The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Monocular Depth Cue Fusion for Image Segmentation and Grouping in Outdoor Navigation

Wenhui Zhou, Lili Lin, Bin Lou and Xuehui Wei

***Abstract*— This paper proposes an efficient fusion strategy of monocular depth cue and other image features for natural image segmentation and grouping. The main idea is to improve the performance of image clustering via fusing depth cue, color, spatial location, and edge confidence in six-dimensional color-depth feature space. It integrates the monocular depth cue estimation, mean shift filtering and graph cuts algorithm together. Firstly, the dark channel prior based atmospheric transmission estimation is employed to recover monocular depth cue. Then the mean shift filtering in the weighted color-depth space is proposed to obtain cluster regions with correct boundaries. Finally, graph cuts algorithm is applied to achieve the final regional grouping. Experimental results indicate the proposed method has excellent performance in outdoor natural environments.**

I. INTRODUCTION

Natural image segmentation and grouping is a crucial task of visual navigation for outdoor autonomous robot [1] and driver assistance system [2]. Because of the complexity of outdoor natural environments and the lack of scene geometry information, 2D (two-dimensional) image segmentation and grouping is still a challenging task.

Most early works are dedicated to find solutions from one image character, such as color [3], feature [4], texture [5], etc. G. DeSouza [1] and Z. Sun [2] had made excellent overviews. Since it is difficult to depict outdoor natural scenes with single image character, these methods are limited to some specific scenes that are conformed to certain scene model assumptions.

Many recent works focus on multiple image features grouping and supervised/unsupervised classification learning to solve the image segmentation and grouping problems in ill-structured environments, such as obstacle detection, road following [6], [7], [8]. Their main disadvantage is the lack of guidance from scene geometry and all road models should be learned in advance. Therefore, these algorithms are not suitable for the first exploration or roving task in unknown environments.

In order to overcome the disadvantages of visual methods, many multi-sensor fusion based complementary algorithms have been developed. Lidar range data are commonly used to aid in lane detection and obstacle avoidance [9]. In the US Army's DEMOIII project [10] and the DARPA Grand Challenge [11] / Urban Challenge [12] programs, these algorithms had been widely used and successfully demonstrated their effectiveness and performance. However, the joint calibration and data fusion between different sensors are difficult, and have great impact on the system performance.

Wenhui Zhou is with the College of Computer and Software, Hangzhou Dianzi University, Hangzhou, China. zhouwenhui@hdu.edu.cn

Lili Lin is with the College of Information and Electronic Engineering, Zhejiang Gongshang University, Hangzhou, China. sunshine@hzcnc.com

Bin Lou is with the College of Computer and Software, Hangzhou Dianzi University, Hangzhou, China. loubin@hdu.edu.cn

Xuehui Wei is with the College of Computer and Software, Hangzhou Dianzi University, Hangzhou, China. weixuehui@hdu.edu.cn

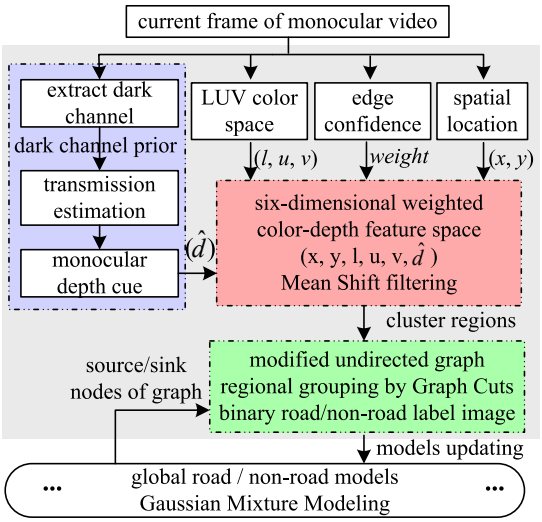


Fig. 1. System Components.

