Integration of Vehicle Detection and Distance Estimation using Stereo Vision for Real-Time AEB System

Byeonghak Lim, Taekang Woo and Hakil Kim
Department of Information and Communication Engineering, Inha University, Incheon, Korea

Keywords: Stixel, Vehicle Detection, Distance Estimation, Surface Normal Vector, Stereo Vision, Convolutional Neural Network, Autonomous Emergency Braking System.

Abstract: We propose an integrated system for vehicle detection and distance estimation for real-time autonomous emergency braking (AEB) systems using stereo vision. The two main modules, object detection and distance estimation, share a disparity extraction algorithm in order to satisfy real-time processing requirements. The object detection module consists of an object candidate region generator and a classifier. The object candidate region generator uses stixels extracted from image disparity. A surface normal vector is computed for validation of the candidate regions, which reduces false alarms in the object detection results. In order to classify the proposed stixel regions into foreground and background regions, we use a convolutional neural network (CNN)-based classifier. The distance to an object is estimated from the relationship between the image disparity and camera parameters. After distance estimation, a height constraint is applied with respect to the distance using geometric information. The detection accuracy and distance error rate of the proposed method are evaluated using the KITTI datasets, and the results demonstrate promising performance.

1 INTRODUCTION

Recently, driving safety has become more important and interest in intelligent vehicle systems has increased. For this reason, the technology for the Advanced Driving Assistant System (ADAS) and autonomous cars has undergone continual development. In particular, the demand for Autonomous Emergency Braking (AEB) systems is increasing as the Euro NCAP has mandated their installation from model year 2018 onwards.

Many studies on AEB systems have suggested technologies using various sensors to find obstacles and measure the distance to them. Radar and LiDAR-based approaches achieve good performance with high quality point clouds. However, these systems cost thousands of dollars and require periodic maintenance at least every three years. It is also difficult for these approaches to distinguish different types of obstacles. Therefore, studies in recent years have focused on image sensors using camera-based technologies. The advantage of a camera-based approach is that it can obtain detailed obstacle information so that the proper decision can be made with respect to the type of obstacle. Although development of monocular camera-based ADAS has been attempted, the distance error rate is too high to be commercialized. A stereo camera is more commonly used in order to take advantage of the detailed information available through the disparity between the two images.

It is difficult to process whole images in real-time because image processing is a pixel-wise computation. Thus, the processing time increases exponentially as the complexity of the algorithm and input image size increase. We propose a new integrated object detection and distance estimation method based on the stereo camera for real-time AEB systems. The main novelties of our proposed method are as follows:

- We introduce real-time integration for AEB systems by using algorithmically generated, low complexity regions of interest (ROIs).
- We apply a surface normal vector (SNV) validation process to minimize errors in the candidate regions caused by low quality stereo matching.

Lim, B., Woo, T. and Kim, H.
Integration of Vehicle Detection and Distance Estimation using Stereo Vision for Real-Time AEB System.
DOI: 10.5220/0006296702110216
In *Proceedings of the 3rd International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2017)*, pages 211-216
ISBN: 978-989-758-242-4
Copyright © 2017 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

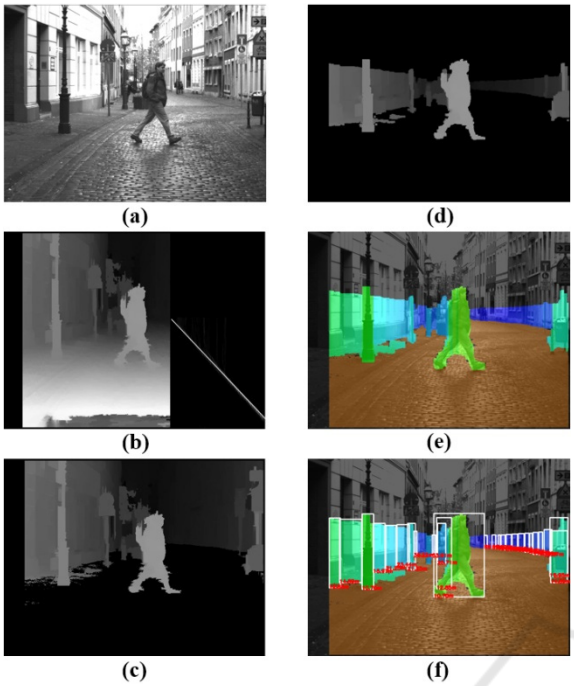

Figure 1: Hypothesis ROI generation flow chart (a) Input image (b) Disparity and V-disparity (c) Removing Ground (d) Height Constraint (e) Stixel Extraction (f) Stixel Segmentation.

