A Video Structural Event Description Model for Traffic Surveillance System

Lei Xu[1, a] and Jianxin Song[1, b]

[1]Nanjing University of Posts and Telecommunications, Nanjing 210000, China.

[a]xl12340000@163.com, [b]songjx@njupt.edu.cn

Keywords: video structural analysis, event description, event detection, camshift algorithm

Abstract. The traditional video event detection only identifies a single event for one model and needs to extract features manually to train a mathematical model. In order to automatically extract features of objects and detect various potential accidents in the video surveillance, a structural analysis based approach to detect traffic events is proposed. In the surveillance video, a structural model is used in this way to collect information of the moving targets and the background things, including objects, attributes, temporal relationships and spatial relationships. To further detect the hidden accidents of traffic video, those above five elements are used to make arithmetical logic expression. It can be adapted for different scenarios easily. The experimental results demonstrate that the approach proposed can detect multiple traffic events in surveillance effectively.

1. Introduction

Surveillance cameras around the city will produce a massive video data in 24 hours every day. Especially, the data volume of all video surveillance devices is up to 1 TB every day in Shanghai, China[1]. Thus, it is critical to describe the video potential content accurately and organize video semantic in order to detect and analyze related traffic accidents. Unfortunately, the traditional model for video event detection cannot meet these tasks.

The traffic surveillance data are valuable. The critical information extracted through structural analysis of video data, can provide reliable service for event description, such as traffic violation event detection, vehicle flow detection, real-time vehicle abnormal behavior detection and suspect tracking system. Nevertheless, structural description for surveillance video events is not a simple affair, there are two main issues below[2,3]:

1. Video structural description is an emerging solution applied to video datasets whose structure is beyond the ability of commonly used relational database to process the data in a traditional way. But the surveillance video is unstructured data, it is difficult to carry out further analysis to get the results, so a new model needed to transform video data into structured data.
2. How to describe an event, which is the core of video description technology.Traditional method use mathematical model to extract features from video, which can be used to discriminate traffic event occurred or not, but cant not achieve the level of describing the event.

In this paper, an event oriented model for video structural description is proposed. We build objects, attributes, temporal relationship, spatial relationship and events for a given domain. In order to represent and annotate information better, the objects classification and spatial/temporal relations in events are developed. Mathematical logic expression designed for complex events description is used as well.

The organization of this paper is as follow. In Section 2, the related work of the proposed work is given. The proposed event oriented model for video structural description is given in Section 3. Experiment and result is presented in Section 4. At last, conclusions are summarized.

2. Related Work

The fundamental issue in event description from videos is the representation of the semantic content. Many researchers have done a lot from different aspects. A simple description method may

be associated the video events with low level features(texture, shape, color, etc.) using frames from videos[4]. These ways do not take any relations between features such as spatial or temporal relations. Obviously, using spatial or temporal relations between objects in videos is significant for describing the meaning of events. Researches such as BilVideo[5], extended-AVIS[6], multiView[7] and classView[8] used spatial and temporal relations but do not have structural models for video analysis. Video structural description aims at parsing video content into the text information, which uses spatiotemporal segmentation, feature selection, object recognition, and semantic web technology[9]. Domestic studies on video structural description mostly carried out by Zheng Xu[10,11,12]. Those researches are called VSD, but do not pay much attention to describe events.

3. Our Approach

3.1 Analysis of Traffic Video Structure.

In order to solve the problems of structural description for surveillance video, we have done a lot of related work and get some commonalities on surveillance. With analysis of traffic surveillance video, three features are listed as follows.

1. Surveillance cameras are fixed. There is no scene changing information in video.
2. Objects in traffic surveillance video are divided into moving foreground objects and static background objects.
3. Events rely on the temporal and spatial information intensely.

According to the features of surveillance video data, information of objects such as species information, temporal information, spatial information and color information can be extracted from the original video data. Traffic video surveillance structural description model is built on the above information. Fig. 1 shows an example of information extracted from traffic surveillance video.

