Vehicle detection and tracking with affine motion segmentation in stereo video

Vehicle Detection and Tracking with Affine Motion Segmentation in Stereo Video

Masayuki Miyama #1, Yoshio Matsuda #2

Graduate School of Natural Science and Technology, Kanazawa University
Kakumamachi, Kanazawa, Ishikawa 920-1192, Japan
1 miyama@t.kanazawa-u.ac.jp
2 matsuda@t.kanazawa-u.ac.jp

Abstract—**This paper proposes a novel method for vehicle detection and tracking based on stereo vision, motion analysis, and road detection. Combining stereo vision with motion analysis makes object detection more appropriate than each individual method. Object tracking with motion estimation, which follows the next object detection starting with an initial solution given by the tracking, improves detection accuracy. Road detection with motion segmentation, considering a parallax in stereo vision as a motion, enables on-road vehicle and obstacle detection. All elements constructing the proposed method are founded commonly on affine motion segmentation. Software and hardware implementations become efficient because their parts share the common principal procedure. Simulation results show the proposed method successfully detects and tracks vehicles on moving background in a complicated scene of downtown.**

I. INTRODUCTION

In recent years, driving assistance system to increase automotive safety and decrease driver's burden has become practical. Adaptive cruise control to follow a front car and collision mitigation brake to forcedly brake before collision are examples of the system. These technologies require detecting moving cars and obstacles in front of the vehicle. Since the old days, radars and ultrasonic sensors have been used for this purpose. These active sensor based system requires lower computational cost. However, the system has disadvantages such as lower resolution and interference with other active sensors. On the other hand, the passive sensor based system using optical sensors such as video cameras has advantages of higher resolution and non-invasion with other systems, although requiring higher computational cost.

Many researches to detect front cars with video cameras have been studied. A vehicle detection system composed of hypothesis generation (HG) step and hypothesis verification (HV) step was proposed in [1]. Fig.1 depicts the concept of the HG and HV steps. The HG step extracts vehicle candidates with lower computational costs, while it cannot avoid false detection as shown in Fig.1. Then, the HV step classifies objects into cars and non-cars; all those were detected as cars at the HG step. The HV step has an advantage of lower false detection, although the computational costs become higher. Combination of HG and HV enables improving computational efficiency, while maintaining detection precision.

The HV step adopts SVM (Support Vector Machine) and NN (Neural Network) as classification methods. Methods with detection rate more than 90% have already been proposed in the literature. For example, feature extraction using wavelet transform and classification with SVM were proposed in [2]. Feature extraction with HOG (Histograms of Oriented Gradients) and SVM classification were proposed in [3].

The HG step can be divided into three categories; those are a knowledge based method, a stereo vision method, and a motion based method. The knowledge based method detects vehicle candidates with a priori knowledge about a car such as symmetry or strong side edges of a car. However, the method produces many false detections and falls into lower detection rate in a complicated scene of downtown. In the situation, the system finds many edges of non-cars, which leading to higher false detection. Overtaking cars on the side lane and oncoming cars on the opposite lane look like non-symmetry, which resulting in lower detection rate.

The stereo vision method estimates a depth from cameras to the point in a 3D space with a parallax. The parallax is a distance between corresponding points in the left and right images. Then, the depth map is created. The depth map is a 2D image; each pixel in the image has a depth value. The method detects general objects by segmenting the depth map into regions so that the depth changes continually in a region. However, the method could possibly divide one object into some regions with the parallax differences. The method requires higher computational costs and accurate camera parameters through strict calibration to estimate depth correctly.

The motion based method detects moving objects by motion segmentation. The motion segmentation calculates a pixel motion (optical flow). Then, a region is created by gathering pixels whose motions agree with a motion model corresponding to the region. This method can also detect general objects on moving background in a complicated scene. However, the method cannot detect static objects and requires higher computational costs.

