社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－CVIM－161 (24)
2008／1／18

道路環境におけるSfMを用いた3Dテクスチャセグメンテーション

姜 有宣　山口 晃一郎　内藤 貴志　二宮 芳樹
(株) 豊田中央研究所

あらまし： 本稿では，道路環境理解のための技術として，車載カメラによるStructure from Motion(SfM)を利用したテクスチャセグメンテーション手法を提案する．提案手法では，まず時系列画像でSfMを適用し，シーンの3次元構造の推定と移動物体の検出を行う．そして、高次局所自己相関関数からテクスチャの特徴ベクトルを抽出する．その際，SfMから得られた距離情報によって異なるマスクパターンを利用し，テクスチャセグメンテーションを行う．SfMとテクスチャセグメンテーションを統合することにより，各テクスチャパターンを識別するとともにその3次元位置を推定することができる．実画像を用いた実験で，提案手法によりテクスチャの識別と3次元位置の推定を行うセグメンテーションが可能であることを示す．

3D Texture Segmentation for Road Environment Scenes using a SfM Module

Yousun Kang, Koichiro Yamaguchi, Takashi Naito and Yoshiki Ninomiya
Toyota Central R&D Labs., Inc.

Abstract： This paper presents a new texture segmentation method which can represent three dimensional structure for road environment scene using a *Structure from Motion* (SfM) module. The SfM module can reconstruct the three dimensional structure of the road scene and detect moving objects using estimation of ego-motion of the vehicle. According to the depth information which is obtained by SfM module, the texture features can be extracted from higher order local autocorrelation with different size of a mask pattern for the texture segmentation. By integrating the results of SfM module and texture segmentation, each texture pattern can be classified and localized in 3D road scene. Experimental results show that the proposed method can not only effectively classify the texture patterns of structures in 2D road scene but also represent segmented texture patterns as three dimensional structures, which is called 3D Texture Segmentation. The proposed system can expand into a dynamic 3D scene analysis system for vehicle environment perception in the future.
Keywords - Structure from Motion, higher order local autocorrelation, texture segmentation

1 Introduction

The last few years witnessed the development of driving assistance systems and automation technologies in Intelligent Transportation Systems (ITS). Among the many ITS technologies, the researches about vehicle safety have become in great interest and related many applications are expected to be used regularly in near future. The Vehicle Safety Systems (VSS) have investigated to provide the drivers with some information concerning its environment and any potential hazard perception [1]. The tasks related to road environment perception have based on detection and localization of traffic participants in front of a vehicle using vision sensors and pattern recognition techniques.

Typically, stereo vision or optical flow process are needed that cope with 3D localization and scene geometry estimation. As pattern recognition techniques for object detection, shape-based and/or texture-based objects have been recognized for road scene understanding. In particular, the shape-based objects recognition has been the main focus for the prevention of vehicular accidents. Because, the obstacles of a vehicle are recognized by specific patterns such as pedestrians, cars, and bicycles. However, a more complex scene understanding system, recently introduced in [2], relies upon the recognition of not only shape-based objects but also of texture-based objects. The texture-based objects, better described by their texture features rather than the local shape, include buildings, roads, trees, and sky.

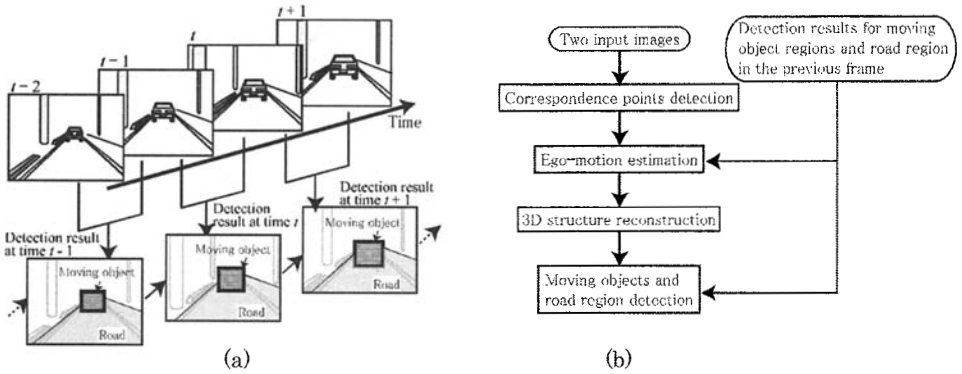


Figure 1: Overview of proposed method (a) Flow of proposed method (b) Process flow for each frame

They are also important objects to be recognized as a background from a moving vehicle. This paper is focused on texture-based object recognition using new texture segmentation scheme.