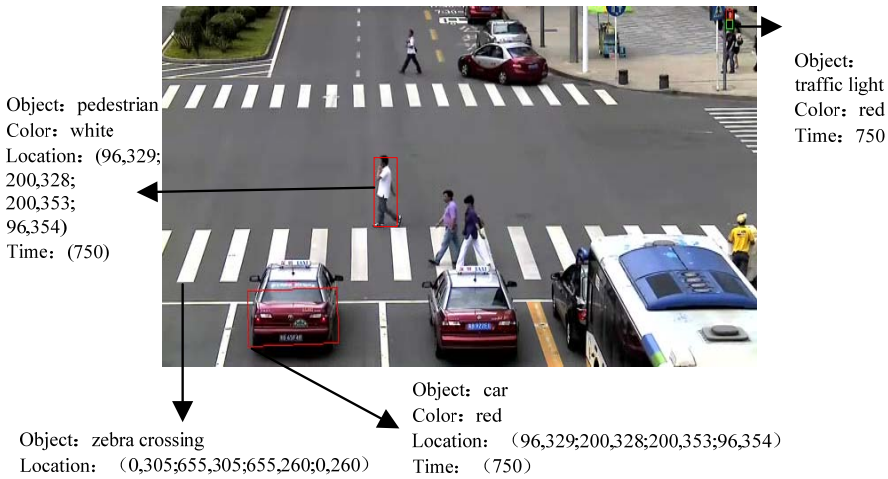

Fig. 1 Information extracted from traffic surveillance video

We make the entire video event structural description model into a hierarchical model with three different semantic data layers, i.e., the objects layer, the attributes layer, and the events layer. Three different layers are described in Fig. 2 as follows.

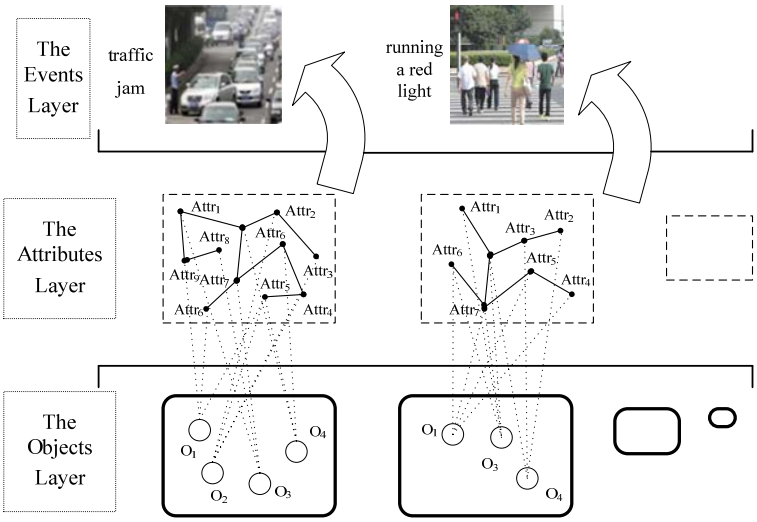

Fig. 2 Three different layers in video structural description model

From Fig 2, the lowest level includes different objects in surveillance video, which are automatically recognized by the pattern recognition. The middle layer includes some basic properties of objects. Each object has both common attributes and unique attributes, and defines several links around attributes.At the highest level, properties relationship can be defined around objects, which directly defines events, and quickly determine whether a certain event occurred or not.

3.2 Definition of Structural Model.

The proposed model is formed by the objects, attributes, spatial relations, temporal relations, and events. Those five basic elements are defined as follows:

Element 1: Objects

Objects element is the most basic unit in the structured description model of video events. In the traffic video, the objects can be vehicles, pedestrians, lane lines, traffic lights, etc. According to the different attributes of the objects in traffic video, objects can be divided into static background objects and dynamic foreground objects (static objects and dynamic objects). Objects can be denoted as:

$$Object = \{so_1, so_2, ..., so_m, do_1, do_2, ..., do_n\} \tag{1}$$

Where so means static objects and do means dynamic objects. m, n means the number of them respectively.

Element 2: Attributes

Attributes element is an important part of the proposed model, which is the description of objects. Spatial attribute and temporal attribute are the most important attribute in the traffic video, which define the basic information of objects. Dynamic foreground objects emphasize on both temporal attribute and spatial attribute, while static objects only focus on temporal attribute due to the fixed location. The remaining attribute such as color, is essential for traffic lights. Attributes can be denoted as:

$$Attribute = \{time, location, color\} \tag{2}$$

$$\{\ do.time = \{t_1, t_2, t_3.....t_n\} \quad do.color = \{c\} \quad do.local = \{l_1, l_2, l_3.....l_n\}\ l_i = (x_{i1}; y_{i1}; x_{i2}; y_{i2})\ \}$$

$$\{\ so.color = \{c_1, c_2, c_3.....c_n\}\ c \in (red, green) \quad so.local = \{l\}\ l = (x_1; y_1; x_2; y_2)\ \}$$

Element 3: Temporal Relations

Temporal relations element is one of the key elements of the model, which is a constraint of two different time points in video, i.e., the relationship between the temporal attribute. Some simple temporal events can be detected by this element. Temporal relations can be expressed as:

$$TR = \{meet, before, after\} \tag{3}$$

$$\{\ \forall meet(t_i, t_j) \rightarrow j - i = 1 \quad \forall before(t_i, t_j) \rightarrow j - i = \tau \quad \forall after(t_i, t_j) \rightarrow j - i = -\tau\ \}$$

Where τ is a positive integer greater than 1, and can be adjusted for different videos.