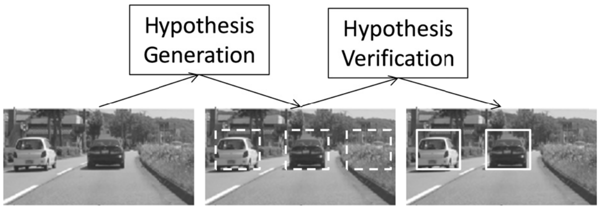


Fig. 1. Hypothesis generation step and hypothesis verification step

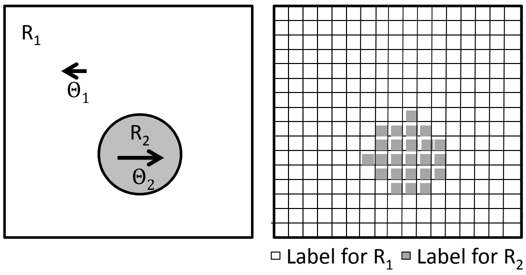
The stereo method can detect static objects; those cannot be detected by the motion based method. Although the stereo method might divide an object into some regions, they can be merged into the object by the motion based method. Thus this paper adopts a novel method for object detection based on both the stereo vision and motion analysis. In addition, object tracking with motion estimation, which follows the next object detection starting with an initial solution given by the tracking, improves detection accuracy. A combination of these methods can be used for general object detection and tracking.

Based on the general method, a novel method for vehicle detection and tracking with road detection is proposed in this paper. The road detection with motion segmentation, considering a parallax in stereo vision as a motion, enables on-road vehicle and obstacle detection. All elements constructing the proposed method are founded commonly on affine motion segmentation, which uses affine motion model to merge pixels into a region. Software and hardware implementations become efficient because their parts can share the common principal procedure. Especially, the hardware implementation can possibly reduce plenty of gates by sharing a circuit for the affine motion segmentation. The affine motion segmentation was difficult to execute in real-time so far, because of the high computational costs. However, a VGA 30 fps VLSI processor has been proposed in [4]. We can expect real-time processing of the proposed method with this processor.

This paper is organized as follows. The next section describes the affine motion segmentation. Then, the vehicle detection and tracking algorithm based on the motion segmentation is proposed in section 3. In section 4, simulation results show the effectiveness. Finally, we conclude this paper.

II. Affine Motion Segmentation

This section describes affine motion segmentation, which is a basis of the proposed method for vehicle detection and tracking. The algorithm was originally proposed in [5]. The aim of the motion segmentation is to extract moving regions in a video sequence. For example, the algorithm extracts the region R_1 moving to the left and R_2 moving to the right in a case of Fig.2. Each motion model of Θ_1 and Θ_2 expresses a motion of the corresponding region. The algorithm assigns a region label to each pixel according to the conformity to the motion model, resulting in a label map e_I for the frame I.



(a) Image I (b) Label map e_I

Fig. 2. Affine motion segmentation and label map

Fig.3 shows a flowchart of the affine motion segmentation. The algorithm adopts the affine motion model with a global illumination change ξ as a motion model. The model can express a motion of the region such as rotation, zoom in/out, and transformation. All pixel flow in a region can be expressed by a set of linear equations. The model is composed of seven parameters as below:

$$\Theta_l = [a_1^l\ a_2^l\ a_3^l\ a_4^l\ a_5^l\ a_6^l\ \xi]^{\mathrm{T}}, \tag{1}$$

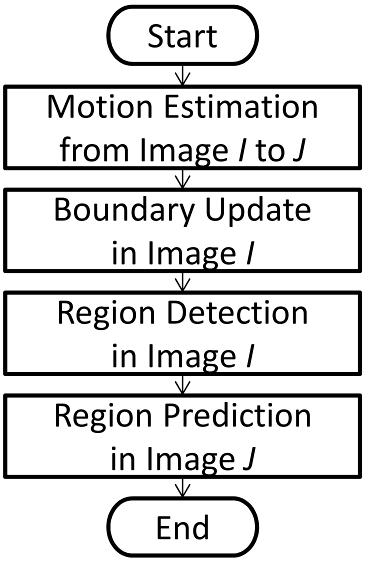
$$\boldsymbol{d}_{\Theta_l}(s) = \begin{bmatrix} u_l(x,y) \\ v_l(x,y) \end{bmatrix} = \begin{bmatrix} a_1^l + a_2^l x + a_3^l y \\ a_4^l + a_5^l x + a_6^l y \end{bmatrix}, \tag{2}$$