Recently, the vision-based intelligent vehicle systems tend to integrate systems of 2D recognition and 3D localization for dynamic scene analysis. Leibe [3] presented a real time system which can detect cars, bicyclists, and pedestrians in 2D image and localized them in 3D using a Structure-from-Motion (SfM) module and calibrated stereo cameras. In addition, Cornelis [4] presented a new framework for 3D urban scene modeling at video frame rate using realistic texture generation. These integration systems showed the better performance in 2D objects recognition and they can reconstruct structures of 2D scene into 3D models.

In this paper, we present a new framework, which is integrated to 2D texture segmentation and 3D localization using SfM module for road environment perception. The SfM module can estimate vehicle egomotion and represent three dimensional structures in road scene. We can obtain depth map of input video frame and the depth map is applied to our texture segmentation algorithm.

Main idea of the proposed algorithm is that feature vectors of a texture pattern can be extracted in consideration of its depth information. The texture features are extracted from higher order local autocorrelation functions. We can change a resolution of texture patterns using variable window size of a mask pattern according to depth. Therefore, the proposed method can effectively recognize texture patterns, in addition, classified texture patterns can be represent in three dimensional structures using the obtained depth map.

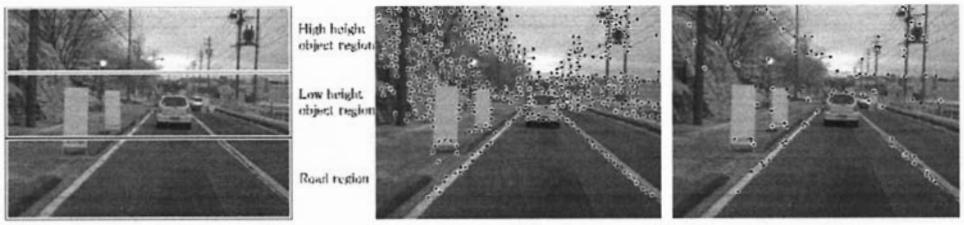
This paper is organized as follows. In Section 2, we explain the method of ego-motion estimation for SfM module. Section 3 describes texture feature extraction method according to depth map for texture segmentation. We show experimental results in Section 4. Finally, we summarize the present work in Section 5.

2 Scene Geometry Estimation

Using the SfM module, we can estimate the ego-motion of the vehicle and detect moving objects on roads by using a vehicle mounted monocular camera. There are two problems in ego-motion estimation. Firstly, a typical road scene contains moving objects such as other vehicles. Secondly, roads display fewer feature points compared to the number associated with background structures. In our approach, ego-motion is estimated from the correspondences of feature points extracted from various regions other than those in which objects are moving. After estimating the ego-motion, the three dimensional structure of the scene is reconstructed and any moving objects are detected.

2.1 The proposed method

The process flow of our SfM module is shown in Figurel. In this method, it is assumed that the camera is calibrated, i.e. the internal parameters of the camera are known. In each frame, our proposed method detects the moving objects and the road region in the current image. As shown in Fig. 1(a), two consecutive images are used at any one time, i.e. the image taken at time $t-1$ and that taken at time t are used at time t. Detection results for moving objects and for



(a) (b) (c)

Figure 2: Feature point selection (a) Image divided into three regions (b) All feature points (c) Selected feature points

the road region at time $t-1$ are also used for estimating the ego-motion and for detecting the road region at time t. In the initial frame, it is assumed that there is no moving obstacle in the previous time and that the road region in the previous time is decided according to the height of the camera that is measured when the vehicle is stationery.

Fig. 1(b) shows the process flow for each frame. First, feature points are extracted with the Harris corner detector [5], and the correspondences of the feature points between the two input images are detected by the Lucas-Kanade method [6]. Next, the ego-motion of the vehicle is estimated from the correspondences of the feature points.

