공학석사학위논문

단일 카메라를 이용한 3D 장면 이해를 위한 다중 지평면 추정

Multiple Ground Plane Estimation for 3D Scene Understanding using a Monocular Camera

2013 년 2 월

서울대학교 대학원

전기 컴퓨터 공학부

윤 상 두

단일 카메라를 이용한 3D 장면 이해를 위한 다중 지평면 추정

Multiple Ground Plane Estimation for 3D Scene Understanding using a Monocular Camera

지도교수 최 진 영

이 논문을 공학석사 학위논문으로 제출함

2012 년 12 월

서울대학교 대학원

전기 컴퓨터 공학부

윤 상 두

윤상두의 공학석사 학위논문을 인준함

2012 년 12 월

위 원 장 : 최 종 호 (印)

부위원장 : 최 진 영 (印)

위 원 : 오 성 회 (印)

ABSTRACT

In this paper, we propose a method for estimation of multiple ground planes using a stationary monocular camera. To estimate multiple ground planes, we perform three major steps. First, to estimate the number of ground planes, we create a histogram of votes with vanishing points and perform mean-shift clustering on this histogram. Second, to recover the active regions of multiple ground planes, we perform back-projection with the votes from the first step to extract trajectories which support each ground plane. We then estimate the active regions of each ground planes with these supporting trajectories. Finally, we efficiently normalize the relative depths of multiple ground planes with the speed of moving objects in the ground planes. In the experiments, we demonstrate that our method successfully estimates multiple ground planes and their relative depths.

Keyword : 3D Scene Understanding, Multiple Ground Plane Estimation, Monocular camera

Student Number : 2011-20888

CONTENTS

1 INTRODUCTION

1.1 Motivation

Recovering 3D structure of a scene from a video sequence is a challenging problem in computer vision. Knowing the 3D structure of a scene is critical in various applications, such as surveillance, behavior analysis, and object detection.

The first step understanding the 3D structure of a scene is to find the ground planes of an image. Using the ground planes, we can estimate 3D structure of an object placed on the ground plane. For this reason, various approaches have been proposed to infer ground planes in a scene and use them to understand 3D structure of the scene.

(a) (b)

Figure 1 : Example of a scene with multiple ground planes. To estimate the multiple ground planes, we have to know the number of ground planes, relative depth information, and their active regions as illustrated in (b).

1.2 Related work

Traditionally, multiple view geometry techniques [6] such as stereo vision or structure from motion have been used for ground plane estimation. 3D depth information and scene structures can be estimated for simple scenes with traditional methods, but not for complex scenes. Especially, the traditional methods are not enough to be applied to a scene with multiple ground planes. To solve this multiple ground plane problem, Hadsell et.al. [5] proposed a fitting method, which divides stereo points cloud into dominant ground plane and obstacle clouds. Lian et.al. [10] presented adaptive homography construction method for estimation of multiple ground planes. However, both methods utilize 3D information from multiple cameras, and therefore cannot be applied to 2D image from a monocular camera.

Recently, various researches have been conducted. Generally, the methods either use a single image or a video sequence from a monocular camera. With single image, the authors in [9, 3, 4] understand 3D scene by extracting vanishing points from edges and corners in the image. Their method assumes that the scene have strong edges and corners which can be easily found by edge detection algorithms. Therefore this assumption is not valid in outdoor scenes which obscure corner points exist. Saxena et.al. [12] built a 3D model that segments an image into many small planar surfaces to estimate 3D structure in outdoor scene. This method cannot obtain accurate depth information since it estimates depth from appearance features.

On the other hand, in [1, 7, 8, 11], 3D scene structure is estimated with video sequence. Breitenstein et.al. [1] proposed 3D scene learning method by assuming that the depth is inversely proportional to the detection window size. However, this approach cannot exactly infer the ground plane since the estimated depth is inaccurate. The authors in [7, 8] used vanishing points from the pedestrian tracking results to estimate camera matrix and structure of the scene. These auto-calibration methods are easy to lose robustness when tracking results are incorrect. Rother et.al. [11] proposed a simple method that learns horizontal line from human tracking results. The above methods work well in the case of the single ground plane, however, not in the case of the multiple ground planes as shown in Figure 1.

1.3 Overview

In this paper, we propose a method to estimate multiple ground planes with a video sequence from a monocular camera. Previous methods are restricted to a scene with single ground plane and do not consider the active region of the ground plane. However, to estimate multiple ground planes, we perform three major steps.

First, we figure out the number of ground planes in the scene automatically. To estimate multiple horizons, we create a histogram of votes for horizons with vanishing points obtained by tracking results. Then, with this histogram, we estimate multiple horizons using mean-shift clustering. Second, we recover the active regions corresponding to each ground plane. We conduct

back-projection on the histogram of votes, and extract the trajectories supporting each ground planes. The active regions of multiple ground planes are inferred with these trajectories. Finally, we normalize the relative depths of multiple ground planes. Since we obtain the relative depths separately, we need to combine the relative depths to understand overall depth information in the scene. We normalize the depths with the speed of the moving objects on the ground planes. The effectiveness of our method is validated through experiments on several video sequences containing multiple ground planes.

