An Intelligent Video Analysis Method for Abnormal Event Detection in Intelligent Transportation Systems

Shaohua Wan, *Senior Member, IEEE*, Xiaolong Xu, *Member, IEEE*, Tian Wang, and Zonghua Gu, *Senior Member, IEEE*

***Abstract*— Intelligent transportation systems pervasively deploy thousands of video cameras. Analyzing live video streams from these cameras is of significant importance to public safety. As streaming video is increasing, it becomes infeasible to have human operators sitting in front of hundreds of screens to catch suspicious activities or detect objects of interests in real-time. Actually, with millions of traffic surveillance cameras installed, video retrieval is more vital than ever. To that end, this article proposes a long video event retrieval algorithm based on superframe segmentation. By detecting the motion amplitude of the long video, a large number of redundant frames can be effectively removed from the long video, thereby reducing the number of frames that need to be calculated subsequently. Then, by using a superframe segmentation algorithm based on feature fusion, the remaining long video is divided into several Segments of Interest (SOIs) which include the video events. Finally, the trained semantic model is used to match the answer generated by the text question, and the result with the highest matching value is considered as the video segment corresponding to the question. Experimental results demonstrate that our proposed long video event retrieval and description method which significantly improves the efficiency and accuracy of semantic description, and significantly reduces the retrieval time.**

***Index Terms*— Intelligent transportation systems, long video event retrieval, segment of interest, superframe segmentation, question-answering.**

I. Introduction

INTELLIGENT transportation system (ITS) can improve the traffic efficiency and effectively guarantee the safety of vehicles and pedestrians on the supervised section. Therefore, it has widely attracted researchers' attention. The road traffic safety situation in the past is facing increasingly severe challenges, and traffic accidents have still frequently happened. It is a huge challenge to detect traffic accidents quickly and accurately, and to avoid the traffic safety problems caused by them. As one of the important sources of video data, video capture cameras can be seen anywhere in all corners of road intersections. Not only that, the number of cameras has also been expanding at an annual growth rate of 20%, accompanied by video analysis derived from video big data. With the rapid growth of the number of applications, video analysis in the intelligent transportation public safety scene has also attracted the attention of academia and industry. In the context of the rapid growth of data processing, how to obtain useful data in videos has become a key goal in the development of ITS to cut down traffic accidents and confirm on the liability of the traffic accidents. An intelligent video analysis method for abnormal event detection is an effective means to achieve this goal, and will determine the degree of intelligence of the entire ITS.

It is easy for humans to watch a long video and describe what happened at each moment in text. However, it is a very challenging task to make a machine capture and extract specific events from long videos and then give descriptive text. The technology that completes such task has received extensive attention in the field of computer vision due to its promising prospects in video surveillance and assisting the blind. Traffic departments analyze video streams from cameras at intersections for traffic flow control, vehicles recognition, vehicle properties extraction, traffic rule violations, and accidents detection. Different from the simple task of the semantic description of static images, the description of video content is more challenging, because it needs to understand a series of consecutive scenes to generate multiple description segments. At present, most of the existing research focuses on the description of short videos or video segments. However, the videos that record actual scenarios are very long, which may be hundreds of minutes in length. So, it takes a lot of time and cost to achieve video retrieval and information selection.

Event retrieval and description of long videos are generally driven by the advances in segment of interest (SOI) recognition, key frame selection, and image semantic description and generation. Sah *et al.* [1] extracted the SOI based on the quality

Manuscript received March 28, 2020; revised July 29, 2020; accepted August 13, 2020. This work was supported in part by the National Natural Science Foundation of China (No.61672454, No. 61762055); in part by the Fundamental Research Funds for the Central Universities of China under Grant 2722019PY052 and by the open project from the State Key Laboratory for Novel Software Technology, Nanjing University, under Grant No. KFKT2019B17. The Associate Editor for this article was A. Jolfaei. *(Corresponding author: Shaohua Wan.)*

Shaohua Wan is with the Department of Computer Science and Engineering, Shaoxing University, Shaoxing 312000, China, also with the School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China, and also with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (e-mail: shaohua.wan@ieee.org).

Xiaolong Xu is with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China.

Tian Wang is with the College of Computer Science, Huaqiao University, Xiamen 361021, China.

Zonghua Gu is with the Department of Applied Physics and Electronics, Umeå Universitet, 90187 Umeå, Sweden.

Digital Object Identifier 10.1109/TITS.2020.3017505

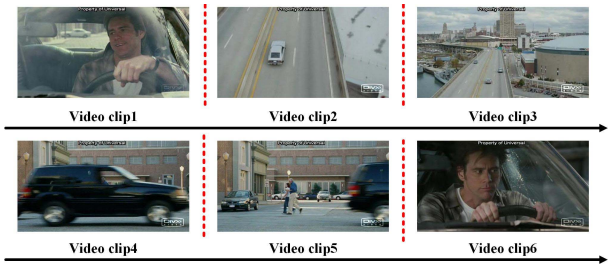


Fig. 1. Long video segmentation and specific event retrieval.

of video frames, and then used deep learning algorithms to encode and decode the video segments, thereby converting the key frames of valid video segments into text annotations, and finally humans were asked to perform information selection and semantic evaluation on the text annotations. Lu and Grauman [2] proposed a video summary generation algorithm that used image quality factors to select representative sub-videos from a given long video to describe basic events. Wolf [3] used the key frame sequence in the video segment to represent the change in video content to replace the corresponding video, which not only reduced the data to be processed, but also greatly improved the efficiency of video retrieval. All the above methods select the key frames in long videos and use these frames instead of long videos to describe the video content. But the above methods all rely on a single modality (video) as the reference for retrieving video content. In practice, videos are often associated with other forms such as audio or text, such as the subtitles of a movie/TV show or the audience words accompanying a live video. These related modes may be an equally important source for retrieving user-related moments.