- We use a convolutional neural network (CNN) with efficiently reduced input data for classification.

Many vehicle detection algorithms based on stereo vision have been developed in recent years. Vehicle detection using sparse density stereo matching with image features (Cabani et al., 2005) and geometry (Chang et al., 2005) was mainly used in the mid-2000s. Some studies used additional motion information like optical flow (Franke et al., 2005) to find differences between the object and the background. In the late-2000s, mid-level representation methods (Elfes, 1989, 2013; Badino et al. 2009; Qin, 2013) appeared and became widely used. The main approach in these methods was to design algorithms to cluster pixels into a hypothesis segments (Barrois et al., 2013; Barth et al., 2009; Broggi et al., 2010). To overcome the technical limits of processing only image pixel values, motion from object tracking has been used and yielded great progress since 2010 (Danescu et al., 2011; Erbs et al., 2011).

As the parallel processing capability of hardware has improved, CNN-based visual processing algorithms have showed outstanding performance in many visual recognition and detection challenges (Everingham et al., 2012; Russakovsky et al., 2015).

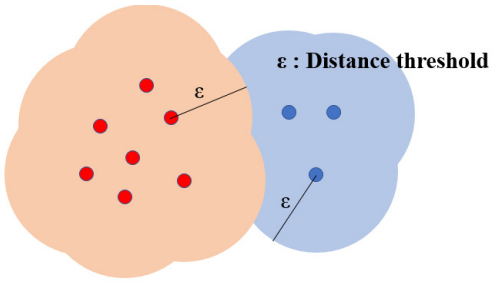

Figure 2: Stixel clustering.

For this reason, CNN detection algorithm development has exploded (Ren et al., 2015). A single shot object detection approach has also appeared, and revealed the possibility that CNN can be used in real-time systems (Redmon et al., 2015; Liu et al., 2015).

2 INTEGRATED MODULE FOR AUTONOMOUS EMERGENCY BRAKING SYSTEM

The detection and distance estimation modules share a disparity map extraction algorithm, which accounts for half of the entire processing time for the real-time integrated system. Each module uses this disparity map as needed. The stereo matching method used for generating the disparity map is a local matching with low computational cost. The details of the stereo matching algorithm will not be covered in this paper because the algorithm can be replaced as necessary.

The vehicle detection algorithm consists of two sequential parts. One is the hypothesis ROI generator based on stixels, and the other is a classifier based on CNN. In region-based CNN detection algorithms, blob detection is commonly used as a region proposal method (Van de Sande et al., 2013; Matas et al., 2004). However, this usually requires a lot of time because there are hundreds of different blobs in an image, and various scales of blobs should be considered. Therefore, we replaced this algorithm with our ROI generator, which only requires a few milliseconds when using a disparity map, to enable our integrated system to run in real-time.

2.1 Stixel Hypothesis ROI Generation

The main concept behind our hypothesis ROI generation is to use a low cost stereo matching algorithm, though the resulting disparity map quality

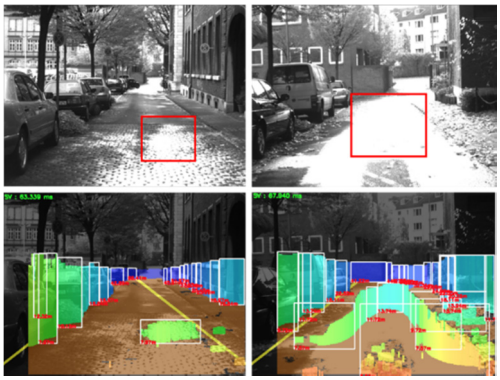

Figure 3: Example of hypothesis ROI generation flow.

is poor. In order to remedy the disparity map errors caused by low quality stereo matching, we use surface normal vectors (SNV) to validate the object candidate regions. Figure 1 shows a flow chart for the stixel ROI generating algorithm.