2 PROPOSED METHOD

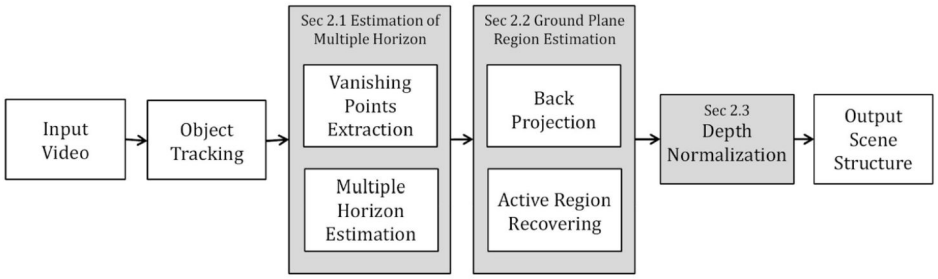


Figure 2 : The overall scheme

The overall scheme of the proposed method is shown in Figure 2. The contributions of this paper are denoted by shaded blocks. In our work, we assume that the height of a moving object in a video is constant and the camera is installed parallel to the horizon. Our method is based on vanishing points from moving objects as in [11], but our method is not restricted to a single ground plane. The method consists of three steps.

The first step is to find the multiple horizons corresponding to the ground planes. We build a histogram of votes and estimate multiple horizons within the scene. The relative depth information in the scene can be obtained using these horizons (detailed in Sec 2.1). The second step is to recover the regions of each ground plane in the scene. We determine the ground plane regions on the image plane using back-projection approach. In this step, the trajectories of moving objects on the ground plane are used to infer the regions for each

ground plane (detailed in Sec 2.2). The final step is to normalize the depth information from the estimated planes (detailed in Sec 2.3).

2.1 Estimation of Multiple Horizons

As the first step to understand a scene with multiple ground plane, we estimate the multiple horizons from vanishing points of tracking results. These horizons let us know the relative depth in their corresponding planes. To robustly estimate horizons regardless of the number of planes and with the inaccurate tracking results, we use voting-based method and mean-shift clustering with extracted vanishing points.

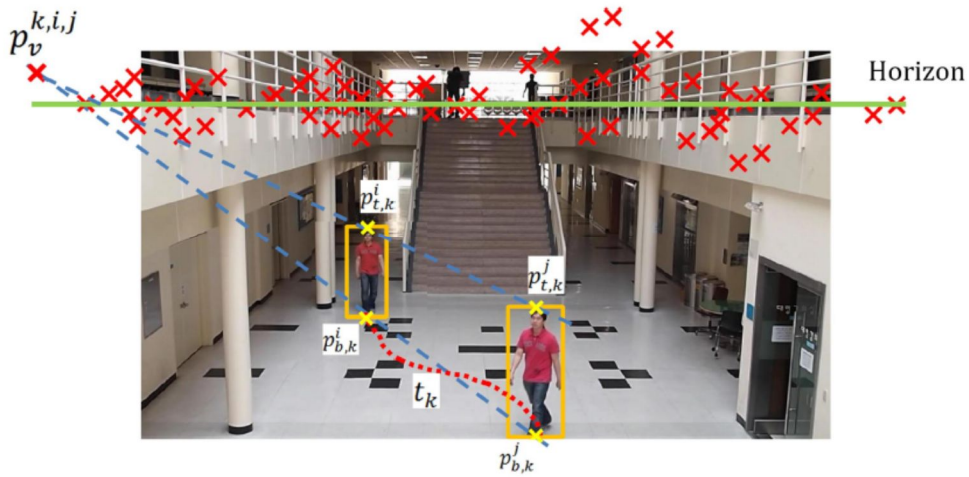


Figure 3 : Example of extracting vanishing points from trajectory and estimated horizon. The red cross maker means vanishing point and the green line means horizon.

2.1.1 Vanishing Points Extraction

We use the vanishing point extraction method based on the work of Rother et. al. [11]. We assume that the height of object is constant regardless of the object's position, and top center point and bottom center point of bounding box represents the pedestrian's head position and foot position respectively. We denote $p^{i}_{t,k}$, $p^{i}_{b,k}$ as the head position and foot position in i-th frame of k-th trajectory, and $t_{k,i:j}$ as the trajectory of index k from i-th frame to j-th frame. As shown in Figure 3, for the same person in different frame i and j, the vanishing point $p_v^{k,i,j}$ means the intersection of two straight lines connecting two head positions $p^{i}_{t,k}$ and $p^{j}_{t,k}$ and two foot positions $p^{i}_{b,k}$ and $p^{j}_{b,k}$ respectively. To reduce the influence of noise generated from inaccurate tracking result due to occlusion or shadows, a sufficiently large number of vanishing points are needed. To extract sufficient vanishing points from a moving trajectory, we sample trajectories set $t_{k,i:j}$ from k-th trajectory with randomly selected i and j ($i \neq j$). The sampling parameter η set the number of samples to $\frac{N_k}{\eta}$, where N_k is total frames of k-th trajectory. This constructs the vanishing points set P ($p_v^{k,i,j} \in P$), and we estimate the horizon lines of the scene as described in the next section.

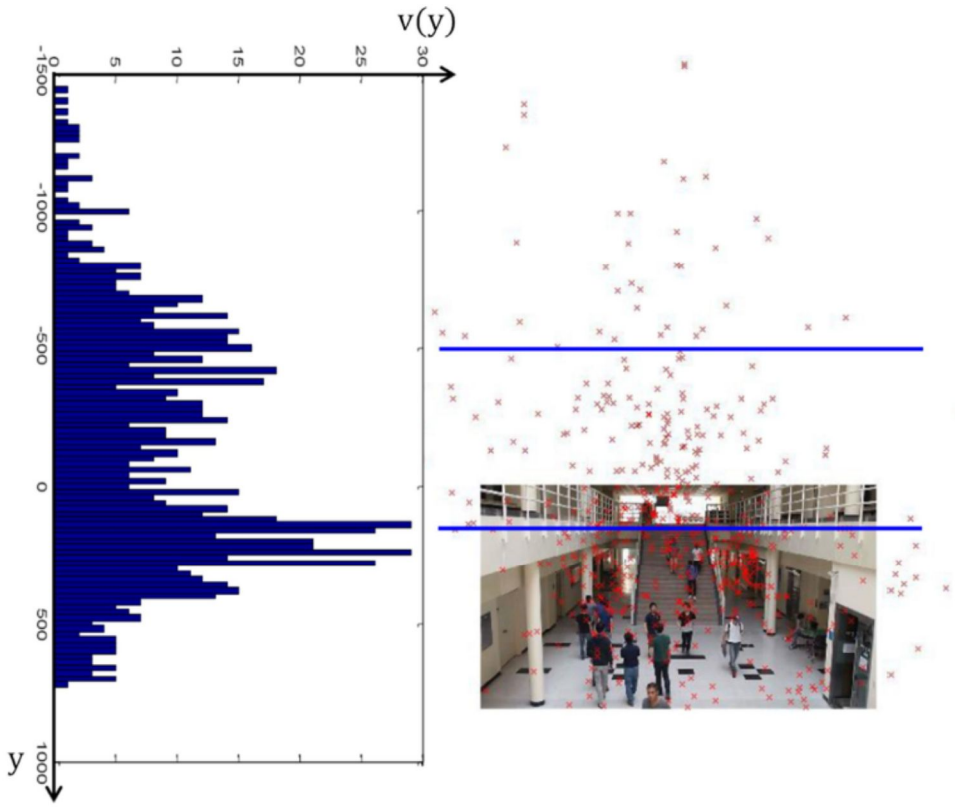


Figure 4 : Example of building voting histogram and estimating multiple horizons. The red cross maker means vanishing point and the blue line means horizon.

2.1.2 Multiple Horizon Estimation

In this step, we use a voting-based method to infer multiple horizons. With vanishing points in Sec 2.1.1, we transform vanishing points into voting space and build histogram $v(y)$ as

$$\mathrm{v}(\mathrm{y}) = \sum_{p_v \in \mathrm{P}} f(p_v, y),$$

$$f(\mathrm{p_v}, \mathrm{y}) = \begin{cases} 1 & if\ |p_v(y) - y| < \tau \\ 0 & otherwise \end{cases},$$

where $\mathrm{p_v}(\mathrm{y})$ is the y-coordinate position of a vanishing point $\mathrm{p_v}$, τ is the voting histogram bin size. This histogram represents how the vanishing points are distributed along the y-axis. We use this histogram to estimate the multiple horizons. After building the histogram on the voting space, we perform mean-shift clustering [2] to find local maxima from the histogram. Then, each local maxima are regarded as positions of horizons. For the mean-shift kernel, we use a Gaussian kernel to reduce the influence of noise in the histogram due to inaccurate vanishing point estimates. The shifted mean $\mathrm{m}(\mathrm{y})$ and kernel function K are given by

$$\mathrm{m}(\mathrm{y}) = \frac{\sum_{y_i \in \mathrm{N}(\mathrm{y})} K(y_i - y) y_i}{\sum_{y_i \in \mathrm{N}(\mathrm{y})} K(y_i - y)},$$

$$\mathrm{K}(\mathrm{y_i}, y) = \exp\left(-\frac{\left|\left| y_i - y \right|\right|^2}{c} \right),$$

where $\mathrm{N}(\mathrm{y})$ means the neighbors of y and the Gaussian parameter c is kernel window size. We conduct mean-shift clustering iteratively and find multiple local maxima of the histogram. An example of estimated positions of multiple horizons is illustrated in Figure 4.

2.2 Ground Plane Region Estimation

The estimated horizons in Sec 2.1, however, do not consider the actual scene structure and only infer coarse ground planes covering the whole scene. For this reason, the regions of inferred ground planes should be recovered as described in Figure 5. With sufficiently large number of object trajectories, we can infer the ground plane regions. For ground plane region estimation, we first assign trajectories on each horizon with respect to their positions, and then we recover active region of ground plane by voting method using those trajectories.

2.2.1 Back Projection

The region we would like to know in the ground plane is the region where the objects are actually moving on. However, since we use the trajectory information to recover the ground plane regions and some trajectories might have noises, we need to discard improper trajectories which are not accurate for describing the ground plane. The back projection scheme is proposed to extract the trajectories from vanishing points on the horizon. In this back-projection process, we define function g as below and find trajectory set $\mathrm{T_n}$ which is a set of trajectories related to the n-th horizon. The function g is

$$\mathrm{T_n} = \{t_{k,i:j} | g(p_v^{t,i,j}, l_n) = 1, \forall \mathrm{p_v^{k,i,j}} \in \mathrm{P}\},$$

$$\mathrm{g}(\mathrm{p_v^{k,i,j}}, l_n) = \begin{cases} 1 \;\; if \;\; \left| p_v^{(k,i,j)}(y) - l_n \right| < \delta \\ 0 \qquad otherwise \end{cases},$$

where $p_v^{k,i,j}(y)$ is the y-coordinate position of vanishing point $p_v^{k,i,j}$ and δ is the threshold for determining membership. Examples of set T_n are shown in Figure 5 (a) and (b). We estimate ground plane regions with these trajectories as described in the next section.

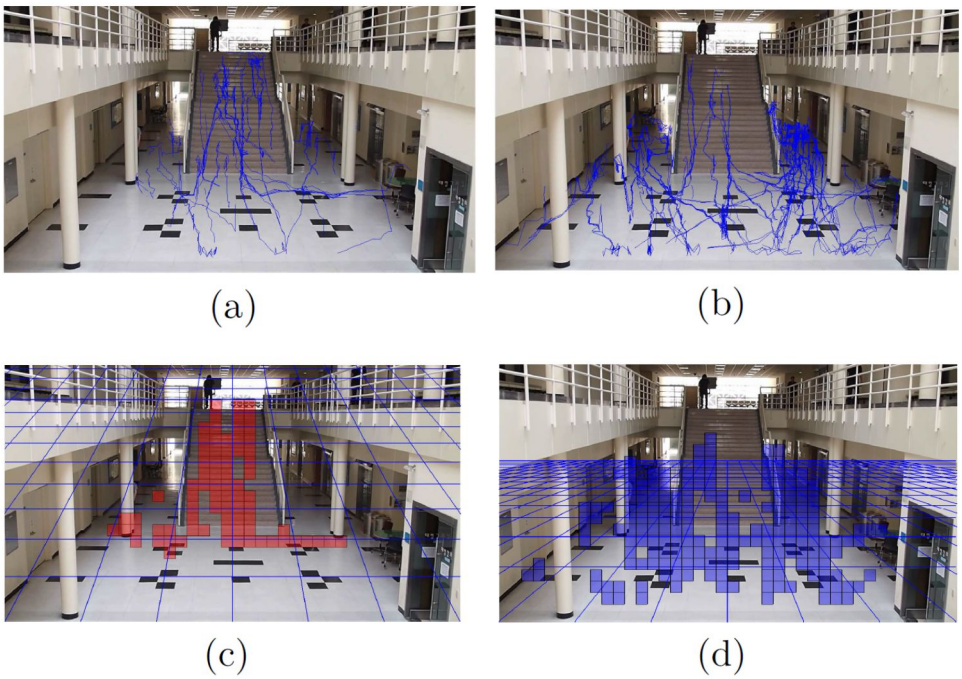


Figure 5 : Back-projected trajectories are shown in (a) and (b). The active regions of ground planes with relative depth are shown in (c) and (d).

2.2.2 Active Region Recovering

With the obtained trajectory set T_n, we use a simple method for estimating actual ground planes. We first divide the image plane into image cells of size

m ∗ m pixels. For the n-th ground plane, the active region voting map M_n is defined as

$$M_n[i,j] = \sum_{t \in T_n} h_{i,j}(t),$$

$$h_{i,j}(t) = \begin{cases} 1 & if\ t\ passes\ though\ image\ cell\ (i,j) \\ 0 & otherwise \end{cases},$$

where $M_n[i,j]$ means the (i,j)-element of M_n. Image cells whose votes are larger than a predefined threshold (γ) are defined as **active region**. Example of the recovered active region of ground planes with relative depth is shown in Figure 5 (c) and (d). In this way, we find areas that represent each multiple ground plane.

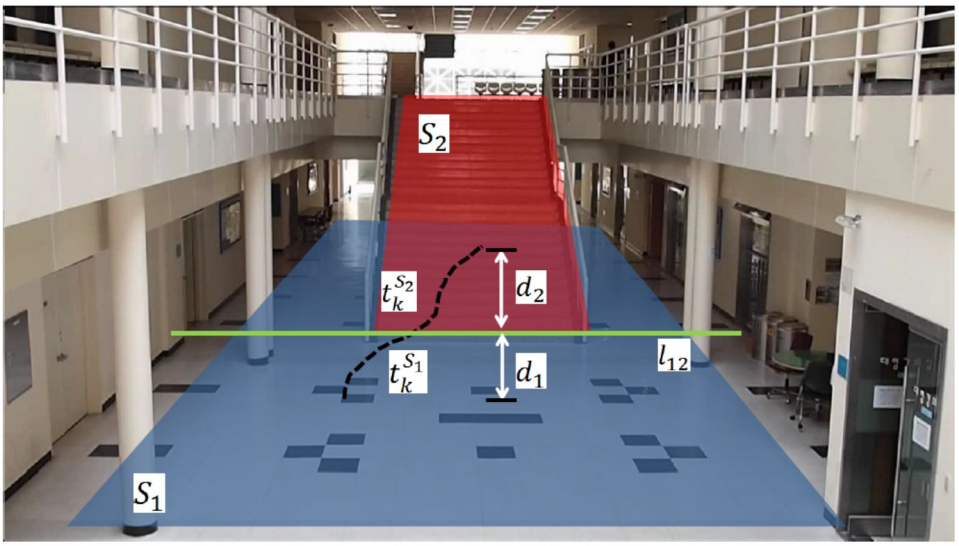


Figure 6 : Example of relative depth normalization. The depth interval d_1 and d_2 are estimated by the trajectory segments $t_k^{S_1}$ and $t_k^{S_2}$ from the boundary line l_{12} to ground planes S_1 and S_2.

2.3 Relative Depth Normalization

In Sec 2.1 and 2.2, multiple ground planes have been estimated in a coarse manner. However, we cannot compare the depth of objects in different planes since we only know the relative depth of each plane. To understand the whole 3D scene, it is necessary to normalize the relative depth scales of estimated ground planes. We need to find the same depth interval in each plane to normalize the depth scale of ground planes. We first assume that the average speed of moving objects in the ground planes are equal to each other. For the plane S_i and S_j, we find the trajectory set $\{t_k\}_{k=1,\dots,K}$ that passes the boundary l_{ij} of the planes. To find the same depth interval, we calculate the average depth distance d_n in n-th plane using

$$d_n = \frac{1}{K}\sum_{k=1}^{K} dist\left(t_k^{S_n}\right),$$

where $t_k^{S_n}$ means trajectory segments from the boundary to the plane S_n in unit frame duration, and the function $dist$ returns the depth distance of the trajectory. Using the distances d_i and d_j, which mean the same depth interval in ground plane S_i and S_j, we can normalize the relative depth of ground planes. The depth interval and normalized ground planes are illustrated in Figure 6. For the case of more than two ground planes, this normalization process is repeated.

3 EXPERIMENTAL RESULTS

For the experiments of our proposed method, we have set the following parameters. The trajectory sampling rate η was set to 30, the voting histogram bin size τ was set to 10, and the mean-shift kernel window size c was set to 10. The back projection range δ was set to 20, the active region voting threshold γ was set to 5, and the active region image cell size m was set to 30.

We performed our experiments with three video sequences. Since the appropriate dataset to evaluate our method does not exist, we took the video sequences of multiple ground plane scenes with a monocular camera. The **SNU301** is an indoor scene and has two ground planes of a floor surface and a stairway surface. The **YJC** is an outdoor scene and has two ground planes of a floor surface and an inclined surface. The **DDHS** is an outdoor scene and has tree ground planes of a floor surface, a lower slopes, and a high slopes. The scenes of video sequences **YJC** and **DDHS** do not have strong edges or corners, so the edge detection based methods [9, 3, 4] cannot be applied to the scenes. Any tracking methods can be applied to our algorithm. The numbers of tracked objects are 246 (in 8 min) in **SNU301**, 164 (in 13min) in **YJC**, and 905 (in 22 min) in **DDHS**.

For quantitative evaluation of our method, we computed the accuracy of ground plane active regions and mean error of estimated relative depth in the scenes. We denote the estimated active region as $R = R_T + R_F$, where R_T

means correctly estimated region, and R_F means the falsely estimated region and the ground truth active region as GT. The ground truth region was annotated by human hand as shown in Figure 9. Then, we calculated the precision A_P and recall A_R of the ground plane active region as $A_P = \frac{area(R_T)}{area(GT)}(\%)$ and $A_R = \frac{area(R_T)}{area(R)}(\%)$. For the ground truth of relative depth, we denoted 10 points in the actual scene with same depth interval and calculated the relative depth interval d_i as shown in Figure 7. To compute the error of relative depths, we assumed d_1 as criteria interval, and defined the mean error of relative depth as $m(e_d) = \frac{1}{N-1}$ (in this experiments, N=9).

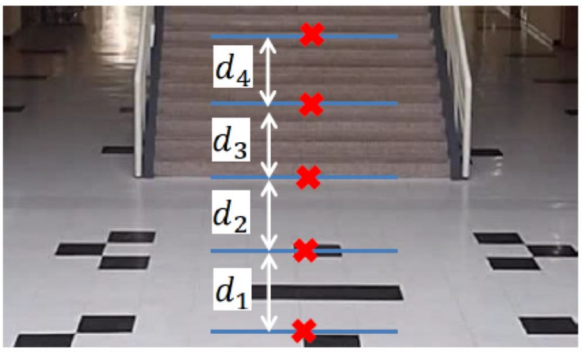


Figure 7 : The example of depth intervals for evaluation of relative depth error. The red-cross makers are the ground truth points with same depth interval and di means estimated depth interval. d_1 and d_2 are relative depths in the floor surface and d_3 and d_4 lie in the stairway surface.

The quantitative results are shown on Table 1, and the qualitative results are shown in Figure 9. To represent estimated ground planes, we drew the grids toward the horizon with same relative depth interval and tinted the active regions of multiple ground planes as shown in Figure 9. The results show that our algorithm exactly estimates the number of ground planes in the scenes. We evaluated the proposed method in two parts of active region estimation and relative depth estimation.

	GP #	A_P	A_R	$m(e_d)$
SNU301	1	75%	67%	0.28
	2	78%	80%	
YJC	1	69%	73%	0.32
	2	61%	67%	
DDHS	1	79%	84%	0.27
	2	93%	48%	
	3	50%	70%	

Table 1 : Accuracy of the estimated ground plane(GP) active region (precision : AP , recall : AR) and mean error of the estimated relative depths.

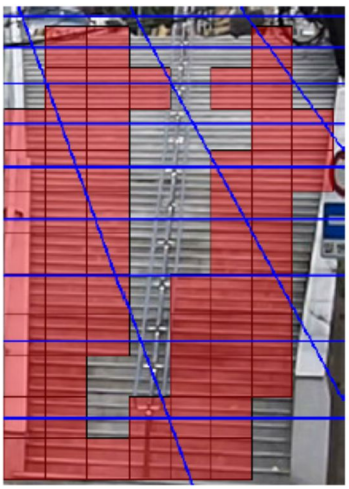


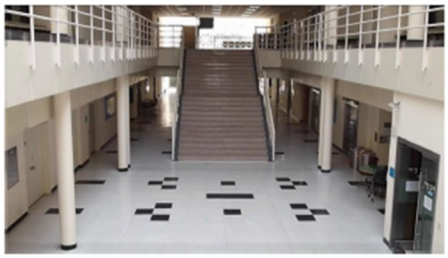
Figure 8 : Estimated active regions of ground plane #1 of DDHS. The handrails regions are not estimated as active region because objects rarely pass through this regions.

As shown in Figure 9, the estimated active regions are well separated in the scenes. Since we use back-projection approach to estimate the active regions of multiple ground planes, the regions that objects rarely pass through are not detected.

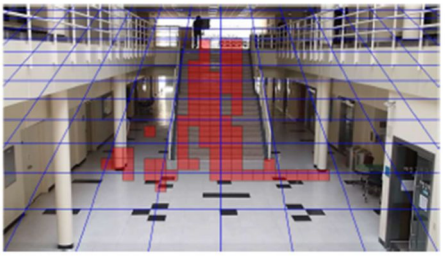
The example of this is the region of stairway handrail in Figure 8. As shown in Table 1, our algorithm performs well in the experiments. **SNU301** dataset has accurate tracking results and less variation of tracking bounding box due to influence of shadows compared to the other datasets. Therefore the extracted vanishing points are accurate and the results of estimate active regions are accurate. But ground plane #2 of **YJC** and ground plane #3 of

DDHS have relatively low precision values. These two ground planes are floor surface and closer to the camera than other ground planes. The variation of bounding boxes of tracked objects is large in the area close to the camera. Therefore the vanishing points in this area are inaccurately extracted, so these regions are not detected. But as shown in Figure 9 (g) and (l), a little farther parts of these regions are correctly detected as active region, therefore these estimated active regions still represent the ground planes well. The result of ground plane #2 of **DDHS** show that the precision value is high but the recall value is low. Since the number of trajectories pass through the ground plane #2 of **DDHS** is very large compared to the number of trajectories of other ground planes, the estimated active regions is likely to invade the area of the other planes.

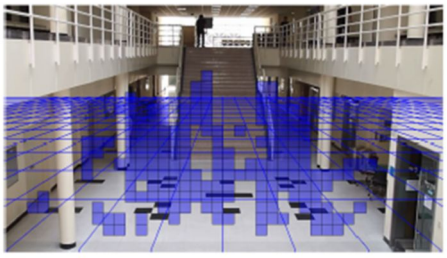
The proposed method robustly estimates relative depths as shown in Table 1. This means our algorithm can accurately estimate the position of the multiple horizons and normalize the relative depths of the multiple ground planes. With our normalized relative depths, we can robustly infer the 3D depth information of multiple ground planes as shown in Figure 9.



(a) Original Image



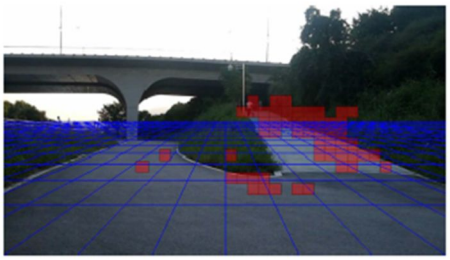
(b) Estimated GP #1



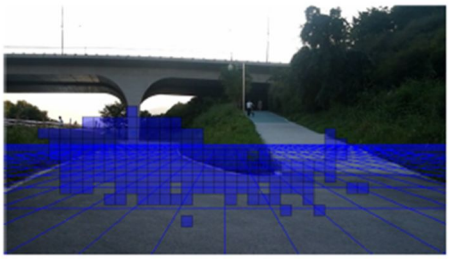
(c) Estimated GP #2

(d) Ground Truth

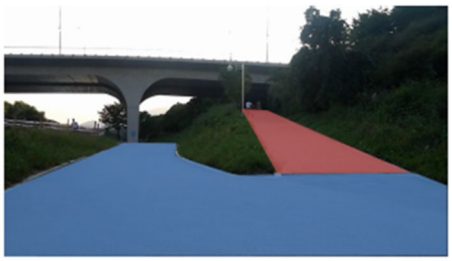
(e) Original Image



(f) Estimated GP #1



(g) Estimated GP #2



(h) Ground Truth

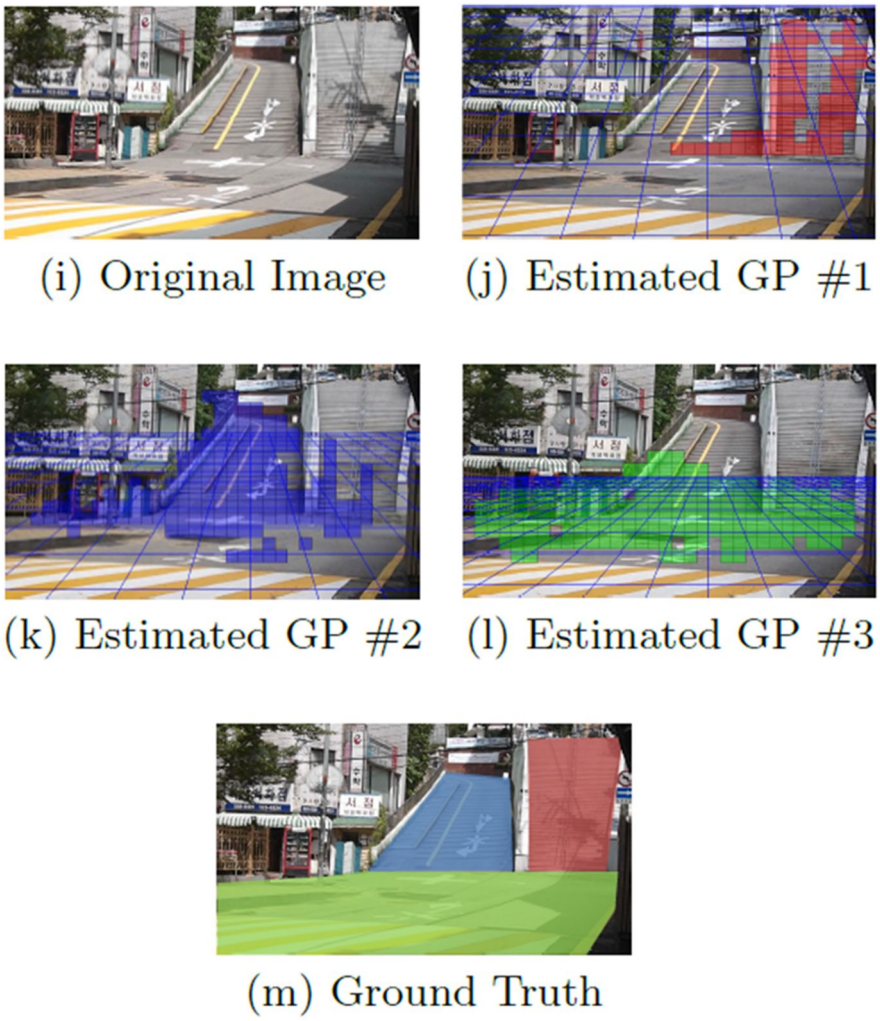


Figure 9 : Experimental results of estimated ground planes(GPs) . The first row is the result of SNU301 dataset, the second row is the result of YJC dataset, and the third row is the result of DDHS dataset.

4 CONCLUSION

We presented a method to estimate multiple ground planes with a video sequence from a monocular camera. Traditional methods for estimating ground plane are restricted to a scene with single ground plane. We extended the ideas of the previous work to multiple ground planes through a method consisted of three steps. First, to know the number of ground planes in the scene automatically, we used voting histogram of horizon candidates with vanishing points obtained by tracking results and estimated multiple horizons using mean-shift clustering. Second, the active regions corresponding to multiple ground planes are recovered through the back-projection algorithm with the trajectories of moving objects. Finally, to normalize the relative depths of multiple ground planes, To normalize the relative depths, we took advantage of the idea that the average speeds of moving objects in the ground planes should be equal to each other. The proposed method was demonstrated through experiments on three video sequences including multiple ground plane. The experimental results showed that the proposed method performs well both quantitatively and qualitatively.

REFERENCES

[1] M. D. Breitenstein, E. Sommerlade, B. Leibe, L. van Gool, and I. Reid. Probabilistic parameter selection for learning scene structure from video. In Proceedings of the 19th British Machine Vision Conference (BMVC'08), September 2008.