where s is a coordinate of (x, y), $\boldsymbol{d}_{\Theta_l}(s)$ is a motion vector at the coordinate s. The algorithm estimates an affine motion model Θ_l corresponding to the region R_l. A symbol $\{\Theta_l\}_I^J$ means a set of motion models from the frame I to J. The algorithm estimates Θ_l with accumulation of $\widehat{\Delta\Theta_l}$ obtained by the iterative re-weighted least square method minimizing sum of a residual $r(s, \Theta_l)$ for each pixel in R_l as the following:

$$\Delta\widehat{\Theta}_l = \underset{\Delta\Theta_l}{\arg\min} \sum_{s\in \mathrm{R}_l} \frac{1}{2} w_s \left\{r\left(s, \widehat{\Theta}_l\right)\right\}^2, \tag{3}$$

$$\begin{aligned} r\left(s, \widehat{\Theta}_l\right) = {} & \nabla J\left(s + \boldsymbol{d}_{\widehat{\Theta}_l}(s)\right) \boldsymbol{d}_{\Delta\widehat{\Theta}_l}(s) + \Delta\xi_l \\ & + J\left(s + \boldsymbol{d}_{\widehat{\Theta}_l}(s)\right) - I(s) + \xi_l, \end{aligned} \tag{4}$$

where $I(s)$ represents a illumination value at s in I. The illumination gradient $\nabla I(s)$ is defined as $\nabla I(s) = \begin{bmatrix} I_x(s) \\ I_y(s) \end{bmatrix}^{\mathrm{T}}$, using illumination differentials $I_x(s)$ and $I_y(s)$ in the x and y direction. The residual $r(s, \Theta_l)$ is derived from a linear approximation of the illumination conservation law expressed by: $I(s) = J\left(s + \boldsymbol{d}_{\Theta_l}(s)\right) + \xi_l$. The weight w_s is a weight assigned to a pixel at s. To set up small weight for each outlier pixel makes the estimation more robust. Introduction of hierarchical method enables large motion estimation.



Start
Motion Estimation from Image I to J
Boundary Update in Image I
Region Detection in Image I
Region Prediction in Image J
End

Fig. 3. Affine motion segmentation flowchart

The boundary update step obtains a label map e_I by updating region boundary in a prediction map $\tilde{e}_I$ using a set of motion models $\{\Theta_l\}_I^J$. This is defined as an energy minimization problem based on the MRF (Markov Random Field) graphical probability model, a label is assigned to each pixel so that the energy function is minimized. The energy function U is composed of three terms as follows: $U(s, \Theta_l, I, J, e_I, \tilde{e}_I) = U_1(s, \Theta_l, I, J) + U_2(s, e_I) + U_3(s, e_I, \tilde{e}_I)$. The label l_s to minimize this function is assigned to the pixel at s according to: $l_s = \arg\min_{l} U(s, \Theta_l, I, J, e_I, \tilde{e}_I)$.

The term U_1 represents the pixel conformity to the motion model. This term changes according to the error based on the illumination difference between the pixel at s in I and the pixel in J after moving with the model Θ_l corresponding to the candidate label l. The term U_2 expresses the energy of label continuity in a map. This term decreases as many neighbour labels agree with the candidate label at s in the label map e_I. The term U_3 decreases as the candidate label is equal to the label at the same position of the prediction map $\tilde{e}_I$. U_2 and U_3 are regularization terms to make a solution stable.

The region detection step detects newly appearing regions. The sub region in R_l, whose motion does not agree with the motion model Θ_l, becomes a new region. Therefore a region is segmented into a static region and a mobile region; the static region conforms to the model and the mobile region does not. The energy function similar to the function U is defined to label a pixel with static or mobile. The mobile region becomes a new region after labelling with this function.

The region prediction step predicts a label map e_J using a label map e_I and a set of motion models $\{\Theta_l\}_I^J$. The step moves labels in the region R_l according to the motion model Θ_l over all label l, yielding a prediction map $\tilde{e}_J$. This increases segmentation accuracy of the next image using the $\tilde{e}_J$ as the initial map.