In this paper, we make efforts to show how scene geometry cues from a single image can be incorporated into image segmentation and grouping. Although inferring the depth information from a single image is still a longstanding unsolved problem in computer vision [13], some monocular depth cues have already been proposed, such as texture variations and gradients, haze, defocus, etc. [14]. The depth cue extracted from atmospheric haze will be employed here, because it is a fundamental cue for human to perceive depth [15]. The main motivation of this paper is to propose an efficient monocular depth cue fusion strategy, and show its effects in visual navigation.

There are three main stages in our methods, as shown in Fig.1. Firstly, monocular depth cue is extracted based on dark channel prior and atmospheric transmission estimation. Then monocular depth cue, color, spatial location, and edge confidence are combined together into a 6D (six-dimensional) weighted color-depth feature space by mean shift filtering. Finally, regional grouping can be achieved by modified graph cuts.

II. Monocular Depth Cue Estimation

As well known, it is an ill-posed problem to recover 3D structure from a single image. In imaging process, fortunately, the appearance of haze resulted from atmospheric light scattering can provide significant distant or depth cue. Due to atmospheric absorption and scattering, only part of the light reflected from distant objects reaches the camera. Furthermore, this light is mixed with scattered ambient light between the object and camera. Thus, distant objects in the scene typically appear considerably lighter and featureless, compared to nearby ones [13]. This phenomenon always occurs in the natural images and causes different degrees of haze, even if the photographs are taken in the clear-sky conditions.

When the atmosphere is homogenous, the relationship between observed image **I**, haze-free image **J**, and optical distance d_i can be formulated as follows.

$$\begin{cases} I_i^c = t_i J_i^c + (1 - t_i) A^c, c \in \{r, g, b\} \\ t_i = e^{-\beta d_i} \end{cases} \tag{1}$$

where I_i^c and J_i^c is a color channel of the pixel i in **I** and **J**, respectively. A^c is a color channel of global atmospheric light **A**. t_i is atmospheric transmission, and β is the scattering coefficient of the atmosphere.

Equation (1) is the standard image formation model. It indicates that the observed image is a combination of two components. One is exponentially attenuated scene radiance with its optical distance. The other is scattered atmospheric light towards the camera. Obviously, the transmission map **t** (the vector composed of t_i) needs to be estimated to obtain the depth cue of each pixel.

Before taking further steps, we need to discuss the validity of the image formation model in Equation (1) for visual navigation. As pointed out by Kopf [13] and Narasimhan [16], this model assumes single-scattering and a homogeneous atmosphere, and it is more suitable for short ranges of distance and might fail to correctly approximate the attenuation of scene points that are more than a few kilometers away. Since the effective visual distance of an optical navigation camera is usually 5 to 70 meters in front of the robot, we can infer the assumption is satisfied in visual navigation. Fig.2 shows some depth cue estimation results, and the original scene images are selected from CMU/VASC image databases [17], which are taken from various Navlabs.

In order to estimate transmission map **t** in Equation (1), some prior knowledge must be required, because there is an intrinsic ambiguity between local image features and depth variations. Recently, many outstanding prior-knowledge-based algorithms have been proposed for transmission estimation [15], [18], [19].

In this paper, we employ the method of K. He *et al.* [15], because it is simple but effective. Their main contribution is introducing the dark channel prior to estimate the transmission and the atmospheric light.

Dark channel prior is based on an observation that there is at least one color component near zero in outdoor haze-free

Fig. 2. Depth cue estimation results of some outdoor scene images in CMU/VASC image databases, which are selected from CMU/VASC image database: apr10-87-bright, jan25-91, jan31-91-down, july4-92-run6a, respectively.

image. The dark channel of haze-free image **J** can be defined as follows.

$$J_i^{dark} = \min_{c \in \{r,g,b\}} \left(\min_{j \in \Omega_i} (J_i^c) \right) \to 0 \tag{2}$$

where Ω_i is a local window centered at pixel i.

Substituting Equation (2) into Equation (1) and solving the minimization, we can obtain the transmission estimation.

$$\begin{cases} \tilde{t}_i = 1 - \hat{I}_i^{dark} \\ \hat{I}_i^{dark} = \min\limits_{c \in \{r,g,b\}} \left(\min\limits_{j \in \Omega_i} \left(\hat{I}_i^c \right) \right), \ \hat{I}_i^c = \frac{I_i^c}{A^c} \end{cases} \tag{3}$$

where $\hat{I}^{dark}$ is the dark channel of normalized image $\hat{\mathbf{I}}$.

