Robust Traffic Event Extraction via Content Understanding for Highway Surveillance System

Akio Yoneyama
Multimedia Communications Lab.
KDDI R&D Laboratories Inc.
Saitama 356-8502, JAPAN
E-mail: yoneyama@kddilabs.jp

Chia H. Yeh and C. -C. Jay Kuo
Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089-2564, USA
E-mail:{chyeh, cckuo}@sipi.usc.edu

Abstract

A method to extract traffic events by integrating the low-level, middle-level, and high-level feature extraction modules is developed in this research. The low-level module extracts features such as motion, size, and location. The middle-level module builds a bridge between the road surface plane in the real world and the captured image plane via geometric analysis. Finally, the high-level module identifies traffic events such as “traffic jam”, “lane change”, and “traffic rule violation”, which require the understanding of video contents in a specific knowledge domain. In the high-level module, various traffic events are related to motion characteristics obtained from the middle-level module. It is demonstrated by experimental results that the proposed system can achieve robust traffic event extraction.

1. Introduction

Vision-based traffic monitoring systems are widely used in the field of Intelligent Transportation Systems (ITS) because of their powerful ability to extract a variety of information. Traffic event extraction and its corresponding scene analysis have become crucial and attracted much attention for various applications. The goal of traffic monitoring systems is to extract the traffic information such as the vehicle count, traffic events, and traffic flows. Although vision-based traffic monitoring systems have good potential to extract a variety of traffic information, it is still a challenging problem to maintain detection accuracy at all time since vision-based processing is sensitive to environmental factors such as shadow, lighting and occlusion [1][2].

Robust image/video processing is essential to vehicle (object) extraction and tracking. There has been a lot of research work performed along this direction. However, it is desirable to develop a video analysis algorithm to achieve self-calibration based on observed video contents. To reach this goal, there are two important problems to be addressed. First, we should have a clear understanding of the relationship between the vehicle motion on the road in the real world and the captured video content. Second, we should relate traffic events to motion characteristics of the vehicle. Then, traffic events can be determined by matching the predefined vehicle motion pattern with the detected vehicle motion information from the captured video. This is actually a model-based approach. We examine the above two problems by proposing a system that contains three levels: “low-level feature extraction”, “middle-level content semantic analysis” and “high-level traffic content understanding and traffic event extraction.”

We have done a significant amount of work on the low-level feature extraction for highway surveillance system [3]-[5]. In this paper, our focus will be on the middle-level and the high-level processing for traffic event extraction. The motivation of this work aims at extracting traffic events in the scene. We would like to detect “traffic congestion”, “rule violation”, etc. Since traffic events have semantic interpretations, we need to build models for traffic events and relate them to features extracted at the middle-level.

2. Review of Previous Work

Many methods have been proposed to handle errors arising from shadow, occlusion and nighttime detection. In particular, the background subtraction method provides a simple and useful solution to the moving region extraction. Solutions that adopt a frequent update of the background image were also proposed to cope with gradual environmental change and noise. However, regions extracted from the background subtraction method include not only the moving object of interest (i.e. the vehicle) but also the associated moving cast shadow. Some further processing step is needed to eliminate the cast shadow from the extracted moving region. We have investigated various techniques to cope with the above problems and presented a sequence of interesting results in [3]-[5]. These three techniques belong to the low-level image processing, and form the foundation of our current research work. In this research, we would like to obtain high-level semantic traffic events by organizing these low-level features extracted using the techniques presented in our previous work.
3. Proposed Methodology and System

In this research, we propose a system to extract traffic events from video without the knowledge of capturing conditions. Let us give some assumptions first. Both single-camera and multiple-camera solutions are considered. Cameras are set in the nearly horizontal direction so that the camera can capture the traffic in the long range. No geometric information of the real world for captured images is needed. Several traffic events are considered such as lane change, traffic jam and traffic accident. Figure 1 shows the block diagram of the proposed system. The system consists of low-level, middle-level, and high-level feature extraction modules with the help of a traffic rule translation building block. They are described below in detail.

1. **Low-level feature extraction module**: This module detects and tracks objects in the image plane. The output of this module is the tracking object information on the image plane such as motion, size, and location.

2. **Middle-level feature extraction module**: This module attempts to analyze the relationship between the input image plane and the road surface plane in order to coincide with the information from middle-level module.

3. **High-level feature extraction module**: This level matches the information between the middle-level feature extraction module and the traffic rule translation sub-module. The output of this module is the traffic event information and another output is the processing parameters that treated as the feedback to the low-level feature extraction module.

4. **The traffic rule translation sub-module**: This sub-module translates the traffic rules into its corresponding motion information. This translation works according to the analysis of the relationship between input image plane and the road surface plane in order to coincide with the information from middle-level module.

3.1 Low-level feature extraction

The low-level module detects moving objects from captured images. We use the simplified cuboid model $M$ for vehicle in the 3D space that is composed of the width, height and length of a vehicle. The model has six vertices. Moreover, we apply the same model to the joint region of vehicle and its cast shadow region, which is called the “joint vehicle/shadow model”. This model is classified into six types in terms of the geometric relationship among the vehicle, the light source, and the camera as shown in Fig. 2 and more details are in [3]. Then, we can obtain the location and the dimension of each moving object region with moving cast shadow eliminated. The output of this module is the information of tracked objects on the image plane such as their motion, size, and location.

3.2 Middle-level feature extraction

The middle-level module attempts to analyze the relationship between the input image and the road surface in the real world. The output of this module is the information of tracked objects on the road surface obtained from the input image. Note that we do not have the relationship between the real world and the image plane since no knowledge of capturing conditions is available. The most difficult part is that we do not have the line that is perpendicular to the traffic lanes on the road surface, and have to do the approximation of the relationship between the real world and the image plane to solve this problem.