As shown in Fig. 1, in a continuous video of street scene, several volunteers prepared some food for distribution to the homeless. The video contains introductions, food preparation, and distribution of food to the homeless on the street. If we want to quote a specific scene or a certain moment in the video, such as an old man sitting on the street, simply referencing the moment by the keywords like action, object, or attribute may not uniquely identify it. For example, important objects in the scene, such as the elderly, appear in many frames. Based on this example, we consider using natural language to locate the moments in a video. Specifically, for a video and text description, we identify the start and end in the video that correspond to the given text description, which is a challenging task that requires understanding both language and video. It has important applications in video retrieval, such as finding specific moments from a a long video, or finding the desired B-roll stock video segment from large video libraries (such as Adobe Stock1, Getty2, Shutterstock3). Aiming at the problems of large-scale computation and large time consumption in the content analysis and topic retrieval of long videos, this article proposes a novel long video event retrieval and description method which significantly improves the efficiency and accuracy of semantic description, and significantly reduces the retrieval time.

The main contributions of this article can be summarized as follows:

- An intelligent video analysis method for abnormal event detection in intelligent transportation systems is proposed based on VQA. By detecting the motion amplitude of the long video, a large number of redundant frames can be effectively removed from the long video, thereby reducing the number of frames that need to be calculated subsequently.
- By using a superframe segmentation algorithm based on feature fusion, the remaining long video is divided into several SOIs which include the video events.
- The trained semantic model is presented to match the answer generated by the text question, and the result with the highest matching value is considered as the video segment corresponding to the question.
- An extensive experimental validation study has been conducted on some benchmark datasets like SumMe dataset and the Hollywood2 dataset, which get excellent performance.

The rest of the paper is organized as follows. Section II provides background of the closely related work. Then, Section III introduces the proposed long video event retrieval algorithm based on superframe segmentation. In section IV, we discuss the experimental setup and the results obtained respectively. Section V wraps up this article with conclusions and discussions of our on-going efforts.

II. Related Work

A. Long Video Event Retrieval

With the rapid development of Internet technology and the popularization of multimedia equipment, video resources have also greatly boomed. For example, about 100 hours of video resources are uploaded to YouTube every minute. These videos often lack professional annotations and content descriptions, which is not conducive to people's rapid retrieval for required video resources and cannot achieve the real-time surveillance in traffic video. Therefore, the use of natural language descriptions has been proposed to describe events in videos, then people propose the corresponding text questions according to their needs, and finally retrieve and locate the video events through answer matching. At present, the widely-used method [4] to achieve event retrieval is to use the deep video language embedding method proposed by references [5]–[8]. In addition, such methods also rely on the joint embedding of video features and natural language. For example, reference [9] used home video surveillance to retrieve daily events, which included a fixed set of spatial prepositions ("across" and "through"). Similarly, reference [10] considered aligning text instructions with video events. However, the method of aligning instructions with video is only applicable for structured videos because they constrain alignment through the order of instructions. In contrast, the actual surveillance video generally contains unconstrained open scenes.

B. Video Semantic Description

The essence of video semantic description is to separate important events in a video according to time labels and give corresponding description sentences. Earlier research on video

summary did not include natural language input [11]–[14], but some algorithms used video-like text [15] or category tags for event query and content selection [16]. Reference [17] collected the text descriptions of video blocks as a summary of the entire video. The dataset used in the above method does not contain relational expressions and has a limited scope of application, so it is not suitable for the event retrieval in actual monitoring scenarios.

C. Video Captioning With Question Answering

The question answering system is a task system that takes an image and a free, open natural language question about the image as the input, and generates a natural language answer as the output. Since the question answering system involves machine vision and natural language processing, combining the machine vision algorithm with the natural language processing algorithm to build a combined model has become the most common method to solve the problem of the question answering system. This combined structure first uses deep learning architecture to extract visual features, and then uses a recurrent neural network capable of processing sequence information to generate the text descriptions of an image. Ma *et al.* [18] used 3 convolutional neural networks (CNN) to complete the image question-and-answer task. Gao *et al.* [19] used a more complex model structure. Malinowski and Fritz [20] combined the latest technologies in natural language processing and computer vision to propose a method for automatically answering image questions. Ren *et al.* [21], [22] suggested combining neural network and visual semantics instead of preprocessing processes such as object detection and image segmentation to perform answer prediction, and obtained good results on public benchmark datasets. Tu *et al.* [23] jointly parsed the video and the corresponding text content and tested it on two data sets containing 15 video samples. Therefore, a successful VQA system usually requires a more detailed understanding of the image and complex reasoning than a system that generates generic image subtitles. Agrawal *et al.* [24] proposed a free-form open-ended visual VQA model. The model can provide accurate natural language answers by entering images into the model and relevant natural language questions.

III. Proposed Method

A. Detection of Redundant Frames in a Long Video

Traffic surveillance cameras generally collect video data in the surveillance area at a sampling rate of 25 frames per second. This is to ensure that the video can maintain a good smoothness. Because these cameras need to collect the traffic scenes 24 hours in an uninterrupted manner, the total number of generated frames can be hundreds of thousands or even millions. The processing of such a large number of frames will consume a lot of computation time, making it difficult to meet the requirements of real-time traffic monitoring. By observing the behavior events in surveillance videos, it is found that long videos often contain a large number of useless static frames (redundant frames), and the processing of these redundant frames consumes much time.

