Camera Based Moving Object Detection for Autonomous Driving

Introduction

Autonomous driving gives vehicle the ability to sense its environment and hence can navigate without human intervention. Automotive industry is changing with the aim of developing intelligent vehicles. The most significant changes in the future with autonomous driving include

- **Increased vehicle safety**: Systems such as obstacle detection, lane departure warning, lane keep assist, adaptive cruise control and so on help avoid accidents and thus increase the safety attribute of the vehicle

- **Steady traffic flow**: Traffic congestion caused by driver error can be avoided by various autonomous driving applications

- **Reduce fuel consumption**: Automatically adjusting speeds on highways by using the adaptive cruise control feature has a direct effect on increased fuel efficiency
To develop any autonomous driving applications, perception functions play the crucial role. The key challenge for developing such system is to manage and combine the significant amount of data coming from the different sensors and to create a consistent model from this data that can make decisions. In developing self-driving vehicle technology, SAE International, initially established as the Society of Automotive Engineers classified standards of driving automation levels based on the amount of driver intervention and attentiveness required, rather than the vehicle capabilities. In general the definitions include:

- **Partially automated**: The driver must continuously monitor the automatic functions and cannot perform any non-driving task

- **Highly automated**: The automatic system recognizes its own limitations and calls for driver to take control when needed. The driver can perform a certain degree of non-driving tasks

- **Fully automated**: The system can handle all the situations autonomously, there is no need for human intervention. Driverless driving is possible at this level

As the technology is approaching towards fully autonomous driving, the number of sensors in vehicles will increase drastically, but which sensors will provide most value is an important question for automaker. It is in his best interest to push hard for cost optimization without sacrificing safety. Some of the prominent sensors include:

- **LIDAR**: LIDAR is the abbreviation of Light detection and Ranging, it is the technology that uses laser light to measure distances up to 100 meters in all directions and also generating a precise 3D map of the vehicle’s surroundings. The problem with LIDAR is that they generate huge amounts of data and are expensive to implement. At the same time, goal of reaching fully autonomous driving needs the behavior of LIDAR to achieve significant results

- **RADAR**: RADAR is the abbreviation of Radio Detection and Ranging, is a sensor system that uses radio waves to determine the velocity, range and angle of objects. RADAR is computationally lighter and uses far less data than LIDAR. However, its output is less angularly accurate than LIDAR.

- **Camera**: Cameras are the master of classification and texture implementation. A camera sees in 2 dimensions, with much higher spatial resolution. It is also possible
to infer depth information from the camera images. By far they are the cheapest and most available sensors. Self driving functions can be developed by using cameras with smart algorithms.

Camera based moving object detection is the most important functionality for collision avoidance, lane departure warning, etc. Object detection can be defined as identifying objects of interest in the video sequence and to segment pixels of these objects. Various techniques for implementing this include frame differencing and background subtraction.

**Frame differencing**

The moving objects are determined by calculating the difference between two consecutive images. It is simple to implement and also computationally less expensive. The method best works for determining moving objects with static camera. Since the moving camera introduces ego-motion to the video, simple frame differencing would be corrupted by the background motion.

![Frame differencing with static camera](image)

**Figure 1: Frame differencing with static camera**

**Background subtraction**

Developing a reference model is the first step for background subtraction. The reference model developed should be sensitive enough to recognize moving objects. The process is to compare the reference model against each frame extracted from the video to determine possible variation. It is a simple algorithm but very sensitive to external environment and has poor anti-interference ability.
Moving camera based moving object detection

A simple frame differencing or background modeling does not provide the expected results when the video recorded with moving camera is used. Compared to fixed cameras, moving object detection in the video captured by a moving camera is difficult to analyze since camera motion and object motion are mixed. When frame differencing or background subtraction applied to such video, the shapes of moving objects fail to be effectively segmented and detected [Hu et al. (2015)]. The example of frame differencing in such video is shown in figure 2.

Computer vision provides more robust results when coming to moving object detection with moving camera. In the proposed method, the feature points in the frames are detected using Harris corner detector [Shi & Tomasi (1993)] and further the feature points are classified as belonging to foreground or background using epipolar constraint. The region of moving object is segmented by Delaunay triangulation followed by tracking. The process is depicted in the figure 3.