To ensure neighboring pixels have similar transmission values, a soft matting algorithm [20] is employed to refine the estimation result.

With the estimated transmission map, we can extract the monocular depth cue map $\hat{d}$ according to Equation (1).

$$\hat{d}_i = -k \ln \tilde{t}_i \tag{4}$$

where k is a constant scale factor, and its value is 255 in this paper.

(a) Original image frame

(b) Dark channel

(c) Estimated transmission map

(d) Monocular depth cue map

Fig. 3. Depth cue estimation result of a frame image in our experiments.

Fig.3 shows the monocular depth cue estimation result of a frame image in our experiments. Fig.3(a) is the original image frame, and Fig.3(b) to Fig.3(d) are its dark channel, estimated transmission map and extracted depth cue map, respectively. More results of our outdoor experiments are shown in section V.

III. Color-Depth Space Clustering

The key idea of our method is to fuse depth cue with other image features. Mean Shift algorithm [21] is a robust nonparametric clustering approach in high-dimensional feature space. It can deal with unknown number of clusters effectively. In this paper, we construct a 6D weighted color-depth feature space by adding monocular depth cue into the conventional color-spatial space.

A. Weighted Color-depth Feature Space

Let $\mathbf{x}_i^s$ and $\mathbf{x}_i^c$ be the spatial coordinate vector $(x_i,\, y_i)$ and LUV color vector $(l_i,\, u_i,\, v_i)$ of image pixel i, respectively. Let x_i^d be the depth cue of pixel i. The feature vector of each pixel i can be defined by $\mathbf{x}_i = \left(\mathbf{x}_i^s,\ \mathbf{x}_i^c,\ x_i^d\right)$. It is a 6D color-depth feature space. The physical meaning of this 6D feature space clustering is the pixels contained in a cluster are not only similar in color and contiguous in the image, but also continuous in depth cue.

To ensure the clustering stability and accuracy at the edge pixels, edge confidence [22] is considered as weights in mean shift filtering step to refine the delineated region boundaries.

Compared with most gradient based edge detectors, edge confidence can detect the sharp edges with small magnitudes. It constructs a hyperspace by gradient subspace and its orthogonal complement. The gradient subspace is defined by two differentiation masks with $n \times n$ windows. The edge gradient magnitude ρ is the projection of actual edge a onto the gradient subspace. The edge confidence η is the inner product between a and the ideal edge template.

Thus, the weight of pixel i is defined as follows.

$$w_i = 1 - (\alpha_i \cdot \rho_i + (1 - \alpha_i) \cdot \eta_i) \tag{5}$$

where α_i is a ratio. Since pixels close to an edge have small weights, the discontinuity preserving property of the mean shift filtering is further enhanced.

B. Mean Shift in Weighted Color-depth Space

To perform the mean shift clustering algorithm in the weighted color-depth feature space, the mean shift vector $m_{h,K}(\mathbf{x})$ should be redefined as follow.

$$m_{h,K}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i w_i k\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} w_i k\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x} \tag{6}$$

where $\mathbf{x}_i$ and w_i, i=1,... n, are vectors and their weights, respectively. $k(\cdot)$ is the profile of the kernel, and $\mathbf{x}$ is the center of the kernel (window).

In conventional mean shift algorithm, Euclidean metric is used for color and spatial space. However, it is not suitable for depth cue space, because the nature of the depth cue is different from that of the color and spatial space. For the joint color-depth domain, their different nature might be compensated by proper normalization. For this reason, the smooth constraint in the global stereo correspondence is used for depth cue space, and a metric similar with the Potts model is employed as follows.

$$\left\|\frac{x^d - x_i^d}{h_d}\right\| = \begin{cases} 0, \text{ if } \left|x^d - x_i^d\right| < h_d \\ T, \text{ otherwise} \end{cases} \tag{7}$$

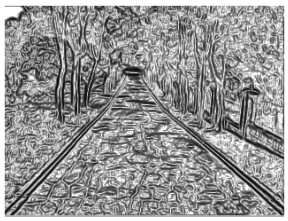
where T is some constant.