The stixel is one of the mid-level representation methods, which uses an occupancy grid and digital elevation map (Elfes, 1989, 2013; Qin, 2013). We borrow the basic concept of the stixel by using the distance and height information of objects.

Ground estimation is important for extracting a high quality stixel. Because the stixel is perpendicular to the ground, the ground information has great influence on the stixel estimation result. Here, we assume the ground is flat to simplify the situation. V-disparity is a graph of the frequency of disparity values along the v-axis. This is used to estimate the ground. The dominant line which represents the ground is estimated using RANSAC, which is an efficient model estimator. After finding the ground, we can remove it from the disparity image and constrain the height. A stixel is obtained by drawing a line from top to bottom such that the disparity pixel value is greater than 0. The stixels are then clustered to generate the object bounding box ROIs. We adopt the clustering method (Ester et al., 1996) to cluster stixels into ROIs. This method first select a seed which is the closest and the leftmost. After that, the stixels within the radius threshold are clustered into one object candidate as shown in Figure 2. We assumed that a car size could not over 2.5 meters.

2.2 Surface Normal Vector Validation

In section 2.1, we introduced the algorithm for ROI generation. The ROIs obtained through local matching are not always reliable because of the disparity error. For example, wet ground and a field

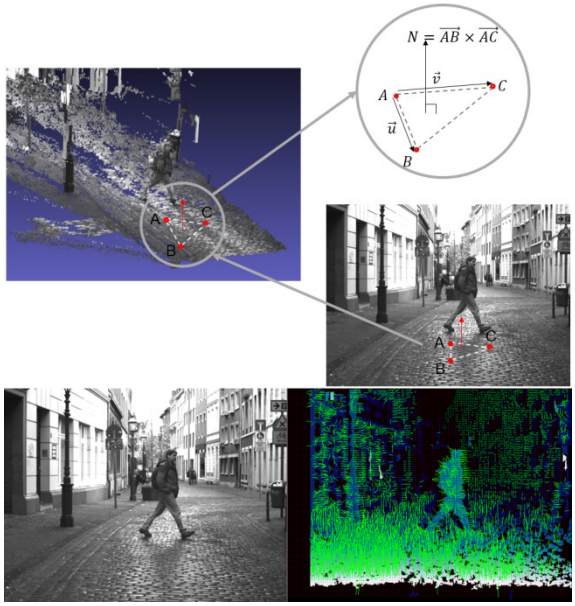

Figure 4: Example of surface normal vector.

that reflects light are incorrectly recognized as objects as shown in Figure 3. The SNV is applied to solve this problem. The SNV can be calculated using the formula below with 3 points, A, B and C, which lie on the plane of the image as shown in Figure 4.

$$N = \overrightarrow{AB} \times \overrightarrow{AC} \quad (1)$$

For this method, the selection of A, B and C affects performance. Thus, we heuristically select pixels with the distance interval, double of the disparity resolution, using the best result from the experiment to determine value change. We determine the dominant normal vector by using the adaptive mean shift. While the original mean shift searches for the mean value with a fixed kernel range, the adaptive mean shift uses a dynamic kernel range. Finally, errors are removed by examining the direction of the dominant normal vector in the ROI.

2.3 CNN Vehicle Classification

In order to separate the candidate regions into the foreground and background, we adopt the CNN as a classifier. Our base CNN network is constructed using only basic 3x3 convolutional filters (Simonyan et al., 2014), which are optimized on the NVIDIA embedded board using their SDK. All fully connected layers in the final section of the network are replaced with convolutional layers to reduce processing time (Lin et al., 2013).

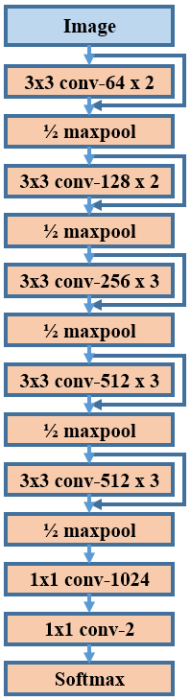

Figure 5: Classifier network model.