Epipolar constraint

Epipolar geometry is the intrinsic projective geometry between two views. It is independent of scene structure and only depends on the camera’s orientation [Hartley & Zisserman (2004)].
A 3D point \( P(X,Y,Z) \) projected in two images planes as \( u_a \) and \( u_d \). The geometric relation between \( u_a \) and \( u_d \) is determined by epipolar geometry. The image points \( u_a, u_d \) along with camera centers \( C_a \) and \( C_d \) with 3D point \( P \) are coplanar and is denoted by \( \pi \) and is depicted in the figure 4.

**Figure 4: Epipolar constraint**

The rays back projecting from \( u_a \) and \( u_d \) intersect at \( P \) and this significant property is used to deal with correspondence problem. If only \( u_a \) is known and \( u_d \) should be calculated, then the epipolar constraint is employed. The corresponding points are related to each other such that, the point from the first view \( u_a \) back projects a ray to \( P \) and this ray is imaged on the second image plane as a line \( l' \), called epipolar line. The search of \( u_d \) can be restricted to this line instead of searching the whole image plane.

The computation of the geometric relation is represented by the fundamental matrix \( F \). From figure 5, it can be stated that to each point \( u_a \) in one image, there exists a corresponding epipolar line \( l' \) in the other image and any point \( u_d \) corresponding to \( u_a \) must lie on the epipolar line \( l' \), \( u_a \mapsto l' \). This mapping from point to line is a projective mapping and is represented by \( F \) for uncalibrated cameras and is represented in the equation [1]. For any
point $u_a$ in the first image, the corresponding line is $l' = F \cdot u_a$

$$u_d^T \cdot F \cdot u_a = 0 \tag{1}$$

The same property can be exploited to detect the feature points present on the moving objects. Consider two images taken by a moving camera at time 1 and 2. $P$ is the position of the 3D point at time 1 and $P'$ is the position of 3D point at time 2. If $P$ is the point on static background and is projected as $u_a$ in the first view then its corresponding point $u_d'$ should lie on its corresponding epipolar line $l' = F \cdot u_a$ else it is projected away from the epipolar line as depicted in figure [6]. Here the orthogonal distance between the point and its corresponding epipolar line is used to distinguish between the foreground and background points.

$$d_{epi} = \frac{(u_d^T \cdot F \cdot u_a)^2}{(F^T \cdot u_a)^2_1 + (F^T \cdot u_a)^2_2 + (F^T \cdot u_a)^2_1} \tag{2}$$

$$u_d = \begin{cases} 
    Foregroundpoint, & \text{if } d_{epi} \geq \text{Threshold} \\
    Backgroundpoint, & \text{otherwise}
\end{cases} \tag{3}$$
Figure 6: Foreground detection by epipolar constraint

This is valid only when the fundamental matrix is estimated using mostly background points and is not a robust solution for autonomous driving scenario in which not only background but various motions involved. A scene is always a mixture of various motions example, the car, pedestrian, camera. Each motion can be represented by a unique F matrix. Points on the same moving object have same F. If a set of points lies on the same moving object in an image, their corresponding points on the other frame lie on the corresponding epipolar lines. However in motion segmentation both motions and fundamental matrices are unknown. To resolve this problem, an iterative approach called randomized voting is employed [Jung et al. (2014)].

First the feature points are randomly clustered and fundamental matrices are estimated for each cluster. Then the sampson distance $d_{epi}$ which is the distance from the point to its corresponding epipolar line is calculated for all the points using F calculated from all the clusters. A vote is assigned for whole points based on the distance value. Based on scores the points are reshuffled in the clusters until the clusters are converged. The detailed algorithm is presented in [Jung et al. (2014)].

Among the separated motion clusters, clusters on the moving object and cluster on the background should be distinguished. Usually the background feature points are scattered more widely than the feature points of moving object except when the background is an extremely smooth plane. Motion clusters can be classified as background or moving objects using scatteredness or dispersion of each cluster. In the proposed method standard deviation of all the points is used to classify background and foreground. Clusters with smaller standard deviation are considered to be present on the moving object. Delaunay triangulation (DT) is
constructed for each cluster to remove the outliers. After constructing the DT, the triangles with large area are rejected and is clearly depicted in the figure [9]. The detected regions are tracked in the subsequent frames using Kanade-lukas-tomasi tracking algorithm.
The functionality can be further developed for applications such as adaptive cruise control, lane departure warning.

I thank Prof. Dr.-Ing. Gangolf Hirtz and Dr.-Ing. Michel Findeisen for giving me the opportunity to work on my thesis at the digital signal processing department and helping me throughout my learning process.
References


URL: http://dx.doi.org/10.1109/CVPR.2014.158