For accurate ego-motion estimation, feature points are dispersedly selected from various regions, except those containing moving objects, by utilizing the detection results from the previous frame. Then, the three dimensional structure of the scene is reconstructed. Finally, the moving objects and the road region are detected. These detection results are utilized recursively for ego-motion estimation and road region detection in the next frame.

2.2 Estimation of vehicle ego-motion

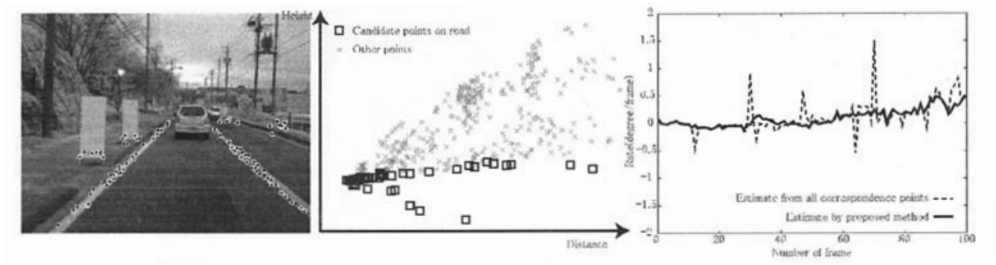
This subsection describes SfM algorithm for estimating the ego-motion of the vehicle. The feature points for ego-motion estimation are selected from the set of feature points. The ego-motion is then estimated from the correspondences of the selected points.

We utilize the moving object detection results from the previous frame to remove feature points on moving objects. Then, for a wide distribution of feature points, each image is divided into three regions; a region which may contain the road, one which may contain low-height objects, and one which may contain high-height objects, as shown in Fig. 2(a). The region that may contain the road is defined at the bottom of an image according to road region detection results from the previous frame. The low-object and high-object regions are then constructed by dividing the remaining region equally into two separate regions.

Feature points are selected from each region, and the number of feature points to be selected from each region is set beforehand. Fig. 2(b) shows a set of feature points extracted from an image. As shown in Fig. 2(b), in the case where the feature points are extracted from the whole image, background structures contribute many feature points, while the road region has a smaller number of feature points. Moreover, some of the extracted feature points are on the vehicle, which is actually a moving object.

On the other hand, feature points are distributed more uniformly throughout the image and some feature points on the vehicle are removed in the image as shown in Fig. 2(c). Although feature points on a moving object cannot be removed in cases where the moving objects are not correctly detected in the previous frame, the contribution of feature points on a moving object is suppressed by selecting points from three separate regions. Therefore, the ego-motion of the vehicle can be estimated accurately and robustly in a road scene by using this selection method.

The essential matrix can now be estimated from the correspondences of the selected feature points using the 8 point algorithm [7] and RANSAC [8]. The motion parameters are calculated from the estimated essential matrix. The motion parameters consist of 3 rotational and 3 translational parameters. The translational parameters are estimated up to scale.



(a) (b) (c)

Figure 3: Candidate points on road and their 3D positions (a) Candidate points on road (b) Side view (c) Estimated yaw rate

2.3 Detection of moving object and road region

After estimating the ego-motion of the vehicle, the positions of feature points in three-dimensional space are calculated by triangulation. Outlier points that are away from their epipolar lines or have negative distance are detected. The set of outlier points consists of feature points on moving objects and false correspondence points. To extract only those points that are on moving objects, the feature points are continuously tracked over consecutive frames. Feature points that are continuously classified as outliers are added to the set of candidate points for moving objects.

Candidates for points on moving objects are grouped according to their position in the image, the direction and the magnitude of their optical flow. Then, a moving object region is defined as a rectangle that includes all points in the same group.

Firstly, the plane of the road is estimated in three-dimensional space from points that are contained in the region detected as being the road in the previous frame, as shown in Fig. 3(a). In this estimation, the LMedS (Least Median of Squares) estimator is used, because some of the points identified as being in the road region in the previous frame may not actually be on the road, and some may have false positions in space due to false correspondences, as shown in Fig. 3(b).

After this, any scale ambiguity in the three dimensional structure can be removed from the position of the estimated road plane and the actual camera height. Then the input image is divided into small patches. Each patch is evaluated to determine whether or not it is a road region from the estimated road plane and the estimated ego-motion. Moreover, the distance of a moving object is estimated from the position of its lower edge by assuming that any moving objects are on the road.