In order to improve the speed of processing large videos, it is necessary to detect and remove a large amount of redundant and meaningless frames contained in long videos. In this research, the method of motion amplitude detection based on local spatiotemporal interest points to achieve the effective detection of redundant frames. Firstly, the improved spatiotemporal interest point detection algorithm is used to calculate the spatiotemporal interest points of each frame in the video. Then, surround inhibition is combined with local and temporal constraints to detect static interest points in the frame. According to the characteristics of spatiotemporal interest points, when the number and position of interest points in a video have not changed, according to experimental observations, it is considered that the content of this video has not changed. Therefore, this characteristic can be employed to remove a large number of unchanged redundant frames existing in a long video. When the number of valid spatiotemporal interest points detected is lower than the threshold value, it means that the current video has a low amplitude of motion or no motion occurs, so it can be determined that the content of this video has not changed and the redundant frames can be removed. In addition, due to the repetitive nature of frames, deleting the redundant frames does not affect the expression of the video content.

B. Extraction of SOI Based on Superframe Segmentation

In the previous section, a large number of redundant frames in a long video can be removed by comparing the changes in the number of motion detection boxes. Since the feature extraction and feature matching of the frames in a long video need to be performed later, the reduced number of extra frames can greatly improve the processing speed. This section will perform video segmentation on the long video with redundant frames removed, and then extract SOI for video event retrieval. Video superframe segmentation divides a video sequence into specific, unique parts or subsets according to certain rules, and extracts the SOI. Reference [25] proposed a method for image quality assessment and applied it to the fast classification of high-quality professional images and low-quality snapshots. Inspired by this, this section chooses to combine low-level features such as contrast, sharpness, and color with advanced semantic features such as attention and face information. This linear combination of these features is used to calculate the interestingness measure of the video segment, and then the long video is segmented based on the interestingness measure.

This article refers to the method in [25] to calculate the contrast score C. Each frame in the video is converted to a grayscale image, and the converted image is processed using low-pass filtering. The converted image is resampled, and the height is adjusted to 64, followed by the adjustment of the width according to the aspect ratio. Since sharpness is an important indicator to describe the quality of a frame, it can well correspond to human subjective feelings. The sharpness score E is obtained by converting a frame into a grayscale image, followed by calculating the square of the difference of the grayscale values of two adjacent pixels. In addition to contrast and sharpness, color is also an important feature

for video segmentation. Biological saliency research holds that color is objectively a stimulus and symbol to humans, and it is subjectively a reaction and behavior. The human visual system is very sensitive to external color changes. In addition, the spatial relationship also affects visual saliency, for example, the high contrast of adjacent areas is more likely to attract visual attention. Similar to the method of calculating the contrast score C, each frame in the video is first converted to the HSV color space, and then is processed using low-pass filtering. The image is resampled, and the height is adjusted to 64, followed by the adjustment of the width according to the aspect ratio. Next, the average color saturation score S of the frame is calculated.

In video segmentation, in addition to the underlying feature information, high-level semantic information also needs to be considered. Here, the method in reference [26] is used to calculate the attention score A. By using a time-gradient-based dynamic visual saliency detection model, frames that may cause visual attention are collected and the corresponding attention score A is calculated. Face information can be used as an important reference for video event retrieval. Similar to the method in reference [27], by detecting the face information in the frame, each score is assigned to each detected face, and then all the scores of detected faces are added as the face score F. Finally, referring to the contribution weights of different features, a linear combination of multi-modal features is used to calculate the score of SOI in the video:

$$I_{score} = \eta(A) + C \cdot E \cdot S + \gamma\,(F) \tag{1}$$

where I_{score} is the final score of interestingness measure, $\gamma = 0.5, \eta = 0.25$. We compute an interestingness score by using non-linear combination of fractions including Attention (A), Contrast (C), Sharpness (E), Colorfulness (S) and Facial Impact (F). Finally, the boundary of long video is determined by the interestingness score. Different colors represent contribution of features — Attention, Contrast, Sharpness, Colorfulness and Facial Score. Empirical testing has shown that Attention, Contrast, Colorfulness and Sharpness are essential feature elements for video segmentation. Facial information is of great importance, however, not every human face appears in every video, thus an influence factor η is added to Facial score. The final measure of superframe cut interestingness score is computed as Equation (1).

The long video is segmented into some segments. These segmented video segments contain the key frames from the original video used to generate video subtitles. As shown in Fig. 2, the video is segmented according to the scores of different feature elements in the video segmentation process, and the long video is segmented into multiple SOIs.

C. Extraction of Visual Features

Through the above processing, the long video is converted into several SOIs, and the redundant frames of these segments have been removed. Video events are included in these segments, so we only need to conduct further processing on these segments. It is required that the video event retrieval model can effectively use natural language-based text question

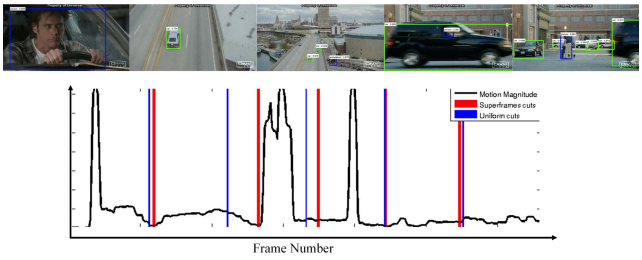


Fig. 2. Segmenting a long video into segments using a superframe segmentation algorithm.

for query localization. For any given SOI $v = (v_t)$, where $t = \in (0, \ldots, T-1)$ represents the length of the SOI, and $\hat{\Gamma} = (\tau_{start} - \tau_{end})$ represents the start and end of the SOI corresponding to the event relative to the entire video. By combining local features and global video context, the temporal context features of the video are extracted to encode the video moments.

$$\hat{\Gamma} = argmin J_0(q, v, \tau) \tag{2}$$

where J_0 represents the joint embedding model which combines the text question q, the features v of SOI, and the given model parameter θ.