Thus, the multivariate kernel of color-depth domain is defined as the product of three radial symmetric kernels.

$$K_{h_{s,c,d}}(\mathbf{x}) = \frac{C}{h_s^2 h_c^3 h_d^1} k\left(\left\|\frac{\mathbf{x}^s}{h_s}\right\|^2\right) k\left(\left\|\frac{\mathbf{x}^c}{h_c}\right\|^2\right) k\left(\left\|\frac{x^d}{h_d}\right\|^2\right) \tag{8}$$

where $\mathbf{x}^s$ is the spatial part, $\mathbf{x}^c$ is the color part and x^d is the depth cue part of a feature vector, h_s, h_c and h_d are kernel bandwidths, and C is the corresponding normalization constant.

Due to limited space, we use the same input image as that of Fig.3, and the comparison results of mean shift clustering in different feature spaces are shown in Fig.4. To be clear, white boundaries are overlayed on the corresponding results. Fig.4(a) is the weight map computed by edge confidence. Fig.4(b) to Fig.4(d) are the results of mean shift in weighted color-spatial space, non-weighted color-depth space, and weighted color-depth space, respectively. Obviously, there are many over-segmented regions and some inaccurate boundaries in Fig.4(b), and Fig.4(c) shows under-segmentation regions with error boundaries. These typical errors are marked by red circles. Comparatively, Fig.4(d) has the best clustering result.



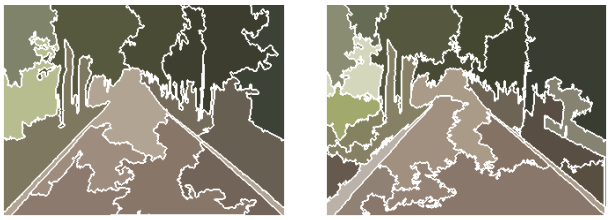
(a) Weight map

(b) Weighted color-spatial space

(c)Nonweighted color-depth space

(d)Weighted color-depth space

Fig. 4. Results of mean shift clustering in different feature spaces with the parameters h_s=7, h_c=6.5, h_d=1.0, α_i=0.3, and minimum region size=2000.



(a) Weighted color-spatial space with different parameters

(b) Weighted color-depth space with different parameters

Fig. 5. Results with different clustering parameters. The first column results with the parameters h_s=3, h_c=3.5, h_d=1.0, α_i=0.3, minimum region size=2000. The second column results with the parameters h_s=7, h_c=6.5, h_d=1.0, α_i=0.3, and minimum region size=1000.

C. Parameters selection

In order to ensure high quality regional grouping, over-segmentation with accurate region boundaries is necessary in the color-depth space clustering. Of course, too many and too small regions would also affect the performance of the regional grouping.

There are five key parameters in our method controlling the level of over-segmentation. They are the ratio (α_i) in Equation (5), three kernel bandwidths (h_s, h_c and h_d) in Equation (8), and the size of minimum region in the cluster fusion process. In Fig.4, the parameters are h_s=7, h_c=6.5, h_d=1.0, α_i=0.3, and the size of minimum region is 2000.

The ratio α_i can improve the accuracy of region boundaries. The size of minimum region can avoid too small regions. The color and space feature kernel bandwidths (h_s and h_c) control the number of over-segmented regions. The depth cue kernel bandwidth h_d aims to reduce the region number and refine the boundaries.

Fig.5 shows the clustering results with different parameters, using the same input image as the above examples. Compared with Fig.4(b) and Fig.4(d), the first column results are obtained with different h_s (=3) and h_c (=3.5), and the second column results are obtained with different minimum region size (=1000). Obviously, the smaller h_s and h_c values or minimum region size are, the more over-segmented regions are. Moreover, the function of monocular depth cue in image clustering is demonstrated in Fig.5 very well. The second row results have fewer regions and better boundaries than those of the first row.