We save approximately one-third of the total processing time by removing the fully connected layers. For additional performance enhancements in the detection rate, we applied the residual connection introduced in (He et al., 2015). Most CNN detection algorithms use an entire image as an input to the convolutional layers for computational sharing. However, in real-world driving environment images, objects rarely take up a huge portion of the image. Therefore, we only use the object regions as inputs. The input image should be resized to 32x32 pixels to form the mini-batches for parallel processing. Figure 5 illustrates the network model used for our method.

2.4 Distance Estimation

The distance to objects is estimated from the disparity information in the ROI. It is determined by the following formula:

$$Z = \frac{fB}{d} \quad (2)$$

where Z is the distance, f is the focal length, B is the distance between the two lenses of the stereo camera, and d is the representative disparity. The representative disparity is selected as the maximum value in the disparity value histogram, which is the shortest distance from the object to the camera.

Figure 6: Result images.

We apply the height constraint again using the more precise distance to reject objects that are too tall to be considered.

3 EXPERIMENTAL RESULTS

The detection accuracy and distance error rate were evaluated using the KITTI dataset benchmark (Geiger et al., 2012). In order to use the distance information in the KITTI annotations, we divided the image set into 3712 images for the training sets and 3769 images of the validation sets. We followed the dividing policy used in the paper by (Chen et al, 2015), which considers the correlation between sequential images. For the CNN training data, we use the ground truth, the stixel candidate regions, and random cropped images. We decided that the threshold for positive sets would be an overlap ratio with the ground truth above 0.5, and the negative sets would be any with an overlap ratio below 0.5.

To prove that the performance of our module is reasonable, we conducted a comparative experiment on detection rate with SSD (Liu et al., 2015) which is the representative real-time CNN detection algorithm. Table 1 displays the average precision rate for 'Car' detection. The precision results for the stixel with SNV validation is 2.18% higher than that without SNV validation. In order to analyze the effect of our separate modules on the precision rate, we also evaluated detection precision on the ground truth regions. This result demonstrated the capability of the CNN classifier and we determined that the stixel clustering must be improved to achieve better detection results.

Table 1: Average Precision Results on Car Class.

Method	AP(%)
Stixel + CNN	60.76
Stixel(SNV) + CNN	62.94
Ground Truth + CNN	91.14
SSD	65.20

The distance error rate was evaluated on objects over 5 m away, and the accuracy was 92.51%. Because we limit the length of the epipolar line to search the corresponding points, distances less than 5 m are not reliable.

Table 2: Distance estimation error rate.

Constraint	Error rate(%)
> 5 m	7.49
> 0 m	8.33

Processing time was also evaluated across several platforms to validate the real-time system. We checked the average processing time over five iterations. The result on PC is 25 fps, and on the NVIDIA TX1 board is 11 fps, as shown in Table 3.

Table 3: Processing Time.

Platform	Processing Time (ms)		
PC (Titan X/i5 4670)	ROI generator	23	39
	Classifier	16	
NVIDIA TX1	ROI generator	51	90
	Classifier	39	

4 CONCLUSIONS

We introduced an integration of vehicle detection and a distance estimation algorithm for real-time AEB systems. Our main innovation is to share disparity map generation, which is the most time-consuming algorithm, for both object detection and distance estimation. To reduce the processing time, we use local matching, which is fast, but not very reliable. We alleviate this problem with SNV. The processing time satisfies real-time requirements on PC, and almost reaches real-time on an embedded board, the TX1. The detection performance is reasonable when compared to the results of other real-time detection modules.

Through the experimental results, we observed that the proposed classifier does not fully utilize its classification capabilities and determined that there is room for improvement in this aspect. Future work will include development of an improved stixel clustering method to enable the CNN classifier model to be fully utilized. Additionally, the CNN classifier model can be re-designed to achieve better performance. Our CNN model has very basic convolutional layers, which could be replaced with a state-of-the-art model (Szegedy et al., 2016). We assumed the ground is flat and is estimated from a straight line in the v-disparity. However, in the real world, the ground is not always flat. Therefore, the estimated line in the v-disparity should be curved to more accurately find the ground.

ACKNOWLEDGEMENTS

This work was supported by the Industrial Technology Innovation Program, “10052982, Development of multi-angle front camera system for intersection AEB,” funded by the Ministry of Trade, Industry, & Energy (MI, Korea).

