Neurocomputing 151 (2015) 700–709

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Integrating 3D structure into traffic scene understanding with RGB-D data

Yingjie Xia [a], Weiwei Xu [a,*], Luming Zhang [b], Xingmin Shi [a], Kuang Mao [c]

[a] Hangzhou Normal University, China
[b] School of Computing, National University of Singapore, Singapore
[c] College of Computer Science, Zhejiang University, China

ARTICLE INFO

Article history:
Received 15 November 2013
Received in revised form
25 March 2014
Accepted 26 May 2014
Available online 31 October 2014

Keywords:
Traffic scene understanding
Depth data
3D structure
Vehicle detection
Pedestrian detection
Overtaking warning

ABSTRACT

RGB Video now is one of the major data sources of traffic surveillance applications. In order to detect the possible traffic events in the video, traffic-related objects, such as vehicles and pedestrians, should be first detected and recognized. However, due to the 2D nature of the RGB videos, there are technical difficulties in efficiently detecting and recognizing traffic-related objects from them. For instance, the traffic-related objects cannot be efficiently detected in separation while parts of them overlap, and complex background will influence the accuracy of the object detection. In this paper, we propose a robust RGB-D data based traffic scene understanding algorithm. By integrating depth information, we can calculate more discriminative object features and spatial information can be used to separate the objects in the scene efficiently. Experimental results show that integrating depth data can improve the accuracy of object detection and recognition. We also show that the analyzed object information plus depth data facilitate two important traffic event detection applications: overtaking warning and collision avoidance.

1. Introduction

In intelligent transportation systems (ITS), traffic flow is one of the most used indices for characterizing traffic conditions to be used in traffic control and transportation management [44]. Traditionally, the data of traffic flow are collected by inductive loop detectors [12], global positioning system (GPS) probe vehicles [31], and remote traffic microwave sensors [43]. However, all these detection devices have their inherent drawbacks [14]. The major disadvantages of inductive loop detectors are high failure ratios and high maintenance costs. The main shortcomings of GPS probe vehicles are poor statistical representation and high error rates in the map-matching, and the main disadvantages of RTMS are high installation costs and inaccurate estimation of traffic state features.

Recently, video devices have been widely deployed for traffic surveillance. The video detectors become the primary sensor to detect traffic flow from roadside or overhead mainly for the following reasons [22]: (1) People are more used to visual information than other forms of sensor data; (2) Video sequences can directly reflect the status of transportation systems by a broad time-varying range of information; (3) Video detectors can be installed, operated, and maintained easily and in low cost. Therefore, the detection, recognition, and tracking of the traffic-related objects, such as vehicles and pedestrians from the captured videos provide the critical basis for ITS applications [41,30].

Significant improvements in traffic scene understanding have been achieved in such 2D image representation based algorithms. However, there are still technical issues remaining to be solved in practice. First, the traffic-related objects cannot be efficiently detected in separation while parts of them overlap; Second, complex outdoor environments increase the difficulty to the vehicle and pedestrian detection since object detection will be influenced by the background; Moreover, it is difficult to design a system robust to detect vehicle movement and drift with 2D image representation.

With the popularization of RGB-D camera, users can now have low-cost and easy-to-use devices, such as Microsoft Kinect, to capture 3D representation of a scene in the format of depth data [11]. Therefore, recent researches in computer vision community have made great efforts on improving the robustness and accuracy of object localization and recognition by integrating 3D representation into the analysis pipeline. Local geometry features from depth data are used to analyze and segment the indoor scene images in high accuracy [35]. In [5], depth kernel descriptors was developed to improve the object recognition accuracy. A recent contribution also investigates how to accurately localize the 3D objects with the assistance of depth data [23]. Inspired by these pioneering research works, it is worth investigating how to use

* Corresponding author.
E-mail address: Weiwei.xu.g@gmail.com (W. Xu).

http://dx.doi.org/10.1016/j.neucom.2014.05.091
0925-2312/

depth data in the traffic scene understanding algorithms to handle the above technical issues.

The major contribution of this paper is a robust, RGB-D data based traffic scene understanding algorithm. The 3D structure information of a traffic scene is captured by Microsoft Kinect. The algorithm starts with the computation of local 2D plus 3D features for the captured RGB-D data. Afterwards, the random forest algorithm is adopted to learn an efficient pixel-level classifier from the features as the basis to low-level understanding of traffic scene [6]. A segmentation and labeling algorithm based on graph-cut is then used to segment the RGB-D images into object-level, which is ready for various high level applications in traffic surveillance.

We have tested our algorithm on a variety of traffic scene images which contains different kinds of traffic objects, such as car, bicycle and pedestrians. Experimental results show that depth data can largely improve the object detection accuracy and facilitate the subsequent high-level traffic surveillance applications.

2. Related work

2D image based traffic scene understanding: The kernel of traffic scene understanding is traffic-related object detection, recognition, and analysis, including vehicle detection [41], pedestrian detection [30], license plate recognition [2], and pedestrian counting [39].