III. Vehicle Detection and Tracking

This section describes a parallax model of a plane in stereo vision, and then the proposed algorithm for vehicle detection and tracking with the affine motion segmentation.

A. Affine Parallax Model

Fig.4 depicts a principal of depth measurement in stereo vision. When a point P belonging to a plane in a 3D space is taken by two cameras whose optical axes are set up in parallel, the depth d from cameras to the point P is expressed by:

$$d = \frac{rf}{p} = ax + b, \tag{5}$$

where r represents a distance between two cameras, f means a focal length of the camera, p expresses a parallax between the corresponding points in the left and the right image. The depth d is inverse proportional to the parallax p. The depth d of an arbitrary point in the plane can be expressed by the linear equation, where x is the coordinate shown in the figure; both a and b are constants. Then the parallax p can be expressed using the image coordinate y as below:

$$p = -\frac{ar}{b}y + \frac{rf}{b}. \tag{6}$$

This equation means each parallax within a plane in stereo vision can be expressed with the affine motion model. We name this model an affine parallax model.

Let us consider applying the affine parallax model to the vehicle detection. The affine motion segmentation usually divides the vehicle surface into one or a small number of regions. The segmented region corresponds to a plane in the vehicle surface such as a rear surface. Unevenness on the surface does not produce many regions in the segmentation. In a stereo image taken by an on-vehicle stereo camera, the segmentation ignores unevenness on the surface because the depth is usually dominant over them. In addition, the affine model can approximate each parallax on a flat road surface as shown in Fig.5. The model is applicable to both vehicle and road detection.

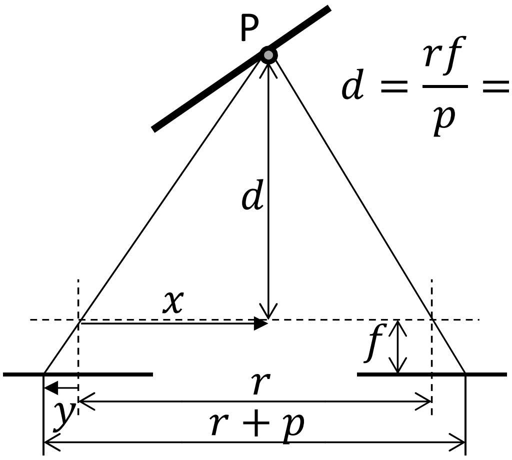


Fig. 4. Stereo vision and plane parallax

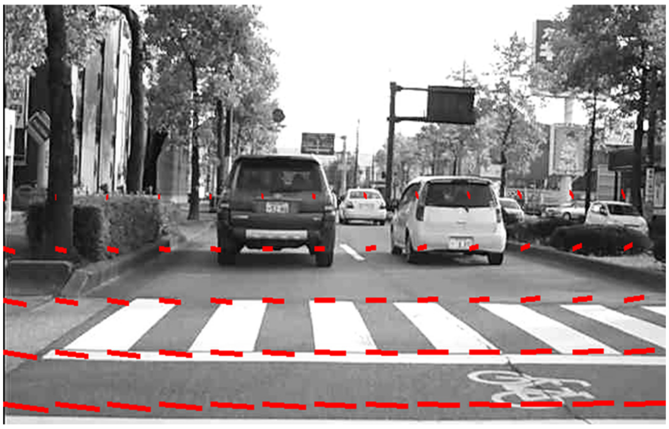


Fig. 5. Affine parallax model on flat road

B. Proposed Method

Fig.6 shows the proposed flowchart of vehicle detection and tracking. In time *t,* the algorithm executes the affine motion segmentation with left and right images, the road detection, and the affine motion segmentation with time *t* and *t*+1 images. Fig.7 shows a conceptual example composed of input images and processing results of every step in the flowchart. Vehicles on the road are detected by the proposed algorithm based on stereo vision, motion analysis, and road detection.

Firstly, object regions are segmented with the parallax in stereo vision at time *t*. An object might be divided into a few planes by the segmentation. Fig.7 (a) is a left image, (b) is a right image, and (c) is a segmentation result. Non-vehicles are also detected in this step. The near right car is segmented into a side surface and a rear surface.