Fig. 3(c) shows the estimated yaw rate between consecutive frames. This sequence contains moving vehicles. Although it is clear that false estimations of yaw rate often occur in the method when using all correspondence points, the proposed method can estimate the yaw rate in a stable manner.

3 Texture Segmentation

Texture gradient is one of the monocular depth cues in natural 2D scene. As the surface from a 2D scene with perspective line gets farther away from us, the texture appears finer. When feature vectors are extracted from a texture pattern, the texture pattern has an adequate resolution according to its depth. However, many researches have been performed regardless of the relation of relative resolution and real depth of a texture pattern in natural scene. Only uniform window size is available to make filter-banks for segmentation [9] and object recognition [10] in image database. Conversely, we propose a new method for extraction of texture features dependence on its depth information. Higher-order Local Auto Correlation (HLAC) functions are employed for feature extraction module in 2D road scene.

3.1 Feature extraction from HLAC

The autocorrelation functions possess uniqueness property for even orders, and they have advantages of being shift-invariant and computational low cost. The HLAC functions have been used in a wide range of applications such as face recognition [11] and texture classification [12]. A local autocorrelation function can be used to assess the amount of regularity as well as the

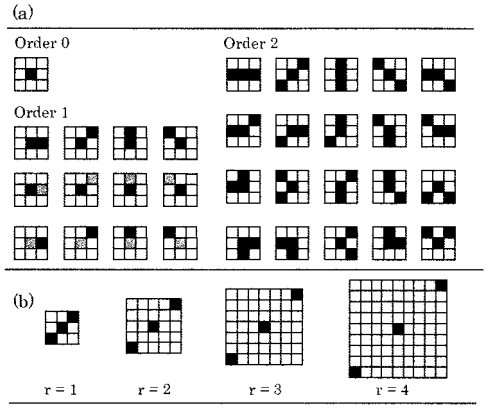


Figure 4: Mask patterns for second order local autocorrelation functions (a) Local mask patterns (b) Variable window size of a mask pattern obtained through dilatation of HLAC features

fineness/coarseness of a texture image. An important property of many textures is the repetitive nature of the placement of texture elements in the image. Due to advantages of low cost and repetitive nature of texture image, a HLAC function is employed in feature extraction module.

The HLAC functions are defined by

$$r_x^n(a_1, a_2, \cdots, a_n) = \int_D f(x)f(x+a_1)\cdots f(x+a_n)dx, \tag{1}$$

where n denotes the order of the autocorrelation function, x is the image coordinate vector, and a_i are the displacement vectors. A function $f(x)$ stands for the image intensity on the retinal plane D. Considering computational cost, we limit the order n to 2.

The feature extraction module computes 33 local autocorrelation coefficients from a texture pattern, using the mask patterns as shown in Fig. 4(a). For each mask pattern, a product is calculated by multiplying pixels in the masks together according to their patterns, whereas a gray pixel is multiplied twice. Because of the different degrees, feature vectors are normalized by taking the power root of the same degree of the product.

In order to extract feature vectors from a texture pattern, which has an adequate resolution dependence on depth, we use the difference window size of mask patterns. For example, mask size of 9×9 is utilized in the closet object, and 3×3 is utilized in the farthest object. Fig. 4(b) shows the ascending window size of mask patterns obtained through dilatation of HLAC features. In Fig. 4(b), r is a distance between neighborhood and center pixel and it can be changed by depth information of a texture pattern.

3.2 Segmentation by depth map

3D coordinates of feature points extracted from an input video frame, the 2D coordinates of moving objects and road region in 2D road scene can be obtained by our SfM module. However, as the feature points have sparse 3D coordinates, all pixels in an input image should be assigned to the interpolated depth value, producing the depth map. The process is illustrated in Fig. 5, and proceeds as follows.

Firstly, an input image is clustered by Mean-shift clustering algorithm [13]. The Mean-shift algorithm does not require prior knowledge of the number of clusters, and does not constrain the shape of the clusters. An image is typically represented as a two-dimensional lattice of p-dimensional vectors (pixels), where $p = 1$ in the gray-level case, three for color images. The space of the lattice is known as the spatial domain, while the gray level or color information is represented in the range domain.