In order to further extract the key frame features and feature information of SOI, a deep convolution network is used to extract features for each video. Local features are extracted by extracting the features of the frame at a specific time, global features are extracted by extracting the features of the SOI, and temporal endpoint features are extracted by extracting the features of the moment of SOI. In order to construct local and global video features, a deep convolution network is used to extract the high-level features of each frame, and then average pooling is performed on the video features within the SOI, that is, averaging all frames in the SOI. When there are scene events in the video that may involve the text questions in this article, we can query and confirm the events by matching. For example, if the text question is "someone is riding a bicycle," the proposed algorithm can locate the scene in the video where the event occurs, and lock the moment in the initial stage of the event, followed by encoding the characteristics of this time period. This is similar to the global image features and context features often used in natural language object retrieval. Locating video events is usually achieved by locating some specific actions, such as "cycling", and "running". This article uses the VGG model pre-trained on ImageNet to extract local features, global features, and temporal endpoint features from frames, which can be represented by F_v^{θ}.

D. Word Vector Transformation of Question Text

The question text consists of natural language and requires to be pre-processed before further used. Firstly, the text questions are pre-processed. Each question is divided into words by spaces and punctuation. Words including numbers are also treated as separate words. Teney *et al.* [28] analyzed the length of the questions in VQA dataset and found that only about 0.25% of the questions were longer than 15 words. Therefore, in order to improve the computation efficiency, this

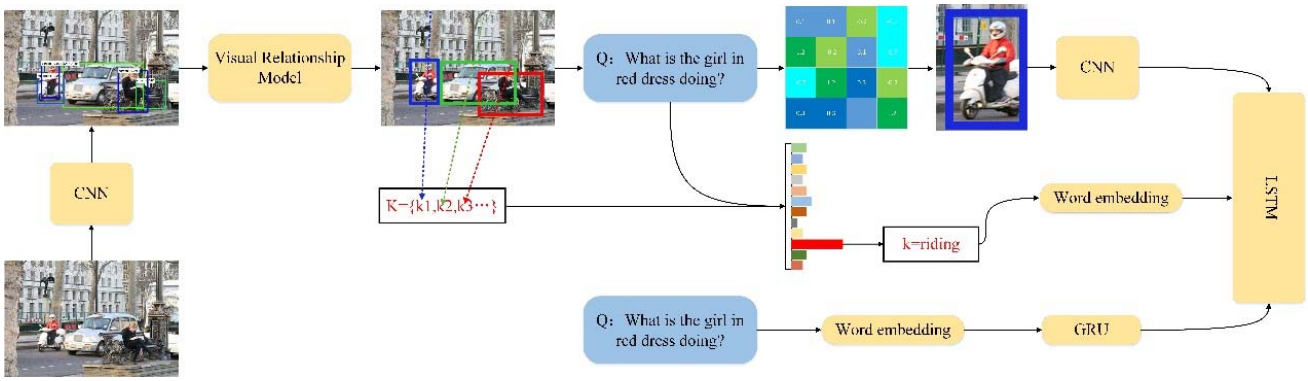


Fig. 3. The pipeline of visual question answering model.

article only retains 15 words when segmenting a sentence. After that, the words are transformed with word2vec into a 300-dimensional word vector. Finally, the word vectors are sent to the LSTM to extract language features, where the embedding sequence of the question sentence has a size of 15,300.

E. Combination of Visual Features and Text Vectors

After the video feature vector P_θ^V is obtained, it is unified with the question text P_θ^L into the same word vector space through a non-linear transformation function. Then the two vectors are combined and represented by:

$$J_\theta(q, v, \tau) = |P_\theta^V(v, \tau) - P_\theta^L(q)| \tag{3}$$

After the model is constructed, it is trained using the loss function. The purpose of training is to obtain the event moment information close to the description of the question text. In order to enhance the robustness of the model, negative samples from different SOIs of the same video and from different videos are added when training the model, so that the model can distinguish some subtle behavioral differences. Herein, we refer to the method proposed by Hendricks *et al.* [29], where the loss function used is defined as:

$$Loss_i^{in}(\theta) = \sum_{n \in \tau} Loss^R(J_\theta(q^i, v^i, \tau^i), J_\theta(q^i, v^i, n^i)) \tag{4}$$

where $L^R(x, y) = max(0, x - y + b)$ represents the loss ranking. In this way, the current video segment is closer to the query result of the question text than all other possible video segments from the same video.

The pipeline of VQA model based on multi-objective visual relationship detection is proposed in this article (as shown in Fig.3), which is inspired by the research on the target relationship in the image. Firstly, the target relationship detection model is pre-trained, and then the appearance relationship feature is used to replace the image features extracted from the original target. At the same time, the appearance model is extended by the word vector similarity principle of the relation predicate, and the appearance features and relationship predicates are sent to the word vector space and are represented by fixed-size vectors. Finally, the integrated vector is sent to the classifier to generate an answer output, through the cascading of elements between the picture feature vector and the question vector. The basic structure of the VQA model is to directly extract image information with CNN, and then send the image features into LSTM to produce prediction results. In this article, the target combination feature vector and the target relationship predicate generated by the image appearance relation model are used to provide image information for the image. The image appearance model consists of two parts: target detection model and target relationship judgment model.

IV. Experimental Results and Analysis

This section verifies the performance of the proposed algorithm through quantitative and qualitative analysis. The experiment is divided into the following subsections according to the algorithm steps. The first subsection mainly verifies the long video segmentation and SOI extraction algorithms based on the detection of motion amplitude. The second subsection is to verify the long video event retrieval algorithm based on text questions. Finally, the accuracy and reliability of the proposed algorithm is analyzed and verified using actual traffic scenarios.