Element 4: Spatial Relations

Spatial relations element is another key elements for the proposed model, which is a relationship between the spatial position for objects in video, i.e. the relationship between the spatial attribute. Some simple spatial events can be detected by this space constrained property. Spatial relations can be expressed as:

$$SR = \{inside, touch, right, left, above, below\} \tag{4}$$

$$\{ \ \forall inside(l_i, l_j) \rightarrow x_{i1} < x_{j1} \wedge x_{j2} < x_{i2} \wedge y_{i1} < y_{j1} \wedge y_{j2} < y_{i2} \quad \forall touch(l_i, l_j) \rightarrow x_{i2} = x_{j1} \vee y_{i2} = y_{j1}$$

$$\forall right(l_i, l_j) \rightarrow x_{i2} < x_{j1} \quad \forall left(l_i, l_j) \rightarrow x_{i1} > x_{j2}$$

$$\forall above(l_i, l_j) \rightarrow y_{i2} < y_{j1} \quad \forall below(l_i, l_j) \rightarrow y_{i1} > y_{j2} \ \}$$

Element 5: Events

Events element is the core of the proposed model, which is a combination of objects, attributes, temporal relations and spatial relations in surveillance. Events can be denoted as:

$$Event = \{Objects, Attribute, TR, SR\} \tag{5}$$

3.3 Event Structural Description in Proposed Model.

To describe events in a video using the results of structural analysis is the focus of this research. Take an event 'car running a red light' as an example, objects in the event are vehicles and stop line, attributes include color, spatial attribute and temporal attribute. Take advantage of their relationship, this event structural description can be represented by the following block diagram:

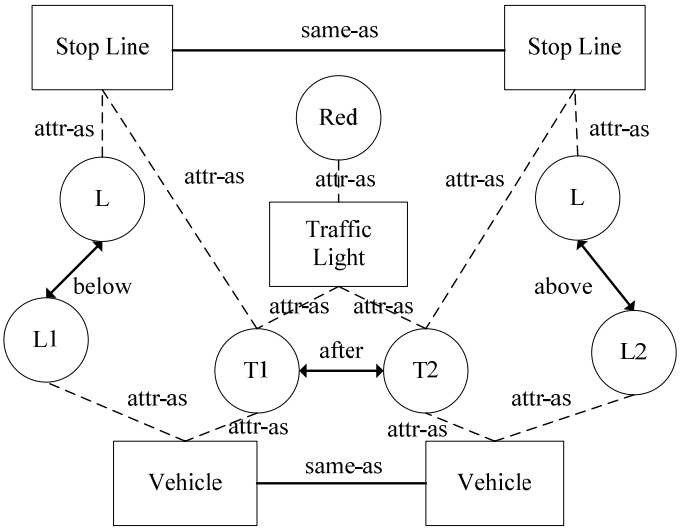

Fig. 3 'Running a red light' Event

4. Experimental Results

4.1 Preparation for Experiments.

Test video is a section of HD traffic video downloaded from youku. Experiments obtain the attributes of objects through the Camshift algorithm for object tracking, using VS2010 and Opencv 2.4.10. Other experimental settings are listed as follows:

1. Take 25 frames as a time interval($\triangle t$).
2. The location of stop line set as (0,295;655,295;;).
3. τ set 3 in temporal relation.

4.2 Vehicle Driving Event Description Experiment.

A method for vehicle driving event description is noted as follows:

Step 1: At the time (t_i) , Using object tracking algorithm (Camshift) to get the car object (do) and traffic light object (so), meanwhile recording their attributes.

Step 2: Update the attributes of *do* and *so* at the time (t_{i+1}), When meet the condition $same(do.l_1, do.l_{1+\tau})$, determined as the car stopping. When meet the condition $below(do.l_1, do.l_{1+\tau})$, determined as the car passing the crossroad.

Step 3: Set $i = i + 1$, go to *Step 2*.