Secondary, a road surface is determined using the parallax model. This step labels every pixel with road or non-road using the similar energy function in the region detection of the affine motion segmentation. Fig.7 (d) is an ideal road detection result, and (e) shows this step distinguishes objects on the road from the others.

Thirdly, the region boundary, which was segmented with the parallax at the first step on the same time, is updated with motion information from time *t* and *t*+1. A small number of planes constructing an object surface are merged with the similarity of their motions. Far moving objects with almost no parallax can be detected at this step. Fig.7 (f) is a result of this step. Two regions constructing a surface of the near right car are merged.

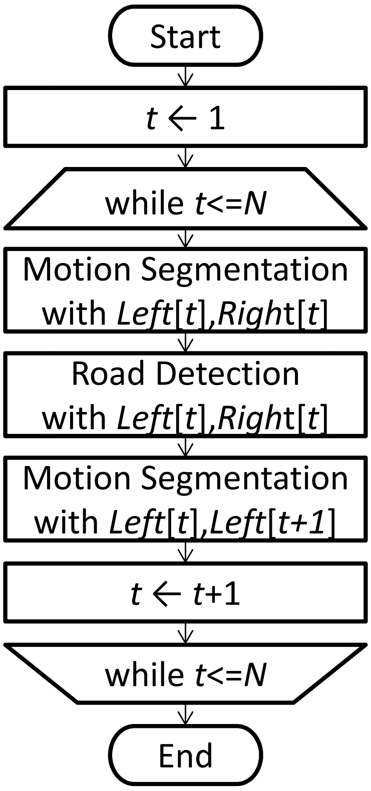

Start
t ← 1
while t<=N
Motion Segmentation with Left[t],Right[t]
Road Detection with Left[t],Right[t]
Motion Segmentation with Left[t],Left[t+1]
t ← t+1
while t<=N
End

