRFG for collision detection under aforesaid circumstances. However, the PVS method cannot detects near-miss. Whereas, the developed Z-STRFG takes care of both collision and near-miss detection. Example results of collision and near-miss detection by Z-STRFG are shown in Figs. 4(a) to 4(c) and Figs. 4(d) to 4(f), respectively. This proves the superiority of Z-STRFG.

b) Reduction ratio: The reduction ratio (RR) is defined as the ratio between the number of frames obtained in the I_s and the total number of frames present in the I. The lower RR indicates the conciseness of the summary. It preserves the information of an anomaly with respect to the number of frames [4]. To preserve the anomaly, the frames with uncertainty may be repeated in the summary. This redundancy can be handled using Z-STRFG through minimization of the roughness score and use of Z-number. In Table IV, the comparison of RR between Z-STRFG and some state-of-the-art is reported. Results are based on the frames present in Figs. 4(a) to 4(c). Here, Figs. 4(a), 4(b), and 4(c) represent the Z-STRFG output for the stages of pre-collision, collision, and post-collision, respectively. Z-STRFG offers preference to the RFGs obtained in I. This is more promising for anomaly detection as the $\underline{RFG}$ must contain the summary (pre, during, and post) of an anomaly occurred in the video, thereby reducing the number of redundant and uncertain frames in the video summary. From Table IV, it is seen that Z-STRFG is superior to others in terms of RR.

c) Detection rate: True positive rate (TPR), false-positive rate (FPR), and accuracy are used as detection rate to evaluate the performance of Z-STRFG. TPR is the rate of truly detected events, whereas, FPR is the rate of wrong detection. Here, the performance of Z-STRFG is compared with three video summarization methods, such as Song, Ki, and PVS and five deep learning-based traffic anomaly detection methods, such as ResNet-50 + CBAM + LSTM (RCL) [23], Local patch learning (LPL) [17], Clustering-based tracking (CBT) [3], Two Stream CNN (TSC) [24], and TPS mapping (TPSM) [28]. Out of these, first three deep models are used for collision detection and rest two are used for near-miss detection. YT8M is used for testing purpose for all aforesaid video summarization and traffic anomaly detection methods. Whereas, PVS is tested over YT8M and An20. UT dataset is used to check whether Z-STRFG can differentiate collision and near-miss from normal traffic or not. The Z-STRFG is used over all the aforesaid datasets, and compared the results wherever applicable, as shown in Table V. From this table, it is seen that state-of-the-art is applicable for either collision (co) or near-miss (nm) detection. Whereas, Z-STRFG is applicable for both collision and near-miss detection. It is also seen that Z-STRFG is superior to some conventional and deep learning-based techniques for both collision and near-miss detection in terms of TPR, FPR, and accuracy.

d) CPU runtime: The computational complexity for the Z-STRFG algorithm is $O(n_t \cdot (4n_t + 2 + n_{FG} + n_{RFG}))$ in the worst case, where n_t, n_{FG}, and n_{RFG} represent the number of FOs, FGs, and RFGs, respectively obtained in f_t. CPU runtime for PVS is 3fps [4], whereas, CPU runtime for Z-STRFG is 11fps. Therefore, Z-STRFG is superior to PVS in terms of CPU runtime.

e) Statistical analysis: A non-parametric statistical test is conducted to ensure the statistical significance of Z-STRFG over some state-of-the-art, namely PVS, RCL, TSC, and TPSM. Out of various non-parametric tests, Wilcoxson signed-rank test is conducted with $\alpha = 0.05$. Here, null hypothesis H_0 is set as 'Z-STRFG is statistically significant than state-of-the-art'. If the p-value is found to be less than 0.05, then H_0 will be failed to reject; otherwise, H_0 will be rejected. Based on the TPR, Wilcoxon signed-rank test is done for each pair-wise algorithms. Results are shown in Table VI, where '×' represents 'Not Applicable'. It is found that each pair-wise (i.e., Z-STRFG vs PVS and Z-STRFG vs RCL for collision detection, and Z-STRFG vs TSC and Z-STRFG vs TPSM for near-miss detection) comparison shows a statistically significant difference (in terms of p-values) between two algorithms at $\alpha = 0.05$. It is evident from this table that Z-STRFG is statistically significant than PVS and RCL for collision detection, and TSC and TPSM for near-miss detection (as $p < 0.05$). All results prove the superiority of Z-STRFG over some state-of-the-art in terms of detection rate, speed, and video summarization.

TABLE V
ANOMALY DETECTION RATE

Algorithm	Data	Event	TPR	FPR	Accuracy
Song	YT8M	co	71%	16%	74.3%
Ki	YT8M	co	69%	21%	72.5%
PVS	YT8M	co	82.3%	10%	86.4%
PVS	An20	co	80.6%	13.8%	84.1%
RCL	YT8M	co	82.6%	14.8%	85.3%
TSC	An20	nm	83.3%	13.1%	86.1%
TPSM	An20	nm	86.2%	10.6%	88.6%
LPL	YT8M	co	81%	13%	86.7%
CBT	YT8M	co	79%	16%	82.8%
Z-STRFG	YT8M	co	90.1%	3.4%	96.2%
Z-STRFG	An20	co	91.2%	3.2%	96.8%
Z-STRFG	An20	nm	91.5%	4.2%	95.7%

TABLE VI
STATISTICAL SIGNIFICANCE TEST RESULTS

Algorithm	Anomalies	PVS	RCL	TSC	TPSM
Z-STRFG	Collision	0.041	0.037	×	×
Z-STRFG	Near-miss	×	×	0.047	0.044

VI. CONCLUSION

A new video summarization algorithm, namely Z-numbers-based spatio-temporal rough fuzzy granulation (Z-STRFG) is developed to handle the uncertainty issue arises in between collision, near-miss, and normal traffic. Various spatio-temporal features are computed and are used for generating the approximate anomaly-prone regions. Only these regions are used for obtaining the two types of rough fuzzy granules (RFGs) along with their roughness scores in discriminating between collision, near-miss, and normal traffic. Further, Z-number is computed over each RFG to obtain the reliability of the detected anomaly class. In Z-STRFG, the process is restricted to the fuzzy granules, rather than considering all foreground objects, which leads to a significant improvement in anomaly detection speed. The use of RFGs and Z-numbers in handling uncertainty is more effective than using only saliency costs. This improves the detection accuracy even in case of dense scenario having occluded objects. Experimental results reveal that Z-STRFG is superior to Ki and Lee [16], Aköz and Karsligil [5], Yun *et al.* [6], Song *et al.* [22], RCL [23], LPL [17], CBT [3], PVS [4], TSC [24], and TPSM [28] when tested over either YouTube8M [4] or Anomaly20, wherever applicable. Anomaly20 is a new real-life traffic data acquired from a plant in India. Z-STRFG is restricted to the monocular vision.

ACKNOWLEDGMENT

The authors acknowledge UAY Project, Govt. of India (GoI), the CoE-SEA, and SAVR Lab of IIT Kharagpur. Sankar Kumar Pal acknowledges his National Science Chair, SERB-DST, GoI.

REFERENCES

[1] ASIRT. (2016). *Road Crash Statistics*. [Online]. Available: https://asirt.org/initiatives/informing-road-users/road-safety-facts/road-crash-statistics

[2] W. Bridges, "Gains from getting near misses reported," Presentation at the 8th Global Congr. Process Saf., Houston, TX, USA, Apr. 2012, pp. 1–4.

[3] Y. Li, W. Lin, T. Wang, Q. Guo, R. Yang, and S. Xu, "Video summarization via cluster-based object tracking and type-based synopsis," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Aug. 2020, pp. 113–116.

[4] S. S. Thomas, S. Gupta, and V. K. Subramanian, "Event detection on roads using perceptual video summarization," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 9, pp. 2944–2954, Dec. 2017.

[5] O. Akoz and M. E. Karsligil, "Video-based traffic accident analysis at intersections using partial vehicle trajectories," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 499–502.