IV. Modified Graph Model

Classification and grouping problem can be described as a labeling process for each pixel in the image. In conventional graph-based image classification and grouping methods, it can be converted to a minimum cut problem of a pixel-based graph model.

Let f=(f_1,...,f_i,...) is a binary vector. f_i is the label of pixel i, which can be annotated as either "object" or "background". The vector set f defines a partition. So it can be formulated in terms of energy minimization in MAP-MRF framework. The standard energy function can be formulated into a data term and smoothness term, as shown in (9).

$$E(f) = \lambda \cdot \sum_{i \in \mathcal{I}} R_i(f_i) + \sum_{i,j \in \mathcal{N}} B_{i,j} \cdot \delta(f_i, f_j) \qquad (9)$$

where $\delta(f_i, f_j) = \begin{cases} 1 & f_i \neq f_j \\ 0 & \text{otherwise} \end{cases}$, R_i represents the characteristic of the region, and it is the cost for pixel i assigned with either "object" or "background". $B_{i,j}$ represents the characteristic of the boundary, it is the luminance or color discrepancy degree between the pixels i and j, which is often reduced to a distance function of pixels i and j. It is proved Graph Cuts algorithm can solve these energy minimization problems effectively [23].

However, this method has two defects. One is the algorithm's computing speed decreases exponentially and the complexity increases exponentially with the size of image. The other defect is there are many isolated small regions or pixels on the segmented region boundary.

We proposed a modified graph model [24] to overcome these two defects. The modified graph model is constructed on the basis of accurate results of clustering. All pixels in a region can be regarded as a whole, and the node set of modified graph model is composed of these cluster regions. Since the number of regions is mainly related with the scene complexity rather than image size, it usually ranges from 20 to 50, which is much smaller than that of pixels. So the modified graph model has much simpler construction, and faster convergence speed in searching for optimal solution.

Followed the mean shift clustering, cluster regions based modified graph model is constructed, then graph cuts algo-