Fig. 6. Flowchart for vehicle detection and tracking

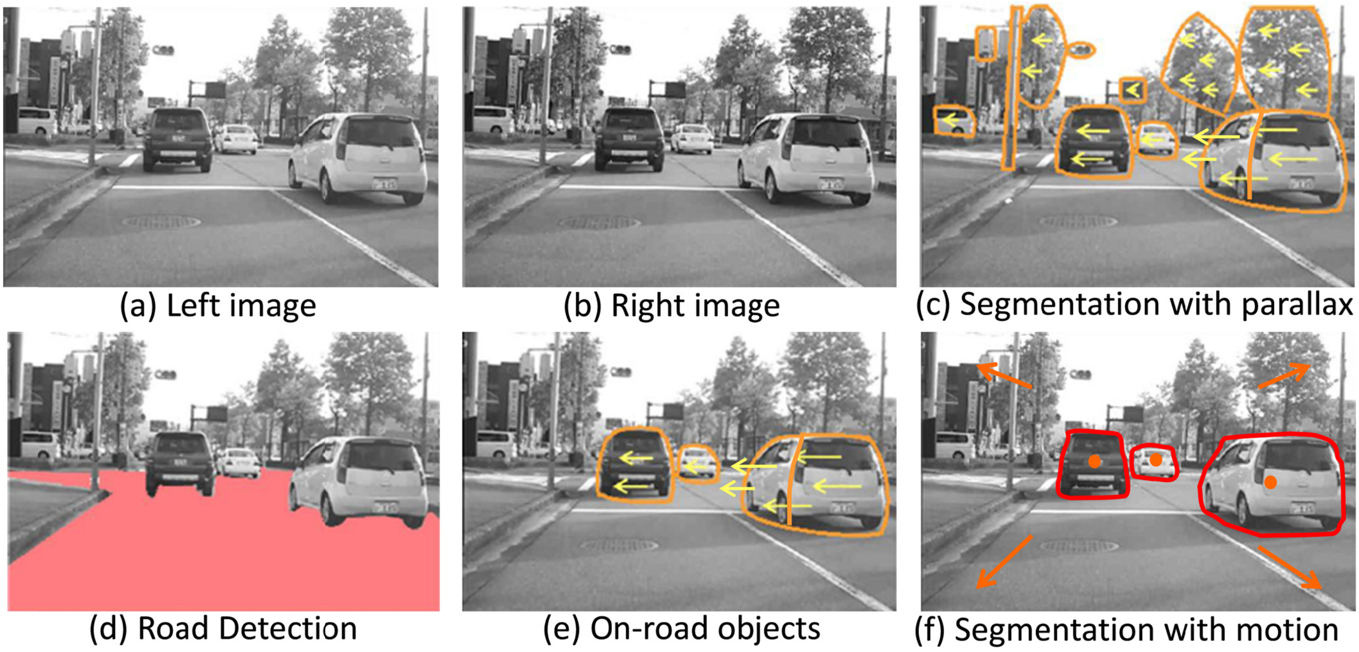


Fig. 7. Example of input images and ideal results for each step

The region prediction in the third step predicts the object regions at time t+1. This is equivalent to the object tracking. The proposed method does not need anything else for the object tracking. The object tracking yields the initial label map at time t+1.The next object detection starts with this label map, leading to improvement of segmentation accuracy.

The proposed method features as follows. Combining stereo vision with motion analysis makes object detection more appropriate than each individual method. Object tracking with motion estimation, which follows the next object detection starting with an initial solution given by the tracking, improves detection accuracy. Road detection with motion segmentation, considering a parallax in stereo vision as a motion, enables on-road vehicle and obstacle detection. All elements constructing the proposed method are founded commonly on the affine motion segmentation. Software and hardware implementations become efficient because their parts share the common principal procedure.

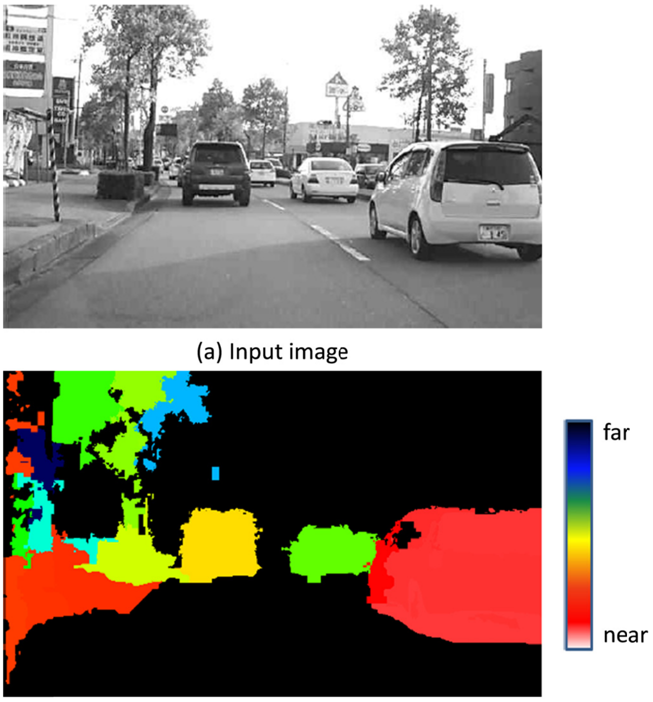
IV. Simulation Results

A. *Simulation Condition*

We took a stereo video which recorded the front of a car running in Kanazawa city at the speed of 20km/h~40km/h with a stereo camera made by ourselves. We made the camera with two video cameras (Exemode DV505, VGA 30 fps, 3.75mm focal length) mounted on a camera platform (SLIK plate II) 34.3cm apart. We mounted two cameras on the platform so that their optical axes became in parallel as possible. In the video, cars close together on the complicated background of buildings and trees. We made a simulation program in C language according to the proposed algorithm and ran it on a PC. In the simulation, we did not either use camera parameters usually estimated through strict calibration, or made any compensation to the video.