For both domains, Euclidean-distance is assumed. Since Euclidean-distance in RGB color space does not correlate well to perceived change in color, we converted the color vectors in RGB color space to L*u*v* color space for clustering process. When the location and range vectors are concatenated in the joint spatial-range domain of dimension $d = p + 2$, their different nature has to be compensated by proper normalization.

The multivariate kernel is defined as the product of two radially symmetric kernels and the Euclidean-distance allows a single bandwidth parameter for each domain

$$K_{h_s,h_r}(x) = \frac{C}{h_s^2 h_r^p} k(\|\frac{x^s}{h_s}\|^2) k(\|\frac{x^r}{h_r}\|^2) \tag{2}$$

where x^s is the spatial part, x^r is the range part of a feature vector, $k(x)$ the common profile used in both two domains, h_s and h_r the employed kernel bandwidths, and C the corresponding normalization constant. In practice, we need to choose the bandwidth parameter $h = (h_s, h_r)$ and maximum pixel size M for best clustering. Fig. 5 (a), shows SfM result of an example image and its clustering result shows in Fig. 5 (b) with clustering parameter $(h_s, h_r, M) = (12, 6, 50)$.

After clustering process, each pixel in each cluster recovers depth using the nearest neighbor search based on Euclidean-distance. The coordinates of road region obtained by the SfM module, are particularly uti-

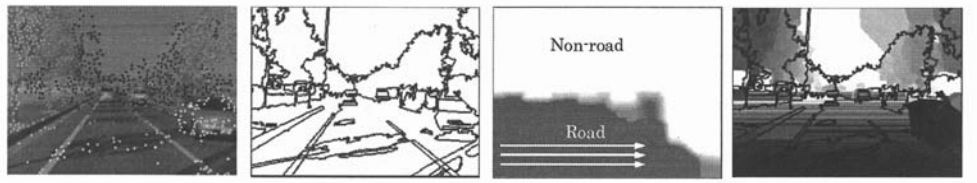


Figure 5: (a) SfM result image with feature points (b) Clustering results by Mean-shift algorithm (c) Road and non-road region obtained by the SfM module (d) Depth map : the closest object-black, the farthest object-white

lized in horizontal-centered distance for nearest neighbor search as shown in Fig. 5(c). Because the depth of road region horizontally changes along the perspective lines in a road scene, feature points of road region are calculated in nearest horizontal point. Finally, each pixel in each clustered region is assigned to the nearest depth value of same cluster, producing depth map as shown in Fig. 5(d).

The window size of a mask pattern can represent r of Fig. 4(b) and is to be decided from final depth map. Through the camera calibration, the distance r is determined linearly in histogram of depth map. Therefore, feature vectors can be extracted from HLCA with an adapted window size of mask pattern dependence on depth. For discriminant function, we employed a widely used simple classifier, Linear Discriminant Analysis (LDA).

4 Experimental Results

This section presents our experimental results for texture segmentation by using the proposed method. We investigate the performance of our system on several video frame, and compare our results with conventional method.

The images were captured using a multiband camera mounted on a moving vehicle. We developed a multiband camera for the purpose of integrating a color camera and a near infrared camera. The developed multiband camera can simultaneously obtain both images of color and near infrared wavelengths [15].

The appearance and structure of the multiband camera is shown in Fig. 6. The multiband camera removes the infrared cut-off filter and a special color filter was designed. Bayer filter [16], which is generally used in most of the conventional digital color camera has a pixel array of 50% green, 25% red and 25% blue patterns, and is referred to as RGBG or GRGB. The special filter of the multiband camera rearranges the 50% green pixel array into 25% green and 25% infrared cell onto a square grid of image sensors. The 25% infrared array acts as a visible light cut-off filter.

One band image of near infrared is utilized in process of the SfM module and a color image of three bands (RGB) and a multiband image of four bands (infrared+color) are utilized in texture segmentation experiment. We also compared the results of multiband image with four bands to the results of color image with three bands. The input image resolution was 320 × 240 pixels, and the frame rate was 30fps.