REFERENCES

I. Cabani, G. Toulminet and A. Bensrhair. (2005). Color-based Detection of Vehicle Lights. *Proceedings of the IEEE Intelligent Vehicles Symposium*.

P. Chang, D. Hirvonen, T. Camus, B. Southall. (2005). Stereo-based Object Detection, Classification, and Quantitative Evaluation with Automotive Applications. *Workshops in the IEEE conference on Computer Vision and Pattern Recognition*.

U. Franke, C. Rabe, H. Badino, S. Gehrig. (2005). 6D-Vision: Fusion of Stereo and Motion for Robust Environment Perception. *Joint Pattern Recognition Symposium*.

A. Elfes. (1989). Using Occupancy Grids for Mobile Robot Perception and Navigation. *Computer*.

H. Badino, U. Franke, D. Pfeiffer. (2009). The Stixel World-A Compact Medium Level Representation of the 3D-World. *Joint Pattern Recognition Symposium*.

A. Elfes. (2013). Occupancy Grids: A Stochastic Spatial Representation for Active Robot Perception. *arXiv preprint arXiv:1304.1098*.

R. Qin, J. Gong, H. Li, X. Huang. (2013). A Coarse Elevation Map-based Registration Method for Super-Resolution of Three-Line Scanner Images. *Photogrammetric Engineering & Remote Sensing*.

B. Barrois, C. Wohler. (2013). 3D Pose Estimation of Vehicles using Stereo Camera. *Transportation Technologies for Sustainability*.

A. Barth, U. Franke. (2009). Estimating the Driving State of Oncoming Vehicles from a Moving Platform using Stereo Vision. *IEEE Transactions on Intelligent Transportation Systems*.

A. Broggi, A. Cappalunga, C. Caraffi, S. Cattani, S. Ghidoni, P. Grisleri, P. Porta, M. Posterli, P. Zani. (2010). TerraMax Vision at Urban Challenge. *IEEE Transactions on Intelligent Transportation Systems*.

R. Danescu, F. Oniga, S. Nedevschi. (2011). Modeling and Tracking the Driving Environment with a Particle-based Grid. *IEEE Transactions on Intelligent Transportation Systems*.

F. Erbs, A. Barth, U. Franke. (2011). Moving Vehicle Detection by Optimal Segmentation of the Dynamic Stixel World. *IEEE Intelligent Vehicles Symposium*.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman. (2012). The Pascal Visual Object Classes (VOC) Challenge. *International journal of Computer Vision.*

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei. (2015). ImageNet Large Scale Visual Recognition Challenge. *International journal of Computer Vision.*

S. Ren, K. He, R. Girshick, J. Sun. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv preprint arXiv:1506.01497*.

J. Redmon, S. Divvala, R. Girshick, A. Farhadi. (2015). You Only Look Once: Unified, Real-time Object Detection. *arXiv preprint arXiv:1506.02640*.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed. (2015). SSD: Single Shot Multibox Detector. *arXiv preprint arXiv:1512.02325*.

J. R. Uijlings, K. Van de Sande, T. Gevers, A. W. Smeulders. (2013). Selective Search for Object Recognition. *International journal of Computer Vision.*

J. Matas, O. Chum, M. Urban, T. Pajdla. (2004). Robust Wide-Baseline Stereo from Maximally Stable Extremal Regions. *Image and Vision Computing*.

M. Ester, H. Kriegel, J. Sander, X. Xu. (1996). A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the AAAI conference on Knowledge Discovery and Data Mining*.

K. Simonyan, A. Zisserman. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

M. Lin, Q. Chen, S. Yan. (2013). Network in Network. *arXiv preprint arXiv:1312.4400*.

K. He, X. Zhang, S. Ren, J. Sun. (2015). Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*.

A. Geiger, P. Lenz, R. Urtasun. (2012). Are We Ready for Autonomous Driving? The KITTI vision benchmark suite. *IEEE conference on Computer Vision and Pattern Recognition*.

X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, R. Urtasun (2015). 3D Object Proposals for Accurate Object Class Detection. *Advances in Neural Information Processing Systems*.

C. Szegedy, S. Ioffe, V. Vanhoucke (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv preprint arXiv:1602.07261*.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS