2D-TO-3D CONVERSION FOR SINGLE-VIEW IMAGE BASED ON CAMERA PROJECTION MODEL AND DARK CHANNEL MODEL

Tien-Ying Kuo, Yi-Chung Lo, and Chia-Chin Lin

Department of Electrical Engineering, National Taipei University of Technology,
Taiwan, R.O.C.
tykuo@ee.ntut.edu.tw, {yclo, cclin}@image.ee.ntut.edu.tw

ABSTRACT

This paper presents a novel technique for the depth estimation from the monocular outdoor images. This technique combines three major components: the camera projection model, object classifier, and dark channel modifier. The camera projection model estimates a reasonable relative depth by the camera parameters from the exchangeable image file format (EXIF). The object classifier categorizes objects into four types; sky, ground, man-made and natural objects, and then assigns initial depth to each object. Finally, the initial depth is further corrected by dark channel model. The experiments show our estimated depth map is close to the ground truth, as well as providing the satisfying stereo visual results.

Index Terms— depth map, perspective projection, object classification, horizon detection, vanishing point

1. INTRODUCTION

3D equipment has entered in our live, such as 3D display, stereoscopic capture, games and so on. The perception of 3D images is due to the parallax between the viewer's two eyes. Therefore, traditional stereo vision generation requires at least two images with slightly different projections. However, most existing digital photos were captured from monocular 2D format, and lack the corresponding depth maps to generate the perception of 3D image. Thus, generating a depth map from a single monocular image becomes an important issue for 2D-to-3D conversion.

There are many depth cues are able to be generated depth maps, such as relative position [1]-[6], linear perspective [1]-[5], atmospheric perspective [6]-[9], and texture gradient [7]-[9]. However, using one depth cue is not suitable for estimating all types of images. The literatures [2][3] classify the images to indoor, outdoor, and closed-up images, and then apply the different algorithms for each image type. In this study, we focus on the outdoor image processing.

Existing 2D-to-3D works are able to estimate acceptable depth, but the relative depth between two objects is not reasonable. Approaches based on statistical training with the ground truth database may approximate to some scenarios of images, but it is easy to fail with caveat that it's hard to adopt representative training features without loss of generality.

This study presents an accurate depth map with relatively low complexity for real practice. The proposed algorithm estimates an initial depth map, and then refines the initial depth map by dark channel model to generate a final depth map for 2D-to-3D conversion.

The rest of this paper is organized as follows. Section 2 describes the proposed algorithm. Section 3 describes the simulation model and gives the results. Finally, Section 4 draws conclusions.

2. PROPOSED ALGORITHM

The proposed algorithm is based on camera projection model, object classifier, and dark channel to create the depth map, as Fig. 1 shows. This algorithm includes two major functions of initial depth creation and final depth generation, to estimate the depth map from a single outdoor monocular image. Initial depth creation assigns an initial depth map to each object based on the image segmentation, the vanishing

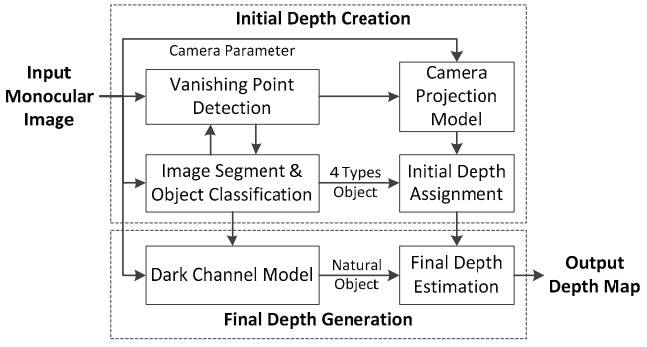

Fig. 1. Depth estimation block diagram

This work was supported by the National Science Council under Grant: NSC100-2221-E-027-080

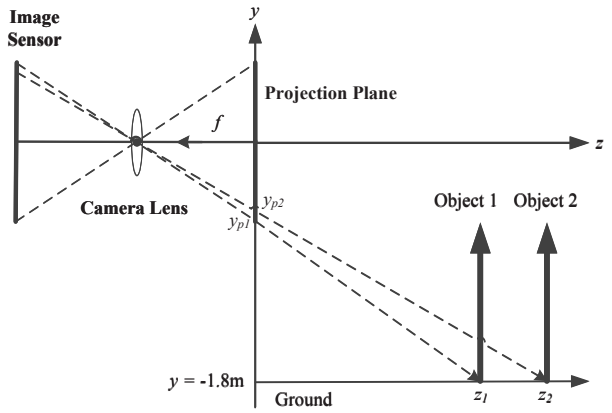

Fig. 2. Camera projection model

point detection, and camera projection model. Final depth estimation modifies the initial depth map by the dark channel information to estimate the output depth map.

2.1 Camera projection model

The purpose of the camera projection model is to estimate the reasonable relative depth by the camera parameters from the exchangeable image file format (EXIF). A. Matessi et al. [10] have proposed a simple model to discuss the transformations between the real world and the projection plane, as (1) shows. Our proposed algorithm improves this model to suit to real applications and correlates it to the camera parameters of EXIF, as Fig. 2 shows.

$$z = \frac{f \cdot y - f \cdot y_p}{y_p} \tag{1}$$

where z is the depth value, y notes the vertical position of a point in the real world, y_p indicates the vertical position of a pixel location in the projection plane, f represents the focal length of the camera lens. Since the projection plane is equivalent to the camera sensor, the vertical range of the projection plane is set as the camera sensor size. The lowest position of the vertical axis is set to a negative 1.8m if a photographer used the handheld camera to take the pictures. The depth value is able to be calculated from real world parameter and the camera lens parameter by (1).

2.2 Image segmentation and object classification

The object classification cooperates with Camera projection model to assign the depth. The stage segments and classifies the objects into the sky, ground, man-made, and nature objects. The sky objects are assigned the farthest depth; the grounds are assigned the relative depth based on the camera projection model; the man-made objects are also assigned the relative depth based on the camera projection model; the nature objects are based on the dark channel.

P. Felzenszwalb et al. have proposed a graph-based image segmentation algorithm [12]. We improve the segmentation performance by changing the preprocessing filter to bilateral filter [13] instead of Gaussian low-pass filter. After processing, the sky objects are classified by the color-based rule [2], the ground objects are classified by the texture variation, and the man-made and the nature objects are categorized by the object contour variation.

The nature objects have varied object contour. In the other word, the contour of the man-made objects usually constructs with straight lines. For example, the tree, flower, and mountain have few straight lines on their contours. Therefore, this algorithm employs the chain code [14] to count the corners and straight lines for each object contour. According to the ratio of the corners and straight lines, the man-made and nature objects are able to be classified.

2.3 Vanishing point detection

The man-made objects can be further classified into the single depth and multi-depth objects. All depths of the single depth object are assigned a depth value which gets from the lowest position of this object, and the multi-depth object has the gradient depths which change along with the vanishing line. The vanishing line connects the lowest point of an object and the vanishing point.

The vanishing point detection algorithm has been proposed by V. Cantoni et al. [15]. We improve the preference of the vanishing point detection by combining Hough Transform [15] and Random Sample Consensus (RANSAC) [16] algorithms. First, our proposed algorithm makes the edge map from the man-made objects. Then the edge map is transformed to the polar parameter space by Hough Transform, and selected the candidate pixels from the top group in the polar parameter space. Finally, we use RANSAC algorithm to find out a reasonable vanishing point.

2.4 Initial depth assignment

Before the initial depth assignment, the objects merging process should be performed. The man-made object usually is not flying in the air, except airplane. Therefore, the object in the air should be merged with the bottom adjacent objects. Almost all objects stand on the ground. Thus, the objects located above the horizon are defined as in the air. The horizon is a horizontal line which the line is tangential to the ground contour at the highest vertical point.

The initial depth assignment is based on the object types, and the relative depth is based on the camera projection model. First, the sky objects are assigned the farthest depth, the ground objects are assigned the relative depth which the depths in the object are calculated form (1) for each pixel. The depth of the single depth object is set by

the lowest position of this object. The depth values in multi-depth object are changed along with the vanishing line.

2.5 Dark channel model

The dark channel model is based on the light intensity reflected from the atmospheric haze commonly observed in outdoor images [17]. However, this model may contain halo effect near the depth discontinuities. S. Fang et al. [11] have improved this problem which the algorithm is similar to the superpixels [9],[18]. Normally, the nature objects have higher saturation values while the man-made objects usually are with lower saturation. Thus, the proposed algorithm masks the man-made objects, and only applies the dark channel model to the nature objects.

2.6 Final depth estimation

Our previous study [6] used the sigmoid function to combines the initial depth, saturation, and the dark channel to generate the depth map. However, the man-made objects usually are lower saturation and lead to the depth estimation error, such as road and building. The previous study [6] does not classify the nature and the man-made objects, and assigns wrong depth to the man-made objects when they are located in the middle of an image.

The object types have been classified in the initial depth assignment stage. Thus, the proposed algorithm assigns different weights to the initial depth results and the dark channel results, and fuses both results by the sigmoid function. The key technology for fusing the initial depths and the dark channel depth is the unit normalization. We assume that the middle distance locates at the same region of the initial depth and the dark channel depth. Lastly, the final depth is generated by applying the cross bilateral filter [3] to the fused depth map.

3. EXPERIMEMTAL RESULTS

The test inputs in this study were taken from Stanford University's 3D image database (Mark3D) [7]-[9], [19]. This image database includes a lot of outdoor images and provides the corresponding ground truth depth maps. The resolution of the outdoor images is 1704x2274. The ground truth depth maps are acquired by a laser distance scanner. The gray values of the ground truth ranging are from 1-81 with a resolution of 305x55. All outdoor images have been divided into two sets, learning set and test set. The learning set includes 400 outdoor images, and the test set includes 134 images. Our test-bed was implemented in C and run on a PC with Intel® Core™ i5-2400 3.1GHz CPU and 2GBytes ram.

Figure 3 presents the visual performance of four test images as examples. Fig. 3(c) shows the depth maps by Saxena's algorithm, in which the object boundary is indistinct; the depth assignment to ground objects is almost

(a) Original image (b) Ground truth (c) Saxena's (d) proposed

Fig. 3. Comparison of depth estimation results

the same; and the depth on the building area is almost wrong. Those problems could negatively affect the stereo vision quality after 2D-to-3D conversion. In contrast, Fig. 3(d) shows that our proposed algorithm provides a better depth map with sharper object shapes and better gradient effect. Some estimated depth given by our algorithm is even better than the ground truth if the object is located in the farther distance. The reason is that the farther distance areas are out of the working range of the laser distance scanner.

The proposed algorithm is much faster than Saxena's method. Although, Saxena's algorithm is wrote by the Matlab and C language, the proposed algorithm is a great speed up, above fourteen times.

This study also collects and tests images from handheld cameras as our next test set, named group 2 test. Figure 4 presents the visual performance of four test images as examples. The scenario in the group 2 is different from Stanford University's 3D image database, which include the ocean, bridge, and mountain. These estimated depth maps using Saxena's algorithm show obvious performance difference between Fig. 3 and Fig 4. The performance of Fig. 4(b) becomes worse than Fig. 3(c) in Saxena's method. Figure 4(c) shows that the proposed method provides much sharper depth on the boundaries of the tree, mountain, and

(a) Original image (b) Saxena's (c) proposed

Fig. 4. Depth estimation result comparison

bridge than Saxena's results, as Fig. 4(b) shows. Thus, Saxena's algorithm is only workable in the similar image composition as training database. On the other hand, the depth maps of Fig. 4(c) are assigned appropriate depth values from near end to far end, and display a clear shape for each object. In addition, the proposed method recognizes and assigns the gradient depth value to the multi-depth object, such as bridge in the Fig. 4(c). However, Saxena's algorithm assigns inaccurate depth to this kind of object, as the bridge is cont considered in training database.

4. CONCLUSION

The ground truth depth map can be obtained from the laser scanners, and the disparity map. However, the resolution of laser scanners is lower than the associated 2D image [20] and the disparity map is also affected by different algorithms. Therefore, 2D-to-3D conversion has become a mathematically ill-posed problem which admits an infinite number of solutions since true depth information of the scene are unknown [21]. In this study, we employ camera projection model to estimate a reasonable relative depth by the camera parameters from EXIF, and generate the final output depth from the initial depth results and the dark channel model results. The experimental results show the proposed algorithm is better than Saxena's. The proposed algorithm is also with much lower complexity than Saxena's.

5. REFERENCES

[1] Y. S. Huang, F. H. Cheng, and Y. H. Liang, "Creating Depth Map from 2D Scene Classification," *3rd Int. Conf. on Innovative Computing, Information and Control*, pp.69, 2008.

[2] S. Battiatoa, S. Curtib, M. L. Casciac, M. Tortorac, and E. Scordatoc, "Depth-map Generation by Image Classification," *Proc. SPIE on Three-Dimensional Image Capture and Applications*, vol. 5302, 95, April 2004.

[3] C. C. Cheng, C. T. Li, and L. G. Chen, "A Novel 2D-to-3D Conversion System Using Edge Information," *IEEE Trans. on Consumer Electronics*, pp. 1739-1745, Aug 2010.

[4] K. Han and K. Hong, "Geometric and Texture Cue Based Depth-map Estimation for 2D to 3D Image Conversion," *IEEE Int. Conf. on Consumer Electronics*, pp. 651-265, 2011.

[5] J. I. Jung and Y. S. Ho, "Depth map estimation from single-view image using object classification based on Bayesian learning," *Proc. IEEE Conf. 3DTV* (*3DTVCON*), pp.1-4, 2010.

[6] T. Y. Kuo and Y. C. Lo, "Depth Estimation from a Monocular View of the Outdoors," *IEEE Trans. on Consumer Electronics*, vol. 57, pp.817-822 , no. 2, 2011.

[7] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning Depth from Single Monocular Images," *Neural Information Processing System* 18, 2005.

[8] A. Saxena, M. Sun, and A. Y. Ng, "Learning 3-D Scene Structure from a Single Still Image," *In ICCV workshop on 3D Representation for Recognition*, 2007.

[9] A. Saxena, M. Sun and A.Y. Ng, "Make3D - Learning 3D Scene Structure from a Single Still Image," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, pp. 824, 2009.

[10] A. Matessi and L. Lombardi , "Vanishing Point Detection in the Hough Transform Space," *Lecture Notes in Computer Science*, pp. 987 - 994, 1999.

[11] S. Fang, J. Zhan, Y. Cao, and R. Rao, "Improved single image dehazing using segmentation," *IEEE Int. Conf. on Image Processing*, pp.3589, 2010.

[12] P. Felzenszwalb and D. Huttenlocher, "Efficient Graph-Based Image Segmentation," *Int. Journal of Computer Vision*, vol. 59, no. 2, pp. 167-181, 2004.

[13] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *Int. Conf. on Computer Vision*, pp. 839, 1998.

[14] H. Freeman, "Computer Processing of Line-Drawing Images, " *ACM Computing Survey*, vol. 6, pp.57-94, 1974.

[15] V. Cantoni, L. Lombardi, M. Porta, and N. Sicari, "Vanishing Point Detection: Representation Analysis and New Approaches", *Int. conf. Image Analysis and Processing*, pp. 90, 2001.

[16] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, pp. 381-395, 1981.

[17] K. He, J. Sun, and X. Tang, "Single Image Haze Removal Using Dark Channel Prior," *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1956-1963, June 2009.

[18] D. F. James, "Computer graphics: principles and practice," *Addison Wesley*, 1995.

[19] "Make3D," Available: http://make3d.cs.cornell.edu/code.html

[20] J. Zhu, L. Wang, R. Yang, and J. Davis, "Fusion of Time-of-Flight Depth and Stereo for High Accuracy Depth Maps," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[21]]W. Chenglei, G. Er, X. Xie, T. Li, X. Cao, and Q. Dai, "A novel method for semi-automatic 2D to 3D video conversion," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, pp. 65-68, 2008.