VEHICLES SEGMENTATION VIA SPATIOTEMPORAL SALIENCY IMPROVEMENT IN CROSSROADS MONITORING

WEI LI[1], LING ZHANG[1], HUA-AN ZHAO[2]

[1] Department of Transportation and Logistics Engineering, Shandong Jiaotong University, Jinan 250023, China
[2] Department of Computer Science and Electrical Engineering, Kumamoto University, Kumamoto 860-8555, Japan
E-MAIL: liwei1195@163.com, campuls@sina.com, cho@cs.kumamoto-u.ac.jp

Abstract:

This paper proposed a moving vehicles segmentation approach combined two kinds of classifiers which are Gaussian mixture model and spatiotemporal saliency map. We extracted new spatial and temporal saliency features and improved spatiotemporal consistency optimization model to calculate more exact saliency map and to speed up the processing. Because misclassification appears when Gaussian background cannot be correctly updated on time, we combined saliency map and Gaussian mixture model foreground to solve the problem. Experimental results showed that, compared to current methods, our approach had higher efficiency and accuracy.

Keywords:

Foreground segmentation; Gaussian mixture model; Spatiotemporal saliency map; Classifier combination

1. Introduction

Every year a lot of traffic accidents happened at crossroads around the world, crossroads monitoring systems become more and more important in traffic safety management. With the popularity of intelligent traffic monitoring, image processing techniques are essential for many purposes such as vehicle detection or objects tracking, etc.. Crossroads monitoring videos have complex and easily changed backgrounds, which make it very difficult to exactly detect moving vehicles in scenes. Now a commonly used method to detect objects in complex scenes is GMM (Gaussian Mixture Model), which can adaptively update the background [1,2]. GMM is robust when there is slight movement or light change in background. However, when there are dramatic changes in the scene, especially when car moving out of the scene and the color of the car is similar to the background, the updating of GMM background will be delayed, which results in "afterimage" of the left car. That is an innate disadvantage of GMM which can not be fundamentally improved.

Recently, research on visual saliency detection is attracting many interests [3-6]. Researchers try to recognize what objects are most noticeable in images and videos, which are called salient objects. Salient objects in videos should have spatiotemporal features that can be extracted for foreground segmentation. In crossroads monitoring videos, moving cars are obvious salient objects, we could detect them by using saliency map, which is always used to segment foreground and background. Since it does not need background updating in saliency detection, it may solve the problem of "afterimage" caused by GMM method. Based on this idea, we proposed vehicles detection via spatiotemporal saliency improvement, an overview of our approach can be seen in Figure 1.

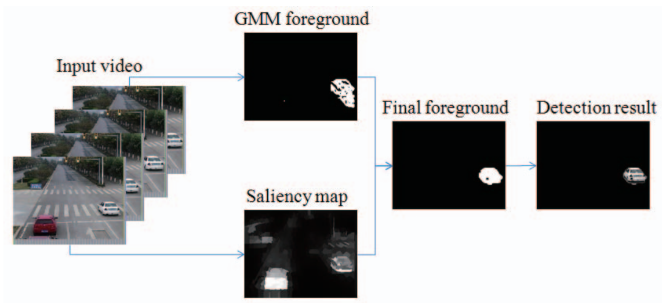

FIGURE 1. Overview of vehicle segmentation enhanced by saliency map

2. Spatiotemporal saliency detection

Most of the video saliency detection models combine spatial and temporal salient features. Zhou et al. [4] carried out superpixel segmentation on each image, and then measured spatial saliency and temporal saliency using appearance contrast and motion contrast of the superpixels. Finally, spatial saliency, temporal saliency, and prior location were linearly fused to obtain spatiotemporal saliency map. Wang et al. [5] extracted edge and optical flow gradient for objects location, then calculated the

978-1-5386-0408-3/17/$31.00 ©2017 IEEE

geodesic distance between superpixels to get spatial saliency and temporal saliency. At last, an energy function was applied to combine spatiotemporal saliency, appearance, and location to get the final spatiotemporal saliency map. Zheng et al. proposed a novel spatiotemporal consistency optimization model as well as a new spatiotemporal saliency region detection model, which is robust with complex scenes and motion conditions [6]. Our work was mainly based on [6], we proposed an improved spatiotemporal consistency optimization model, which enhanced the contribution of temporal saliency at producing the final saliency map. Therefore, moving objects in videos could be detected easier.