A. Evaluation of Long Video Superframe Segmentation Algorithm

This section uses the SumMe dataset [30] and the Hollywood2 dataset [31] to evaluate the effectiveness of the proposed algorithm for superframe segmentation. The SumMe dataset contains 25 videos. The Hollywoo2 dataset contains 3,669 samples, including 12 categories of actions and 10 categories of scenes, all from 69 Hollywood movies. As shown in Fig. 4, the video is selected from the Hollywood2 dataset and it describes the process of the male protagonist driving home through the street in a movie clip. By detecting the number of points of interest, we could decide whether the video contents at different times change, and then the redundant video frames can be optimized according to the changes in the number of points of interest in a certain time horizon.

As shown in Fig. 5, the video is selected from the Hollywood2 dataset, and it describes an outdoor street scene. Unlike the previous two videos, outdoor scenes are often more complex and changeable. The characters and events contained in the video are no longer unique, and different events may overlap or partially overlap on the time axis. Therefore, it is a very challenging task to screen useful

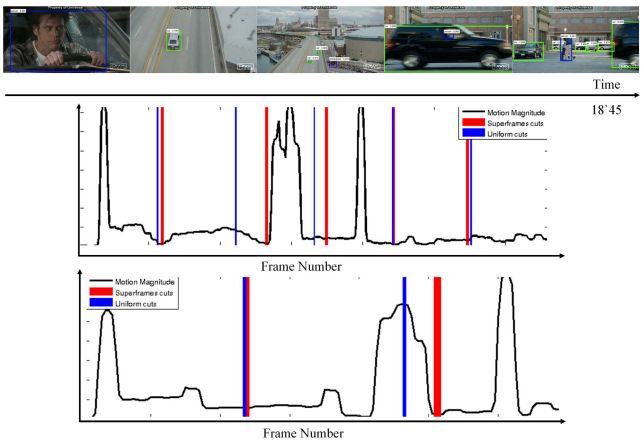


Fig. 4. Identification of interesting segments from the long video.

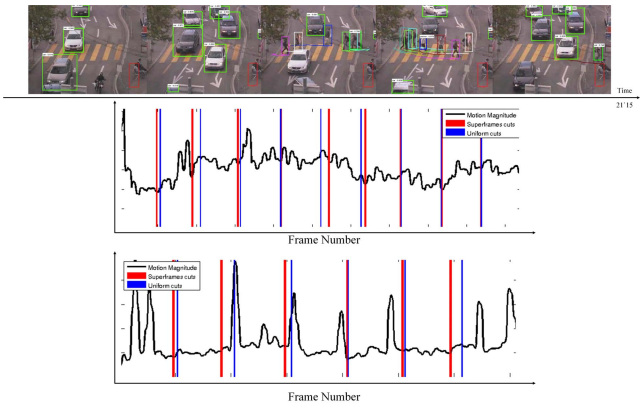


Fig. 5. Identification of SOI from the long video on street scene.

video segments from complex outdoor environments and give appropriate descriptions. The number of frames in this video is 7,373. After motion amplitude detection, the number of frames is reduced to 1,700, and the entire video is divided into 29 SOIs.

As shown in Fig. 6, the video content explains a road scene in campus. The road conditions on campus are relatively simple compared to external transportation, and the main task is on the detection and description of pedestrians. The video contents vary with the movement of the vehicle, and mainly records the state of the vehicle and pedestrians in the front. There are a large volume of redundant video frames in this kind of video, and optimization algorithms should be used to remove more redundant video frames. Table I shows the influences of different features on the results of video segmentation. The average value of each feature in superframe segmentation is used for feature influence analysis together with the average benchmark correlation score. The mean square error of the linear regression model is used as the fitness criterion that affects the score. It can be seen from the evaluation results that all features have a significant role in superframe segmentation. Although the contrast features and facial features have the lowest scores, the overall performance of each feature is well balanced. Although reference [32] considered facial features as the most important parts in the detection of key frames, they would be greatly affected by

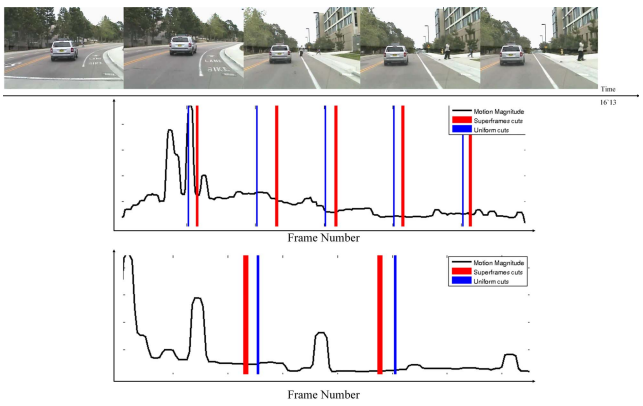


Fig. 6. Identification of SOI from the long video on school scene.

TABLE I
FEATURE EVALUATION OF COMPLEX STREETSCAPE ON SUMME DATASET

Feature	Mean rank	Top-1	Top-2	Top-3
Contrast	0.358	2	2	2
FaceImpact	0.336	1	3	5
Sharpness	0.428	2	4	4
Saturation	0.438	4	5	9
Attention	0.395	2	3	4

the sharpness and angle of the video in the actual detection process, which is an unstable factor.

B. Question Text Matching and Event Retrieval

After optimizing and segmenting the long video, several video segments can be obtained, each of which contains potential video events. As shown in Fig. 7, a traffic accident video shows several steps, including pre-accident, post-accident and the crowd reaction of the accident. These steps can be divided into different video segments through preprocessing, and then the semantic description of these video segments is extracted using a semantic model. Next, when querying a question or searching for a specific event, we only need to match the question text with the extracted description sentence of the video segment to locate the time of the event and obtain the corresponding event description.