One of the input video frames and its ground truth image are displayed in Fig. 7(a) and Fig. 7(b), respectively. The result image obtained by our SfM module is shown in Fig. 7(3). The pink rectangle and the green region represent a moving object region and the road region, respectively. The feature points with 3D coordinates represent in colored point such as closer feature points change color into red from black. Selected fea-

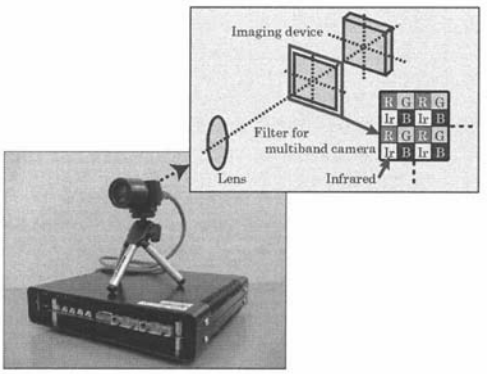


Figure 6: Appearance and structure of multiband camera with special color filter array

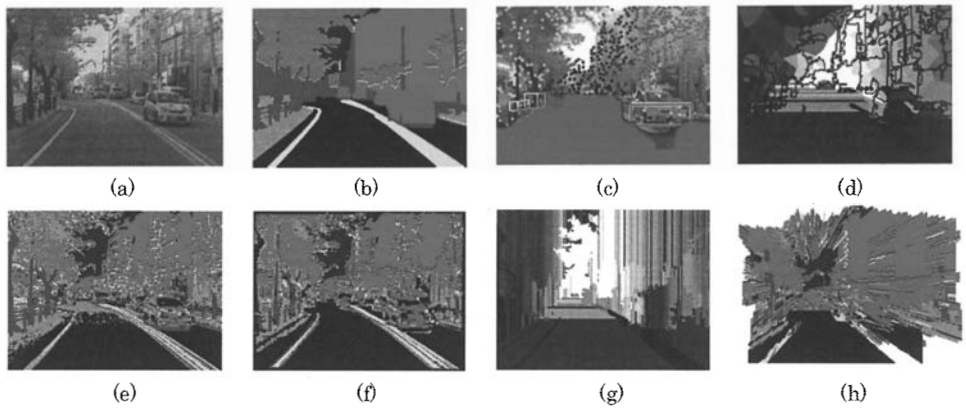


Figure 7: Experimental results (a) Input color image (b) Ground truth input image (c) SfM result image (d) Depth map after clustering (e) Segmentation result of conventional method (f) Segmentation result of the proposed method (g) Last depth map after segmentation (h) 3D texture segmentation

Table 1: Recognition rate of several experiments

	Conventional method (2nd-order HLAC)	Proposed method			
		Only color image(RGB)		Multiband image(infrared+RGB)	
Window size (r)	1	1 to 4	1 to 12	1 to 4	1 to 12
Recognition rate	67.8%	72.6%	74.1%	76.1%	78.4%

ture points are approximately 600 up to 1000 points per one video frame. Fig. 7(d) shows the depth map using Mean-shift algorithm with clustering parameters, (h_s, h_r, M) = (6, 8, 50) and nearest neighbor search.

To get the training patterns, we sampled features from a first video frame. We then made the feature vectors consist of the 99 dimensions for color image, (33 dimension × 3 bands-RGB) and 132 dimensions for infrared and color band (33 dimension × 4 bands-infrared+RGB). Each object in an input image was classified into eight classes and assigned a color. The assigned colors are : road-black, tree-green, trunk of tree-brown, building-pink, sky-blue, lane-yellow, sidewalk-gray, car-red, redundancy-violet.

Fig. 7(e) shows a result of test image from a next video frame. Fig. 7(e) results from conventional HLAC features, which is limited in second order and extracted the feature vector from uniform mask size (r=1, window size = 3 by 3). Fig. 7(f) shows the result of the proposed our segmentation method with different size of mask pattern according to depth (r=1 to 4, window size = from 3 by 3 to 9 by 9).