The detection, recognition, and analysis of vehicles and pedestrians have broad applications in ITS. Vehicle detection and recognition are used for identifying cases of traffic violation, which is the main cause of traffic accidents [29]. One of the vehicle detection methods is designed to divide video frames into sub-regions and extract local features from sub-regions to enable the detection less susceptible to the variance of vehicle poses, shapes, and angles [41]. In order to precisely separate a vehicle with its neighboring vehicles, Sivaraman and Trivedi integrate active-learning and particle filter tracking to implement an on-road vehicle detection system [37]. Cherng et al. propose a dynamic vehicle detection model which visually analyzes the critical motions of nearby vehicles in video [9]. However, these work have not efficiently solved some special cases, such as vehicle overlapping happens. To detect the vehicles in complex traffic scenes is very useful for multiple ITS applications.

The pedestrian detection is also very important for the effective traffic scene understanding. For example, pedestrian detection can reduce the occurrence of pedestrian-and-vehicle-related cases, such as collision accidents. Cao et al. use a classifier to identify the risky regions based on vehicles from the video data, and evaluate the risk of pedestrians by the estimated distances between pedestrians and risky regions to avoid accidents [7]. Munder et al. utilize a Bayesian method on multiple features of shape, texture, and 3D information to detect pedestrians in urbans [30]. In night time, Ge et al. use a monocular near-infrared camera to detect and track pedestrians in real-time [17]. Night-vision systems are also used to model the pedestrian detection by the probability calculated through a function with various pedestrian features [4]. In these related work, RGB-D camera can be effectively used to detect pedestrians because it uses infrared light to capture depth information, while it still has not been employed for pedestrian and vehicle identification in their mixed occurrence.

From video data, license plate recognition is a basic module in ITS aiming to identify and locate the vehicle. Typical license plate recognition consists of two steps, license plate location and characters recognition. Morphological and chromatic processing on frames of traffic video is widely used in license plate location. The morphological processing is based on morphological features of license plates [21]. Some approaches utilize histograms of gray-scale images, which are not available when incomplete characters exist or the background is too complicated [24]. As for chromatic processing, some work uses the specific color to locate the license plate region, while it is fragilely interfered by the illumination changes and other similar colors in the image [1]. In the characters recognition, the template matching method is widely used. This method does not work well for the images with a lot of noise, and the recognition results heavily depend on the chosen templates [8]. Neural network is another common-used approach to recognize characters of license plates [28].

As the statistical traffic data analysis on video data, pedestrian counting is particularly useful in some special cases, such as emergency evacuation [10]. Video is a low-cost and effective device to implement the pedestrian counting. Zhang et al. extract high-dimensional statistical features from the pedestrian video data, and adopt the supervised dimension reduction technique to select the representative features [18,45]. Tan et al. propose a semi-supervised elastic net model based on the relationship between each frame and its neighboring frames to achieve pedestrian counting [39].

In addition to the aforementioned related work, the traffic signs [3] and lanes detection [26] by traffic videos are also very important for ITS applications. Since 2D cameras have been widely deployed on roads, there are few applications using 3D traffic video data. However, the depth information is very useful under some special circumstances, such as detecting the overlapping traffic-related objects. It is valuable to investigate how to use RGB-D data to interpret traffic scenes in ITS applications.

RGB-D data based scene understanding: Depth data can be exploited to the learning of discriminative object features and the analysis of the 3D scene structure, which has proven to be successful in scene understanding applications.

A typical application of RGB-D data is indoor scene understanding. Silberman et al. [35] collected a database of indoor scene RGB-D images, and developed RGB-D SIFT descriptor to improve the segmentation and labeling accuracy of indoor scene RGB-D images. Koppula et al. learned a highly accurate indoor scene object classifier through mixed integer optimization [27], and achieved around 80% accuracy of depth data labeling. In computer graphics, RGB-D data has been used in 3D indoor scene reconstruction applications. [36,25,34]. Depth data is important in correct analysis and reconstruction of the 3D indoor scene layout in such applications. Besides geometric properties, there is also research work on how to derive physical interactions between objects in the scene with depth data, such as structural stability and supporting relationship analysis [50]. These algorithms can be combined with super-pixel or graphlet representation to accelerate the RGB-D image segmentation [32,46,48,49].

RGB-D data can also be used in object recognition and retrieval [40,16]. Research efforts have been devoted to view-invariant 3D shape or depth data feature descriptors [42,15]. Histogram of oriented depth is used in human detection in RGB-D images [38]. Depth kernel descriptors developed by Bo et al. applies match kernel to the local patch based geometric features to generate highly discriminative and robust geometric features [5]. Integrating it into various classifier algorithms, such as support vector machine and random forest, results in more accurate object recognition results. With the assistance of depth data, accurate 3D object localization in captured RGB-D images can be realized by learning a segmentation mask in the 2D bounding box of an extracted object in the image [23].

3. Traffic scene segmentation and labeling algorithm

The goal of segmentation and labeling is to obtain an object-level traffic scene image understanding. That is, the captured