As shown in Fig. 7, after the long video is subject to the removal of redundant frames and divided into several video segments, a semantic-based VQA model can be used to obtain a natural language description sentence that can represent each video segment. When we want to query the video content or a specific event, we only need to convert the question into a text vector and assign it to different video segments. The part marked by the red box is the video segment closest to the text question. For example, our problem is “two men are talking”, then the model will automatically retrieve for the moment when the two men start talking in the video in chronological order, and record the video segment and moment where this content is located. Similarly, if our problem is “the moment when a white car appears”, then the model will locate the first video segment of the white car based on the existing retrieval results. With the above method, as long as the following two conditions are met, the video event retrieval task can be successfully completed. The first is an accurate frame

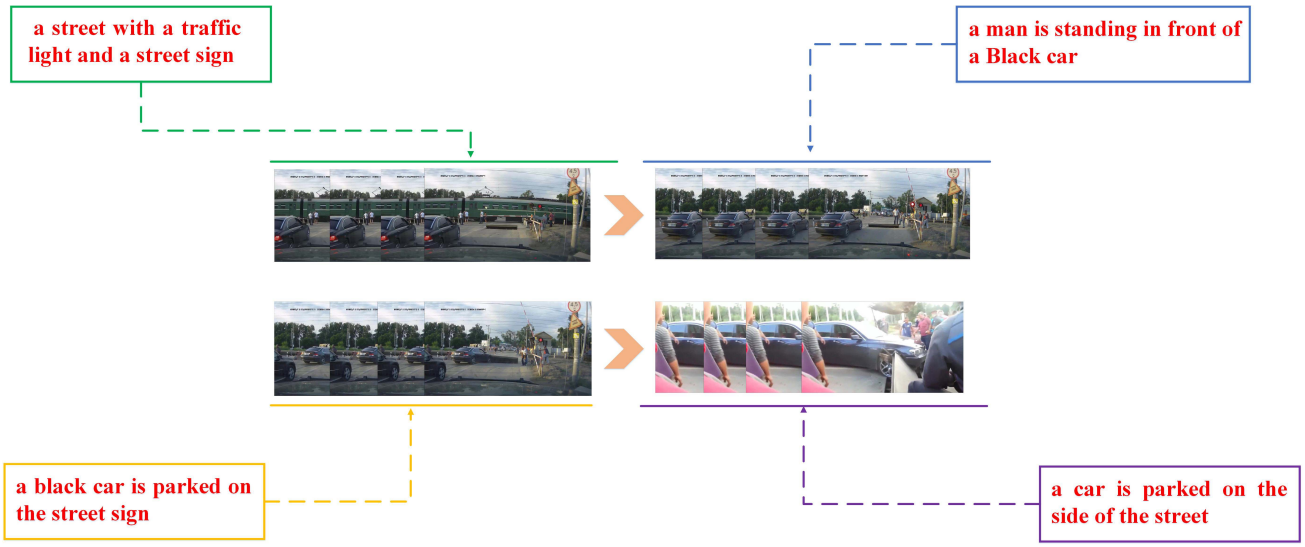


Fig. 7. Steps of extracting event description from several video segments.