As a result, we can see that road, tree, and lane regions are recognized more accurate in proposed method. Because, texture patterns of these regions include variable features, which are sensitive to perspective structure in a 2D road scene. In addition, all pixels in classified region still have depth value using the depth map of Fig. 7(d), so that we can localize classified texture patterns in 3D. Image filtering makes segmented region smooth and the mean value of depth is computed in representinf of 3D reconstruction. The road patterns are assigned to horizontal mean value, other patterns are assigned to vertical mean, and the sky patterns is assigned to most large value of depth as shown in Fig. 7(g). The result of 3D image segmentation shows in Fig. 7(h).

In Table 1, the recognition rate of proposed method is higher than that of the conventional method with 10 video frame. It should be noted that the result of the multiband image is better than that of only color image with three bands. Our best recognition rate (78.4%

from multiband image) is comparable or better results than other segmentation work [9], which gave overall pixel-wise accuracy, 72.2% with their image database. Therefore, we can confirm that the proposed system is expected to play an important role in complex scene understand system for road environment perception.

5 Conclusion

This paper presented a new framework to integrate SfM module and texture segmentation scheme for road environment perception. The SfM module presented a novel method of estimating the ego-motion of a vehicle and of detecting moving objects on a road. Our SfM algorithm can accurately estimate the vehicle ego-motion in severe situations, such as when there are moving objects and when the camera is moving nearly parallel to its optical axis. In texture segmentation, depth map of an input image was generated by Mean-shift and nearest neighbor search algorithm. The relative resolution of a texture pattern was determined by obtained depth map. Texture features can be extracted from HLAC functions with variable mask size. The proposed system can not only effectively classify the texture patterns but also represent in 3D texture segmentation. By integrating other scene interpretation system, the proposed system can expand into a dynamic 3D scene analysis system for vehicle environment perception in the future.

References

[1] Bertozzi, M., Broggi, A., Cellario, M., Fascioli, A., Lombardi, P., Porta, M.: Artificial vision in road vehicles. Proc. of the IEEE - Special issue on Technology and Tools for Visual Perception. (2002) 90(7):1258-1271

[2] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust Object Recognition with Cortex-Like Mechanisms. IEEE Trans. Pattern Anal. Mach. Intell. (2007) 29(3) 411-426

[3] Leibe, B., Cornelis, N., Cornelis, K., Gool, V. L.: Dynamic 3D Scene Analysis from a Moving Vehicle. IEEE Conf. Computer Vision and Pattern Recognition (2007)

[4] Cornelis, N., Leibe, B., Cornelis, K., Van Gool, L.: 3D Urban Scene Modeling Integrating Recognition and Reconstruction. International Journal of Computer Vision, 2007 (in press)

[5] Harris, C. and Stephens, M.: A combined corner and edge detector. Proc. Alvey Vision Conference. (1988) 147-151

[6] Lucas, B.D. and Kanade, T.: An iterative image registration technique with an application to stereo vision. International Joint Conf. on Artificial Intelligence. (1981) 674-679

[7] Hartley, R.: In defence of the eight-point algorithm. IEEE Trans. Pattern Anal. Mach. Intell. (1997) 6(19) 580-593

[8] Fischler, M.A. and Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM. (1981) 6(24) 381-395

[9] Shotton, J., Winn, J., Rother, C., Criminisi, A.: TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation. Proc. of the 9th European Conf. on Computer Vision. (2006)

[10] Winn, J., Criminisi, A., Minka, T.: Object Categorization by Learned Universal Visual Dictionary. Proc. of the 10th IEEE International Conf. Computer Vision. (2005) 1800-1807

[11] Toyoda, T., Hasegawa, O.: Extension of higher order local autocorrelation features. Pattern Recognition. (2007) 40 1466-1473

[12] Kurita, T., Otsu, N.: Texture classification by higher order local autocorrelation features. Proc. of Asian Conf. on Computer Vision (1993) 175-178

[13] Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. (2002) 24(5) 603-619

[14] Christoudias, C., Georgescu, B., Meer, P.: Synergism in low-level vision. 16th Int. Conf. on Pattern Recognition (2002) IV 150-155

[15] Kidono, K., Ninomiya, Y.: Visibility Estimation under Night-time Conditions using a Multiband Camera. Proc. of IEEE Intell. Vehicles Symposium. (2007)

[16] Gunturk, B. K., Altunbasak, Y., Mersereau, R.: Color plane interpolation using alternating projections. IEEE Trans. Image Process. (2002) 11(9) 997-1013