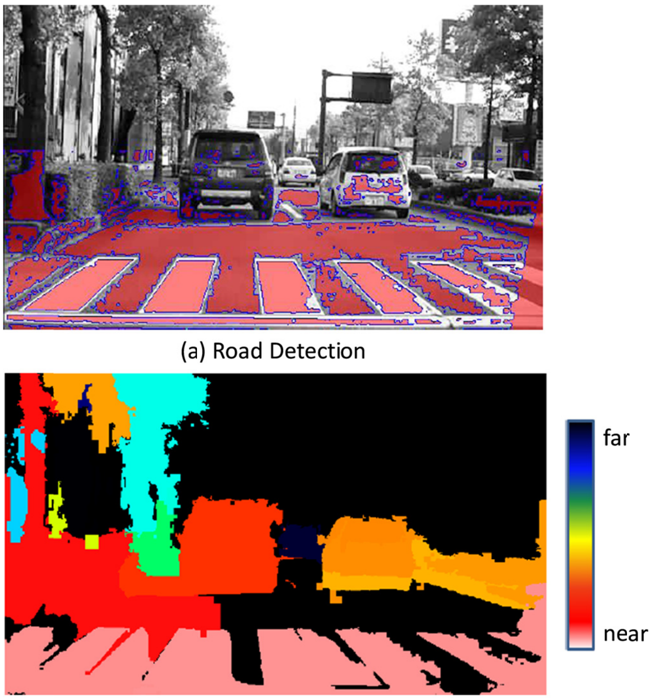
B. *Results*

Fig.8 (a) shows an input image and (b) shows the corresponding depth map. The depth map was made from an estimated label map and a parallax model for each region. The proposed method detected vehicles and trees with relative depths correctly. Fig.9 (a) shows a result of road detection and (b) shows the corresponding depth map. The proposed method detected a pedestrian crossing in the depth map. The road detection also detected the crosswalk as a road surface. These indicate that the crosswalk can be taken away in later processing. In addition, the proposed algorithm could keep tracking three front cars appeared in both scenes. The farthest car colored in deep blue can be found in Fig.9 (b). The two scenes are 10 seconds apart in time.


(a) Input image

(b) Depth map

Fig. 8. Simulation results


(a) Road Detection

(b) Depth map

Fig. 9. Road detection and depth map

C. Discussion

In Fig.9 (a), there are pixels labelled with road on vehicles. They can be removed by appropriate thresholding of label density in the region. The pixels on the crosswalk boundary are not labelled with road. They will be improved by applying an individual parallax model to each divided area or adjusting the energy function parameters. Another promising approach is to investigate the region conformity to a parallax model, which assumes the object to be standing vertically on the road. In the case of the object included in the road surface such as the crosswalk, the total energy within the region applying the vertically standing model increases in comparison with that of the road surface model.

V. Conclusion

This paper proposed a novel method for vehicle detection and tracking based on stereo vision, motion analysis, and road detection. All elements constructing the proposed method are founded commonly on the affine motion segmentation. We also proposed the affine parallax model for the segmentation in stereo vision. In the simulation results, the proposed method could detect and track moving vehicles on the road with relative depths in a complicated scene of downtown. The future works are quantitative investigation of segmentation and depth accuracy, investigation of camera calibration influence to the detection accuracy, and robustness evaluation of the proposed method. We plan to implement the real-time system with our VLSI processor for the affine motion segmentation. Application of object detection and tracking with the affine motion segmentation to the gesture recognition is another challenge.

Acknowledgment

This research was supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 22560325, 2010-2011.

References

[1] Zehang Sun, George Bebis, Ronald Miller, “On-Road Vehicle Detection: A Review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, No.5, May 2006.

[2] Z. Sun, G. Bebis, and R. Miller, "Quantized Wavelet Features and Support Vector Machines for On-Road Vehicle Detection," *Proc. IEEE Int’l Conf. Control, Automation*, Robotics, and Vision, Dec. 2002.

[3] F. Han, Y. Shan, R. Cekander, “A Two-Stage Approach to People and Vehicle Detection with HOG-Based SVM,” *PerMIS*, pp. 133-140, 2006.

[4] Masayuki Miyama, Yoshiki Yunbe, Kouji Togo, Yoshio Matsuda, “A VLSI Architecture for VGA 30 fps Video Segmentation with Affine Motion Model Estimation,” *ISIC-2009 - 12th International Symposium on Integrated Circuits*, pp. 449-452, 2009.

[5] Jean-Marc Odobez, Patrick Bouthemy, “Direct incremental model-based image motion segmentation for video analysis,” *Signal Processing*, Vol.66, pp.143-155, 1998.