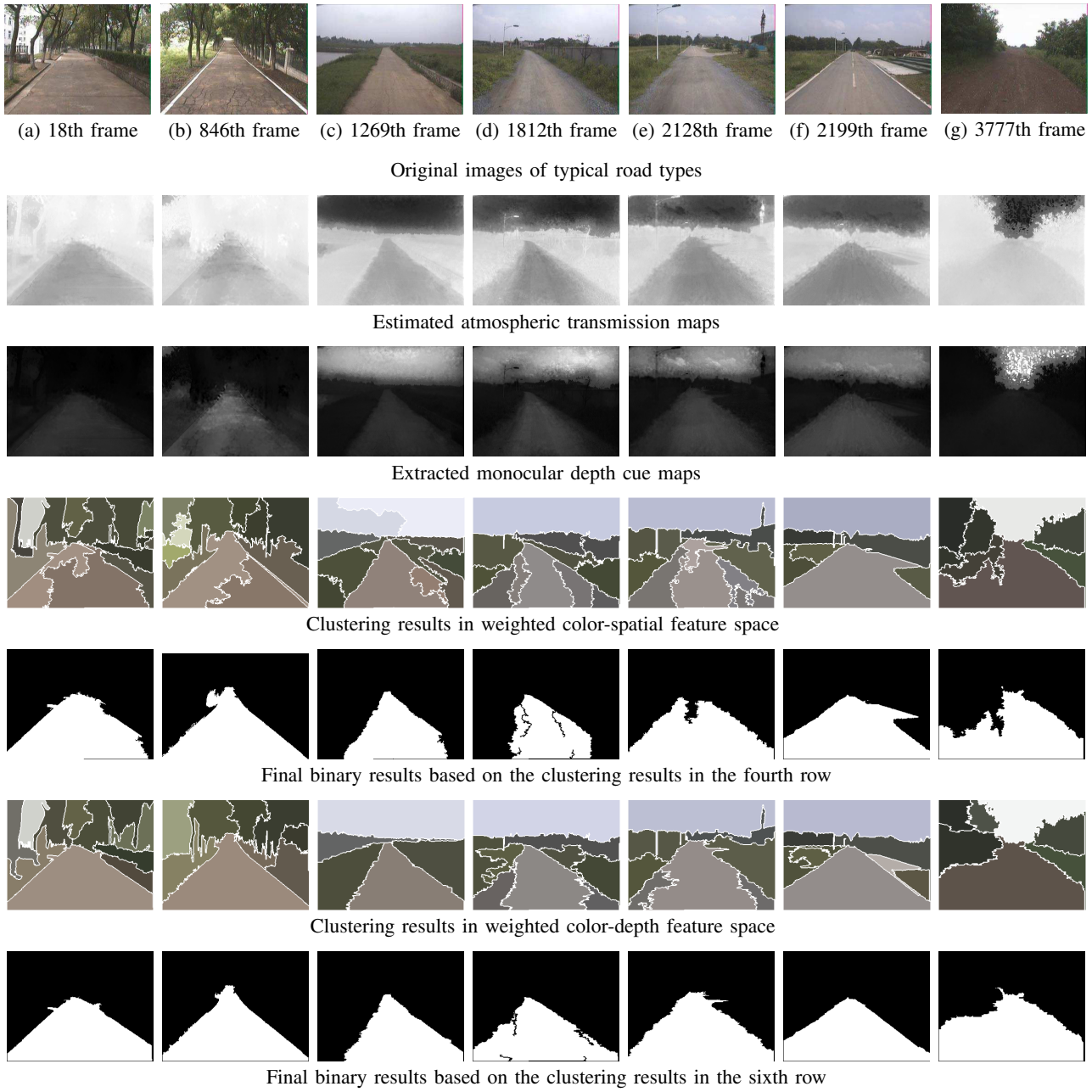


Fig. 6. Comparison results of several road types.

rithm is used to obtain the final binary object / background label image.

V. Experiments

In outdoor road following experiments, a circular experimental field, with a perimeter of about 2 kilometers, is selected to test our proposed algorithm. This experimental field contains four types of roads as follows: sandstone road, concrete road, avenue, and off-road. Moreover, the road regions and non-road regions often contain several color models. The onboard camera (SONY XC-555P) sits at 2.05 meters of altitude. Totally, 5527 frame images are taken in the experiments, and the image resolution is 320×240 pixels. The parameters of mean shift clustering are h_s=7, h_c=6.5, h_d=1.0, α_i=0.3, and the size of minimum region is 2000 pixels.

The typical frames of these road types are shown in the first row of Fig.6. The second row is the estimated transmission maps, and third row is the extracted monocular depth cue maps from transmission maps. The clustering and the final binary results of the proposed method are show in the sixth and seventh rows, respectively. For comparison, the corresponding results in weighted color-spatial feature space are shown in the fourth and fifth rows. To be clear, white boundaries are overlayed on the corresponding clustering

results.

Apparently, the clustering performance is crucial in our method. The correctness of region partition, especially the precision of the boundary, will directly affect the performance of road following. The comparison results indicate the monocular depth cue derived from dark channel prior can effectively guarantee both accuracy of region partition and discontinuity preserving property.

VI. Discussion and Conclusion

This paper proposes a natural image segmentation and grouping method based on an efficient fusion strategy of monocular depth cue and other image features. The main idea is fusing monocular depth cue, color, spatial and edge information into a six-dimensional weighted feature space. The experimental results demonstrate this method possesses excellent performance in complicated environments.

During the extensive experiments, we find the characteristics of monocular depth cue derived from atmospheric transmission map are only partly equivalent to those of depth range or disparity in traditional sense. In addition, it also represents many attributes of atmospheric transmission, scene radiance, surface material and reflection factor, etc. The main reason maybe is the parameter k in Equation (4) is treated as a constant. Fortunately, this does not seriously affect the performance of our method. Future work will further explore the character of monocular depth cue and focus on integrating more cues such as texture together.