Take the car on left side as object do , descriptions of vehicle driving event are noted as follows:

Table 1 Results of vehicle driving event description

Time	Spatial Attribute	Color	Event
t_1=705	l_1= (98,332;199,330;200,352;99,354)	c_1= red	Car Stopping
t_2=730	l_2= (98,332;198,330;199,352;99,354)	c_2= red	
t_3=755	l_3= (96,329;200,328;200,353;96,354)	c_3= red	
t_4=780	l_4= (97,329;199,328;199,353;97,354)	c_4= red	
t_5=805	l_5= (97,329;199,328;199,353;97,354)	c_5= red	
t_6=830	l_6= (95,329;196,328;196,353;95,354)	c_6= red	
t_7=855	l_7= (97,328;196,327;196,350;97,351)	c_7= green	Car Passing
t_8=880	l_8= (103,305;209,301;210,323;104,327)	c_8= green	
t_9=905	l_9= (120,266;211,263;212,285;121,288)	c_9= green	
t_{10}=930	l_{10}= (128,231;209,224;211,245;130,252)	c_{10}= green	
t_{11}=955	l_{11}= (132,195;205,191;206,211;133,215)	c_{11}= green	
t_{12}=980	l_{12}= (134,161;199,157;200,175;135,179)	c_{12}= green	
t_{13}=1005	l_{13}= (145,130;196,128;197,144;146,146)	c_{13}= green	
t_{14}=1030	l_{14}= (158,102;204,100;205,116;159,118)	c_{13}= green	

The results show that the car stopped before time t_7 and then began to move forward. With noting the color of traffic lights simultaneously, we can speculate that the car passed the road normally.

4.3 Running Red Light Event Description Experiment.

A method for 'running red light event' description is noted as follows:

Step 1: At the time (t_i) , Using object tracking algorithm (Camshift) to get certain dynamic object (do) and traffic light object (so), meanwhile recording their attributes.

Step 2: Update the attributes of *do* and *so* at the time (t_i t_{i+1} ... $t_{i+\tau}$), When meet the condition $(left(do.l_i, so.l) \wedge so.color = red) \wedge (inside(do.l_{i+\tau}, so.l) \wedge so.color = red)$, determined as running red light event occured.

Step 3: When the condition is not satisfied, set $i = i + 1$, go to *Step 2*.

Test 19 dynamic objects in three periods, the results show that 15 objects determined correctly. A futher analysis of more events will be done in the future.

5. Summary

Surveillance cameras around the city will produce a large volume of video data in 24 hours every day. It is important to describe the video potential content accurately and organize video semantic in order to detect and analyze related traffic accidents. The proposed model for video structural event description is formed by the objects, attributes, spatial relations, temporal relations, and events. In addition, we define mathematical logic expression to detect event efficiently. A more complete analysis of video structural event description is suggested for further studies.

References

[1]. Xu Z, Mei L, Liu Y, et al. Video structural description: a semantic based model for representing and organizing video surveillance big data[C]//Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on. IEEE, 2013: 802-809.

[2]. Wu L, Wang Y. The process of criminal investigation based on grey hazy set[C]//Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on. IEEE, 2010: 26-28.

[3]. Liu L, Li Z, Delp E J. Efficient and low-complexity surveillance video compression using backward-channel aware Wyner-Ziv video coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2009, 19(4): 453-465.

[4]. Xu Z, Mei L, Liu Y, et al. Semantic enhanced cloud environment for surveillance data management using video structural description[J]. Computing, 2016, 98(1-2): 35-54.

[5]. Dönderler M E, Şaykol E, Arslan U, et al. BilVideo: Design and implementation of a video database management system[J]. Multimedia Tools and Applications, 2005, 27(1): 79-104.

[6]. Sevilmiş T, Baştan M, Güdükbay U, et al. Automatic detection of salient objects and spatial relations in videos for a video database system[J]. Image and Vision Computing, 2008, 26(10): 1384-1396.

[7]. Fan J, Aref W G, Elmagarmid A K, et al. MultiView: Multilevel video content representation and retrieval[J]. Journal of electronic imaging, 2001, 10(4): 895-908.

[8]. Fan J, Elmagarmid A K, Zhu X, et al. ClassView: hierarchical video shot classification, indexing, and accessing[J]. IEEE Transactions on Multimedia, 2004, 6(1): 70-86.

[9]. Xu Z, Liu Y, Mei L, et al. Semantic based representing and organizing surveillance big data using video structural description technology[J]. Journal of Systems and Software, 2015, 102: 217-225.

[10]. Xu Z, Zhi F, Liang C, et al. Semantic annotation of traffic video resources[C]//Cognitive Informatics & Cognitive Computing (ICCI* CC), 2014 IEEE 13th International Conference on. IEEE, 2014: 323-328.

[11]. Hu C, Xu Z, Liu Y, et al. Video structural description technology for the new generation video surveillance systems[J]. Frontiers of Computer Science, 2015, 9(6): 980-989.

[12]. Xu Z, Hu C, Mei L. Video structured description technology based intelligence analysis of surveillance videos for public security applications[J]. Multimedia Tools and Applications, 2015: 1-18.