2.1. Spatiotemporal consistency optimization model

Spatiotemporal consistency is defined as that, regions that have similar spatial or motion feature should also have similar saliency performance, the higher of similarity, the more smoothness of saliency. This characteristic could be applied to optimize the original saliency features such as edge and optical flow gradient. This optimization model can be formulated as follow.

$$\min L = \lambda_1 B + \lambda_2 F + \lambda_3 V + \lambda_4 W \tag{1}$$

As shown in Equation (1), the optimization model contains 4 constraints, background constraint B, foreground constraint F, intra-frame consistency constraint V, and inter-frame consistency constraint W, which are separately defined in Equation (2-5).

$$B = \sum_{i=1}^{N} \omega_i^b s_i^2 \tag{2}$$

$$F = \sum_{i=1}^{N} \omega_i^f (s_i - 1)^2 \tag{3}$$

$$V = \sum_{ij} \omega_{ij}^{intra} (s_i - s_j)^2 \tag{4}$$

$$W = \sum_{ik} \omega_{ik}^{inter} (s_i - s_k)^2 \tag{5}$$

In the above equations, N is the number of superpixels, s_i and s_j are intra-frame adjacent, s_i and s_k are inter-frame adjacent. In background constraint B, superpixels of background have lower saliency value, while in foreground constraint F, superpixels of foreground have higher saliency value. Intra-frame consistency constraint V makes superpixel s_i and s_j having smooth saliency value, similarly, inter-frame consistency constraint W makes superpixel s_i and s_k having smooth saliency value. λ_1, λ_2, λ_3, λ_4 are weights that should be estimated.

As described in [6], the procedure of saliency detection is that, firstly, spatial features and temporal features are extracted and then separately optimized by this model to get spatial saliency value and temporal saliency value, next, spatial saliency and temporal saliency are multiplied and optimized by the spatiotemporal consistency model again, finally, saliency map based on superpixel segmentation can be obtained. The procedure is complex and time-consuming.

In our research, we found an object that has high spatial saliency value may be not a moving object. Especially in traffic videos, objects like signal lights, road signs, and stopped vehicles are easily labeled as salient objects, but these are not our targets. So we improved this consistency optimization model, the new approach could reduce the influence of spatial saliency to the final spatiotemporal saliency map. Meanwhile, we only carried out saliency optimization one time, therefore the speed is much faster than Zheng's method.

2.2. Spatial saliency and temporal saliency calculation

It has been proved that boundary connectivity and compactness are efficient in measuring spatial saliency. Boundary connectivity *bdc* shows how much a superpixel connects to image boundary. If superpixel has higher *bdc*, it is more likely that this superpixel belongs to the background. But in traffic videos, there are always vehicles crossing boundaries, at this time using *bdc* to detect moving objects will result in bad performance. So, compactness *comp* was utilized to improve the performance. *comp* indicates the disorderly distribution of superpixels, higher *comp* means the superpixel has higher likelihood of the background. *comp* can not be influenced by the locations of the objects, so it is complementary with *bdc*.

In Zheng's work, the author also used rareness for saliency detection. But in our research, we found rarely appeared colors always represent road signs, traffic lights, buildings beside the road, etc.. Therefore, we didn't use rareness for moving objects detection.

To calculate boundary connectivity *bdc* and compactness *comp*, image was cut into superpixels using SLIC method at first [7]. Then according to color similarity, superpixels were clustered into several object regions that represent both foreground and background. Next, using Equation (6) and (7) to calculate *bdc* and *comp* of each region, where *len(i)* means how many pixel are on the boundary of region *i* and area(i) means the total number of pixel in region *i*. In Equation (7), *K* represents the number of superpixel in region *i*, *pos(j)* is the mass point of superpixel *j* and *mpos(i)* is the mass point of region *i*.