Since the main aim of this paper is to explore the effect of monocular depth cue, no optimization strategy is adopted. So we do not discuss the average computational time of the proposed algorithm. Based on analytical result, the most time consuming computation is the soft matting algorithm, and the image pyramid scheme should be used to speed up the algorithm. This is also the important task of future work for the real time visual navigation.

VII. Acknowledgement

This work is supported by the National Natural Science Foundation of China (No.60902077), Zhejiang Province Social Development Projects of Major Science and Technology Projects of China (No.2008C13076), and Zhejiang Provincial Natural Science Foundation of China (No.Y1080967, Y1091074). The authors are grateful for the anonymous reviewers who made constructive comments.

References

[1] G. DeSouza and A. Kak, "Vision for mobile robot navigation: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 237–267, 2002.

[2] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 694–711, 2006.

[3] Y. He, H. Wang, and B. Zhang, "Color-based road detection in urban traffic scenes," *IEEE Trans. on Intelligent Transportation Systems*, vol. 5, no. 4, pp. 309–318, 2004.

[4] C. Rotaru, T. Graf, and J. Zhang, "Extracting road features from color images using a cognitive approach," in *IEEE Intelligent Vehicle Symposium*, London, UK, pp. 298–303.

[5] C. Rasmussen, "Texture-based vanishing point voting for road shape estimation," in *British Machine Vision Conference*, London, UK, 2004.

[6] C. Tan, T. Hong, T. Chang, et al., "Color model-based real-time learning for road following," in *IEEE Intelligent Transportation Systems Conference*, 2006, pp. 939–944.

[7] G. Zhang, N. Zheng, C. Cui, et al., "An efficient road detection method in noisy urban environment," in *IEEE Intelligent Vehicle Symposium*, Xi'an, China, 2009, pp. 556–561.

[8] M. Blas, M. Agrawal, A. Sudaresan, et al., "Fast color/texture segmentation for outdoor robots," in *IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems*, 2008, pp. 4078–4085.

[9] C. Rasmussen, "Roadcompass: Following rural roads with vision + ladar using vanishing point tracking," *Autonomous Robots*, vol. 25, no. 3, pp. 205–229, 2008.

[10] T. Hong, C. Rasmussen, T. Chang, et al., "Fusing ladar and color image information for mobile robot feature detection and tracking," in *7th International Conference on Intelligent Autonomous Systems*, Marina del Rey, CA, 2002.

[11] B. Siciliano, O. Khatib, and F. Groen, *The 2005 DARPA Grand Challenge: The Great Robot Race*, Springer-Verlag Berlin Heidelberg, 2007.

[12] B. Siciliano, O. Khatib, and F. Groen, *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic*, Springer-Verlag Berlin Heidelberg, 2009.

[13] J. Kopf, B. Neubert, B. Chen, et al., "Deep photo: Model-based photograph enhancement and viewing," *ACM Transactions on Graphics*, vol. 27, no. 5, pp. 116–1 – 116–18, 2008.

[14] A. Saxena, J. Schulte, and A. Y. Ng, "Depth estimation using monocular and stereo cues," in *International Joint Conference on Artificial Intelligence*, 2007.

[15] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 1956–1963.

[16] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather degraded images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 713–724, 2003.

[17] VASC Lab, "Cmu/vasc image database," `http://vasc.ri.cmu.edu/idb/html/road/index.html`.

[18] R. Fattal, "Single image dehazing," *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 72–1 – 72–9, 2008.

[19] P. Carr and R. Hartley, "Improved single image dehazing using geometry," in *Digital Image Computing: Techniques and Applications*, Melbourne, Australia, 2009.

[20] A. Levin, D. Lischinski, and Y. Weiss, "A closed form solution to natural image matting," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 228–242, 2008.

[21] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[22] P. Meer and B. Georgescu, "Edge detection with embedded confidence," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1351–1365, 2001.

[23] M. Kumar, P. Torr, and A. Zisserman, "Obj cut," in *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 18–25.

[24] L. Lin and W. Zhou, "A robust and adaptive road following algorithm for video image sequence," in *Inter. Conf. on Intelligent Computing*, 2007, pp. 1041–1049.