[2] Y. Cheng. Mean shift, mode seeking, and clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 17(8):790 - 799, aug 1995.

[3] R. Cipolla, T. Drummond, and D. Robertson. Camera calibration from vanishing points in images of architectural scenes, 1999.

[4] E. Delage, H. Lee, and A. Ng. A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 2, pages 2418 - 2428, 2006.

[5] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoer, K. Kavukcuoglu, U. Muller, and LeCun. Learning long-range vision for autonomous o-road driving. J. Field Robotics, 26:120 - 144, 2009.

[6] R. I. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[7] I. Junejo and H. Foroosh. Robust auto-calibration from pedestrians. In Video and Signal Based Surveillance, 2006. AVSS '06. IEEE International Conference on, page 92, nov. 2006.

[8] N. Krahnstoever and P. Mendonca. Bayesian autocalibration for surveillance. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, volume 2, pages 1858 - 1865 Vol. 2, oct. 2005.

[9] D. C. Lee, M. Hebert, and T. Kanade. Geometric reasoning for single image structure recovery. In In proc. CVPR, 2009.

[10] G. Lian, J. Lai, and Y. Gao. People consistent labeling between uncalibrated cameras without planar ground assumption. In Image Processing (ICIP), 2010 17th IEEE International Conference on, pages 733 - 736, sept. 2010.

[11] D. Rother, K. Patwardhan, and G. Sapiro. What can casual walkers tell us about a 3d scene? In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1 - 8, oct. 2007.

[12] A. Saxena, M. Sun, and A. Ng. Make3d: Learning 3d scene structure from a single still image. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31(5):824 - 840, may 2009.

초록

본 학위 논문에서는 고정된 단일 카메라로부터 다중 지평면을 추정하기 위한 방법을 제안한다. 다중 지평면을 추정하는 방법은 크게 세 가지 단계로 구성되어 있다. 첫째로, 지평면의 수를 추정하기 위해 소실점을 이용하여 voting histogram 을 생성하고 이 histogram 에 mean-shift clustering 방식을 적용한다. 둘째로, 다중 지평면의 실제 영역을 복원해내기 위하여 back-projection 방식을 voting histogram 에 적용하여 각각의 지평면을 지지하는 물체의 이동 궤적을 추출해낸다. 이러한 지지 궤적들을 이용하여 영역 voting 방식으로 실제 영역을 복원한다. 마지막으로 우리는 다중 지평면의 상대 깊이 정보를 표준화하기 위한 방법을 제안한다. 실험으로 본 방법이 다중 지평면 상황에 대하여 성공적으로 실제 지평면 영역들을 분리하고 깊이 정보를 추정하는 것을 확인하였다.

주요어 : 3D 장면 이해, 다중 지평면 추정, 단일 카메라

학 번 : 2011 – 20888