The perspective projection $P$ from points $p_v = (x_v, y_v)$ on the 2D image plane $I_p$ to their corresponding points $p_r = (x_r, y_r)$ on the road surface plane $I_r$ can be represented as $P_r = P(p_v)$, where

![Figure 1: The proposed system for traffic event extraction.](image1)

![Figure 2: Six types of the joint vehicle/shadow model.](image2)
For the above perspective projection, at least four-pair coordinates of the corresponding points between the 2D image plane and the projected road surface plane are needed to determine all parameters in Eq. (1). One solution is to choose rectangular blocks using painted lane marks on the road. However, this approach may not be suitable if the camera cannot detect the lane painting clearly. We propose an alternative method with the reduced number of pair coordinate using three points $p_{p1}$, $p_{p2}$, and $p_{p3}$, with an approximation as shown in the middle part of Fig. 3.

Then, the projection function Eq. (1) can be rewritten with the reduced set of parameters as

$$x_r = \frac{c_{i0} \cdot x_p + c_{i2} \cdot y_p + c_{i3}}{c_{i1} \cdot y_p + 1}, \quad y_r = \frac{c_{i4} \cdot y_p + c_{i5}}{c_{i1} \cdot y_p + 1}. \quad (2)$$

The locations of $p_{p1}$, $p_{p2}$, and $p_{p3}$ can be obtained by passing vehicles in the image and required to determine projection parameters in Eq. (2). The effect introduced by the approximation appears as the error in the vertical coordinate of vehicles on the projected image plane. The parameters for the projection are given in Fig. 4. We would like to measure the error ratio $y'_r/y'_s$ on the projected image plane.

The approximation is to treat the angle difference $\xi$ as zero. The error on the right side is the largest in the image and the value can be represented as

$$r = \frac{y'_r}{y'_s} = \frac{\sin(\theta + \phi)\sin(\xi)}{\sin(\phi + \xi)\sin(\theta)}. \quad (3)$$

### 3.3 High-level feature extraction

Based on the above study, traffic events can be detected by the combination of the middle-level feature extraction module and the traffic rule translation sub-module. The traffic rule translation sub-module analyzes the traffic scene information such as the location of each lane and the number of lanes. Although there are several approaches for lane information acquisition by texture-based analysis, we do not employ these approaches. This is because texture-based approaches cannot be robust to any kind of road situation. Instead of using them, we obtained the lane location information by accumulating the center positions of detected vehicles since most of vehicles keep the same traffic line.

After the recognition of the traffic scene, the traffic rule translation sub-module translates the traffic rules into the motion and location information on the road plane. Let us study lane change and traffic jam cases as examples. The target vehicle will be detected as each of above mentioned events according to the following criteria.

i) Each vehicle is classified into one of the detected lanes by finding the nearest lane from the center of the vehicle. If the classified lane in the current frame is different from that in the previous frame for a vehicle, it is regarded as the event of lane change.

ii) The traffic jam is determined when the average speed on the lane is lower than a predefined threshold obtained by averaged speed in a day.

Although these conditions for each traffic event need to be defined in advance, the detection of each traffic event can be done automatically once these conditions are settled. Moreover, the high-level feature extraction module calculates parameters, whose calculation formula is fed by the traffic rule translation sub-module, for each passing vehicle. If all of the conditions for an event are met for a target vehicle, it is claimed to be the event detected for the vehicle. Thus, we can extract various kinds of traffic events accordingly.

### 4. Experimental Results

The experimental results of traffic event extraction are shown in this section. Figures 5(a) and 5(b) show results of vehicle detection in the low-level and the middle-level feature extraction modules, respectively. In Fig. 5(a), the low-level feature extraction module detects moving vehicles with their moving cast shadows eliminated. Moreover, each detected vehicle is tracked on the frame base. Red lines in Fig. 5(a) represent tra-
jectories of detected vehicles. The information obtained in this stage is the motion and the size of detected vehicles in the captured image plane.

The projection of the detection result onto the road surface plane is shown in Fig 5(b). This image represents road surface plane, which corresponds to the right-most rectangle in Fig. 3. Although we cannot obtain the line that is perpendicular to traffic lanes on the road surface, the error introduced by the approximation mentioned in the middle-level feature extraction part is less than about 5% as long as the vanishing point locates within the captured image rectangle.

Figure 6 shows the result of the traffic lane detection in the traffic rule translation sub-module. The projected road plane is the same as Fig. 5(b). Traffic lanes are obtained by accumulating the temporal difference. In this case, the duration of this accumulation is approximately 20 seconds. In this figure, five lanes in the middle of the plane and one in the left side are detected as traffic lanes. Then, the position of the lane center position for each lane is obtained.

Two examples of traffic event extraction are given in Figs. 7 and 8. The first example given in Fig. 7 is the lane change detection. In the case of lane change detection, the nearest lane center for each vehicle is examined on the road plane shown in Fig. 7(b) in every frame using the result in Fig. 6. The nearest lane center for the vehicle at the center of Fig. 7(b) changed in the frame. Consequently, it is regarded as a lane change event for the target vehicle. The second example is the traffic jam detection as shown in Figs. 8(a) and 8(b). Since the traffic lane information is analyzed beforehand, the system can detect lane-based traffic information. The traffic jam is detected as the slower speed of passing vehicles than that of a long-term average.

5. Conclusion

In this work, we propose a new method to extract traffic events by integrating the low-level, middle-level, and high-level feature extraction modules for highway surveillance systems. The main feature of our work is that no prior information of capturing conditions is required. By relating traffic events to motion rules, the proposed system can extract traffic events robustly. In the near future, we would like to extend our current work to extract other traffic events under more challenging capturing conditions such as at night and in the rain.

6. References