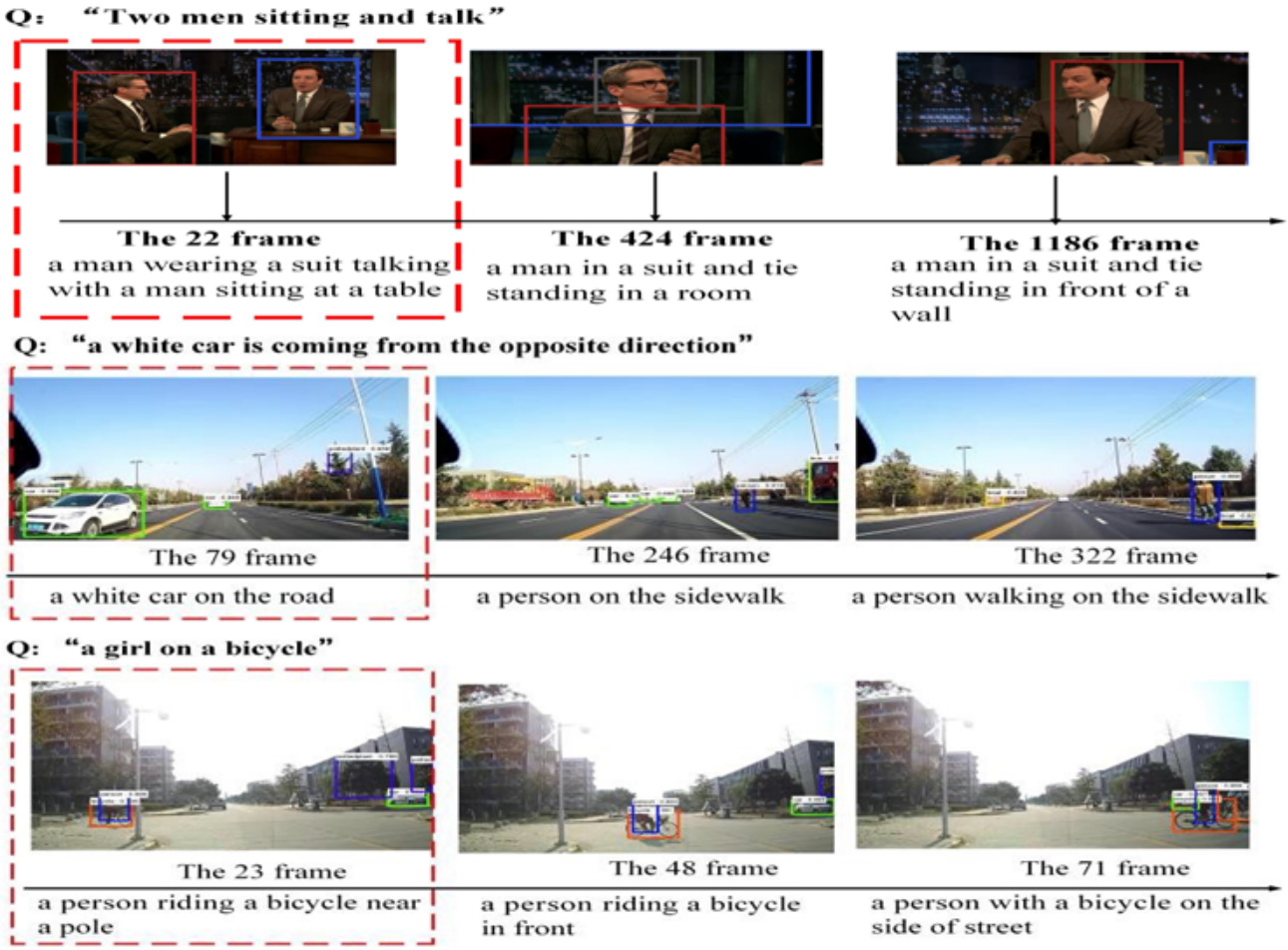


Fig. 8. Process of retrieving and extracting the corresponding video segments from several video segments.

semantic description, and the second is that the question's answer matches the semantic description.

In addition, a quantitative verification is also performed to examine the effectiveness of the VQA model. Comparison is made between the proposed algorithm and the most widely-used question-answering system algorithm on multiple datasets. The performance of the image question-answering system model is mainly evaluated according to Acc and WUPS [33]. Table II compares the experimental results of the proposed algorithm in the standard dataset DAQUAR-ALL. The Acc method is a comparison method referring to image classification problems. As most of the answers are composed

TABLE II
REPORTED RESULTS ON THE DAQUAR-ALL DATASET

DAQUAR-all	Acc(%)	WUPS@0.9	WUPS@0.0
Neural-Image-QA [34]	19.43	25.28	62.00
Multimodal-CNN [18]	23.40	25.59	62.95
Attributes-LSTM [35]	24.27	30.41	62.29
QAM [36]	25.37	31.35	65.89
Bayesian [37]	28.96	34.74	67.33
DPPnet [38]	28.98	34.80	67.81
ACK [39]	29.16	35.30	68.66
ACK-S [32]	29.23	35.37	68.72
SAN [40]	29.30	35.10	68.60
Our proposed	**29.86**	**36.34**	**69.34**

of single or multiple words, the effectiveness of the proposed algorithm can be easily evaluated by examining the accuracy of the words.

V. CONCLUSION

Semantic retrieval of long videos is of paramount importance in the application of traffic video surveillance. This article proposes a long video event retrieval algorithm based on superframe segmentation. By detecting the motion amplitude of the long video, a large number of redundant frames can be effectively removed from the long video, thereby reducing the number of frames that need to be calculated subsequently. Then, by using a superframe segmentation algorithm based on feature fusion, the remaining long video is divided into several SOIs which include the video events. Finally, the trained semantic model is used to match the answer generated by the text question, and the result with the highest matching value is considered as the video segment corresponding to the question.

When preventing and handling traffic safety accidents, people have more requirements for the real-time and accuracy. Processing videos and images at the edge can obviously reduce network bandwidth and lower delay. Therefore, a video preprocessing architecture based on edge computing is presented to remove redundant information of video images, so that partial or all of the video analysis is migrated to the edge or edge server, thereby reducing the dependency for cloud centers, decreasing the computation, storage, and network bandwidth requirements of the network while improving the efficiency of video image analysis. Real-time data analysis and processing play an extremely important role in the prevention of many traffic accidents. The high accuracy and low latency of video analysis tasks require strong computing performance. In order to solve this problem, an architecture of collaborative edge and cloud is proposed, which offloads heavy computing tasks to the edge server or even the cloud, while the small amount of computation tasks are kept locally at the edge. However, some video analysis tasks are long-term and continuous. For example, statistics of traffic volume are used as a reference for the duration of traffic lights, and the demand for delay is not very important. Therefore, for edge computing-driven intelligent transportation video analysis, how to design an efficient integrated cloud, edge and end architecture, perform computing migration at different levels, and reasonably configure edge computing resources is a critical research topic that needs to be solved in the future.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful insights and suggestions which have substantially improved the content and presentation of this article.

REFERENCES

[1] S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prud'Hommeaux, and R. Ptucha, "Semantic text summarization of long videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 989–997.

[2] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2714–2721.

[3] W. Wolf, "Key frame selection by motion analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, May 1996, pp. 1228–1231.

[4] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, "Learning joint representations of videos and sentences with web image search," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 651–667.

[5] S. Ding, S. Qu, Y. Xi, and S. Wan, "Stimulus-driven and concept-driven analysis for image caption generation," *Neurocomputing*, vol. 398, pp. 520–530, Jul. 2020.

[6] S. Ding, S. Qu, Y. Xi, and S. Wan, "A long video caption generation algorithm for big video data retrieval," *Future Gener. Comput. Syst.*, vol. 93, pp. 583–595, Apr. 2019.

[7] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 207–218, Dec. 2014.

[8] Z. Gao, Y. Li, and S. Wan, "Exploring deep learning for view-based 3D model retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–21, Apr. 2020.

[9] S. Tellex and D. Roy, "Towards surveillance video search by natural language query," in *Proc. ACM Int. Conf. Image Video Retr. CIVR*, 2009, pp. 1–8.

[10] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien, "Unsupervised learning from narrated instruction videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4575–4583.

[11] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 17–31, Apr. 2007.

[12] S. Wan, Y. Xia, L. Qi, Y.-H. Yang, and M. Atiquzzaman, "Automated colorization of a grayscale image with seed points propagation," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1756–1768, Jul. 2020.