[6] 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.*, 2014, pp. 3062–3067.

[7] T. Dilber, M. Serdar Guzel, and E. Bostanci, "A new video synopsis based approach using stereo camera," 2021, *arXiv:2106.12362*.

[8] A. Pramanik, S. K. Pal, J. Maiti, and P. Mitra, "Granulated RCNN and multi-class deep SORT for multi-object detection and tracking," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 1, pp. 171–181, Feb. 2022.

[9] C. A. Murthy and S. K. Pal, "Histogram thresholding by minimizing graylevel fuzziness," *Inf. Sci.*, vol. 60, nos. 1–2, pp. 107–135, Mar. 1992.

[10] S. K. Pal and A. Ghosh, "Fuzzy geometry in image analysis," *Fuzzy Sets Syst.*, vol. 48, no. 1, pp. 23–40, May 1992.

[11] A. Pathak and S. K. Pal, "Fuzzy grammars in syntactic recognition of skeletal maturity from X-rays," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-16, no. 5, pp. 657–667, Sep. 1986.

[12] D. Sen and S. K. Pal, "Generalized rough sets, entropy, and image ambiguity measures," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 39, no. 1, pp. 117–128, Feb. 2009.

[13] S. K. Pal and A. Skowron, *Rough-Fuzzy Hybridization: A New Trend in Decision Making*. Berlin, Germany: Springer-Verlag, 1999.

[14] L. A. Zadeh, "A note on Z-numbers," *Inf. Sci.*, vol. 181, no. 14, pp. 2923–2932, 2011.

[15] S. K. Pal, D. Bhoumik, and D. B. Chakraborty, "Granulated deep learning and Z-numbers in motion detection and object recognition," *Neural Comput. Appl.*, vol. 32, pp. 1–16, May 2019.

[16] Y. K. Ki and D. Y. Lee, "A traffic accident recording and reporting model at intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 188–194, Jun. 2007.

[17] W. Lin, Y. Zhang, J. Lu, B. Zhou, J. Wang, and Y. Zhou, "Summarizing surveillance videos with local-patch-learning-based abnormality detection, blob sequence optimization, and type-based synopsis," *Neurocomputing*, vol. 155, pp. 84–98, May 2015.

[18] H. T. Nguyen, S.-W. Jung, and C. S. Won, "Order-preserving condensation of moving objects in surveillance videos," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 9, pp. 2408–2418, Sep. 2016.

[19] S. Chakraborty, O. Tickoo, and R. Iyer, "Adaptive keyframe selection for video summarization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 702–709.

[20] 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.

[21] K. Doshi and Y. Yilmaz, "An efficient approach for anomaly detection in traffic videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4236–4244.

[22] H.-S. Song, S.-N. Lu, X. Ma, Y. Yang, X.-Q. Liu, and P. Zhang, "Vehicle behavior analysis using target motion trajectories," *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 3580–3591, Oct. 2014.

[23] Z. Lu, W. Zhou, S. Zhang, and C. Wang, "A new video-based crash detection method: Balancing speed and accuracy using a feature fusion deep learning framework," *J. Adv. Transp.*, vol. 2020, pp. 1–12, Nov. 2020.

[24] 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, 2020.

[25] A. Albanese, S. K. Pal, and A. Petrosino, "Rough sets, kernel set, and spatiotemporal outlier detection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 194–207, Jan. 2014.

[26] Z. Yuan, H. Chen, T. Li, B. Sang, and S. Wang, "Outlier detection based on fuzzy rough granules in mixed attribute data," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8399–8412, Aug. 2022.

[27] H.-J. Zimmermann, *Fuzzy Set Theory—And Its Applications*. Heidelberg, Germany: Springer, 2011.

[28] X. Huang, T. Banerjee, K. Chen, N. V. S. Varanasi, A. Rangarajan, and S. Ranka, "Machine learning based video processing for real-time near-miss detection," in *Proc. VEHITS*, 2020, pp. 169–179.