[13] M. Gygli, Y. Song, and L. Cao, "Video2GIF: Automatic generation of animated GIFs from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1001–1009.

[14] C. Chen, X. Liu, T. Qiu, and A. K. Sangaiah, "A short-term traffic prediction model in the vehicular cyber–physical systems," *Future Gener. Comput. Syst.*, vol. 105, pp. 894–903, Apr. 2020.

[15] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo, "Multi-task deep visual-semantic embedding for video thumbnail selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3707–3715.

[16] A. Sharghi, B. Gong, and M. Shah, "Query-focused extractive video summarization," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 3–19.

[17] S. Yeung, A. Fathi, and L. Fei-Fei, "VideoSET: Video summary evaluation through text," 2014, *arXiv:1406.5824*. [Online]. Available: http://arxiv.org/abs/1406.5824

[18] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," in *Proc. 31th AAAI Conf. Artif. Intell.*, 2016, pp. 3567–3573.

[19] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2296–2304.

[20] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1682–1690.

[21] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2953–2961.

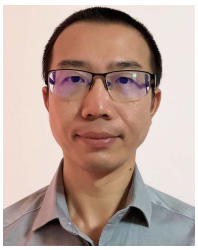
[22] Y. Xi, Y. Zhang, S. Ding, and S. Wan, "Visual question answering model based on visual relationship detection," *Signal Process., Image Commun.*, vol. 80, Feb. 2020, Art. no. 115648.

[23] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu, "Joint video and text parsing for understanding events and answering queries," *IEEE Multimedia*, vol. 21, no. 2, pp. 42–70, Apr./Jun. 2014.

[24] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.
[25] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2006, pp. 419–426.
[26] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using Web-image priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2698–2705.
[27] N. Ejaz, I. Mehmood, and S. W. Baik, "Efficient visual attention based framework for extracting key frames from videos," *Signal Process., Image Commun.*, vol. 28, no. 1, pp. 34–44, Jan. 2013.
[28] D. Teney, P. Anderson, X. He, and A. V. D. Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4223–4232.
[29] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with temporal language," 2018, *arXiv:1809.01337*. [Online]. Available: http://arxiv.org/abs/1809.01337
[30] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 505–520.
[31] G. Guan, Z. Wang, S. Lu, J. D. Deng, and D. D. Feng, "Keypoint-based keyframe selection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 729–734, Apr. 2013.
[32] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image captioning and visual question answering based on attributes and external knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.
[33] C. L. Zitnick, D. Parikh, and L. Vanderwende, "Learning the visual interpretation of sentences," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1681–1688.
[34] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1–9.
[35] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 203–212.
[36] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "ABC-CNN: An attention based convolutional neural network for visual question answering," 2015, *arXiv:1511.05960*. [Online]. Available: http://arxiv.org/abs/1511.05960
[37] K. Kafle and C. Kanan, "Answer-type prediction for visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4976–4984.
[38] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 30–38.
[39] Q. Wu, P. Wang, C. Shen, A. Dick, and A. Van Den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4622–4630.
[40] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 21–29.

Shaohua Wan (Senior Member, IEEE) received the joint Ph.D. degree from the School of Computer, Wuhan University and the Department of Electrical Engineering and Computer Science, Northwestern University, USA, in 2010. Since 2015, he has been holding a post-doctoral position with the State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology. From 2016 to 2017, he was a Visiting Professor with the Department of Electrical and Computer Engineering, Technical University of Munich, Germany. He is currently an Associate Professor with the School of Information and Safety Engineering, Zhongnan University of Economics and Law. He is the author of over 100 peer-reviewed research articles and books. His main research interests include deep learning for the Internet of Things and edge computing.

Xiaolong Xu (Member, IEEE) received the Ph.D. degree in computer science and technology from Nanjing University, China, in 2016. He was a Research Scholar with Michigan State University, USA, from April 2017 to May 2018. He is currently an Associate Professor with the School of Computer and Software, Nanjing University of Information Science and Technology. He has published more than 60 peer-review articles in international journals and conferences, including the IEEE TRANSACTIONS ON INTELLIGENT TRANSACTIONS SYSTEMS (TITS), the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS (TII), the *ACM Transactions on Internet Technology* (TOIT), the *ACM Transactions on Multimedia Computing, Communications, and Applications* (TOMM), the IEEE TRANSACTIONS ON CLOUD COMPUTING (TCC), the IEEE TRANSACTIONS ON BIG DATA (TBD), the IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS (TCSS), the IEEE INTERNET OF THINGS JOURNAL (IOT), the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE (TETCI), the IEEE International Conference on Web Services (ICWS), and ICSOC. His research interests include edge computing, the Internet of Things (IoT), cloud computing, and big data. He received the Best Paper Award from the IEEE CBD 2016, the TOP Citation Award from the *Computational Intelligence Journal* in 2019, the Distinguished Paper Award, and the Best Student Paper of EAI Cloudcomp 2019.



Tian Wang received the B.Sc. and M.Sc. degrees in computer science from the Central South University in 2004 and 2007, respectively, and the Ph.D. degree from the City University of Hong Kong in 2011. He is currently a Professor with the College of Computer Science and Technology, Huaqiao University, China. His research interests include the Internet of Things, edge computing, and mobile computing.

Zonghua Gu (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from the University of Michigan at Ann Arbor under the supervision of Prof. Kang G. Shin in 2004. He worked as a Post-Doctoral Researcher with the University of Virginia from 2004 to 2005, and then as an Assistant Professor with The Hong Kong University of Science and Technology from 2005 to 2009 before joining Zhejiang University as an Associate Professor in 2009. His research interests include real-time and embedded systems. He serves on the editorial board of the *Journal of Systems Architecture*.