Multiresolution and Wide-Scope Depth Estimation Using a Dual-PTZ-Camera System

Dingrui Wan and Jie Zhou, *Senior Member, IEEE*

***Abstract*—Depth information is a critical cue for scene understanding which is very important for video surveillance. Traditional depth estimation approaches based on stereo vision have been well studied in past decades. However, little research is conducted with active cameras. In this correspondence, we discuss the depth estimation problem by employing dual-PTZ (pan/tilt/zoom)-camera system. The contributions of this correspondence includes the following three aspects: 1) we analyze the effect on depth's precision with different PTZ settings; 2) we propose a coarse-to-fine framework to deal with depth estimation problem for complex region; 3) we offer a method to generate mosaic of depth maps to combine depth information from different visual angles and resolutions for large scope. These contributions will widen the usage of dual-PTZ-camera systems to a greater extent. Real-data experimental results show that the proposed approach works well.**

***Index Terms*—Depth estimation, dual-PTZ-camera system, stereo vision.**

I. Introduction

SCENE understanding has always been an important, yet difficult, research problem in computer vision, especially for complex scenes. It is well known that depth information is a good cue to help dealing with the scene understanding problem. In this correspondence, we will discuss the depth map estimation problem by employing a dual-PTZ(pan/tilt/zoom)-camera system, including how to obtain depth information of large visual scope and how to obtain depth information with different resolution.

The PTZ camera is more and more widely used in both research and real applications. In our study, we used two PTZ cameras to constitute a visual system. Since PTZ camera is capable of obtaining multi-visual-angle and multiresolution image information, the flexibility of dual-PTZ-camera system is greater than traditional dual-static-camera system [1]. This system can be used for large area surveillance and cooperative active object tracking with high resolution at the same time. In our study, we use these characteristics to handle the high-resolution depth map and wide-scope depth map estimation problem, which might highly elevate the potential application value of this system.

In this correspondence, we use the dual-PTZ-camera system to get depth information for scene understanding by utilizing the alterability of PTZ parameters, and that can be regarded as an active vision system [2]. In the past decade, many dual-camera active vision systems have been put forward in literature. CeDAR, the Cable-Drive Active-Vision Robot [3], [4] that incorporates a common tilt axis and two pan axes, is regarded as one of the simplest active stereo system. FOVEA, the FOveated VErgent Active Stereo System [5], used the foveated cameras to achieve 3-D scene recovery. Scassellati [6] applied a binocular foveated active vision system for Cog project at the MIT Artificial Intelligence Laboratory which focused on the construction of an upper torso humanoid robot. Park and Lee [7] also proposed a so-called PSI-II system to get panoramic depth image from the panoramic image pairs constructed by a sequence of overlapping image pairs.

System calibration and control strategy are always two of the most common difficulties in these systems, and different systems might have very different solutions. In our study, the used PTZ camera has three DOF: pan, tilt and zoom, which can be acquired from the camera. The calibration of the dual-PTZ-camera system is achieved by using a spherical stereo model to convert it into an equivalent traditional dual-stereo-vision system. The control strategy is task-oriented, which is designed according to the property of cameras mentioned above to obtain high-resolution depth information and wide-scope depth information.

The target of our research is to obtain high-resolution depth map and wide-scope depth map for scene understanding by varying the PTZ parameters. To our knowledge, there is little research in the literature about depth estimation using a dual-PTZ-camera system. In our study, we proposed a coarse-to-fine iterative framework to get a depth map with a higher resolution in each iteration, and realized a mosaicing approach to generate a panoramic depth map that allows dynamic updating. The difficulty of the first application is how to design a control strategy, to make the two cameras move to suitable positions to obtain high-resolution depth information by using the relationship between PTZ parameters and corresponding depth precision. For the second application, as the used camera has limited field of view, a depth map of a wide scene cannot be calculated at once. We use several depth maps to construct a panorama. In this problem, we should consider how to minimize the stitching error caused by the uncertainties of PTZ parameters, and depth fusing error caused by a so called "disparity drift" problem.

II. Stereo Model of Dual-PTZ-Camera System

A. Single PTZ Camera Model

The pin-hole camera model is used in our study

$$\tilde{x}_n = K_0^{-1}\tilde{x} = \kappa Z(z)R(p,t)X \tag{1}$$

where x and X are image coordinates and world coordinates, respectively; x_n is the normalized image coordinates. Symbol $\sim$ means homogeneous coordinates. p, t, and z are the PTZ parameters supplied by camera. κ is a scaling factor. K_0 is a 3×3 matrix to translate image origin to principal point (u_0, v_0). For simplicity, we assume the following.

1) Pixel aspect ratio is 1, skew is 0, and no radial distortion.
2) The rotation axes of pan and tilt are orthogonal and intersect at one point, which is chosen as the center of projection and origin of the world coordinate system. So no translation factor is considered.
3) Principal point (u_0, v_0) is replaced by the zoom center [8]. This approximation has been used in many studies [9], [10].

$R(p,t)$ is a rotation matrix determined by the 'pan' and 'tilt' angles. $Z(z) = \mathrm{diag}\{f(z), f(z), 1\}$ is a scaling matrix determined by 'zoom' parameter. We denote $\theta = [p, t, 1/f(z)]^T$ as the PTZ parameter vector. In our study, we use Sony EVI D70 cameras, for which a detailed calibration procedure can be found in [1].

B. Spherical Stereo Model

In the ideal dual-camera stereo model, two cameras have the same focal length, and the optical axes are parallel and perpendicular to the baseline. However, this model does not hold well for many real systems. Stereo rectification is a way to convert these systems into the ideal ones [11], which will greatly reduce the searching scope of stereo

Manuscript received August 01, 2008; revised November 19, 2008. Current version published February 11, 2009. This work was supported by the Natural Science Foundation of China under Grant 60573062, 60673106, and 60721003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Peyman Milanfar.

The authors are with the Department of Automation, TNList, Tsinghua University, Beijing 100084, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2008.2011178

(a) (b) (c) (d)

Fig. 1. Example of stereo computing [1]: (a) and (b) are the original gray level images with different PTZ parameters; (c) is the feature points matching result; (d) is the estimated disparity map (using region based SSD algorithm), the gray level indicates the relative disparity respected to the minimal disparity.

matching. In our study, we adopt the spherical rectification method proposed in our previous work [12].

1) Spherical Rectification: Assume $x_o(u_o, v_o)$ is a point in original image, we denote its spherical rectification coordinates by (α_x, γ_x), and the corresponding rectified image coordinates by $x_r(u_r, v_r)$.

Here we only introduce some conclusions which are necessary for this study. Assume a point X in the scene has two observations in the two images, x_o^1 and x_o^2. Then the u_r-component of $x_r^i (i = 1, 2)$ is $u_r^i = \left(\gamma_x^i - \gamma_{\min}^i\right)/\lambda$, where $\gamma_{\min}^i$ is a constant translation summand for each image, and λ is a constant scale factor for the given image pair, which determines the unit length in γ-coordinate corresponding to 1 pixel in rectified image (we change the symbol $\Delta\gamma_u$ used in [1] into λ to avoid misunderstanding in later discussion). Denote b as the constant baseline, the depth of point X can be computed by

$$D = \frac{b}{d_r}, \text{ and } d_r = \lambda d_{\text{im}} + \gamma_0 \tag{2}$$

where $d_r = \gamma_x^2 - \gamma_x^1$ is called the *real disparity*; $d_{\text{im}} = u_r^2 - u_r^1$ is called the *image disparity*; $\gamma_0 = \gamma_{\min}^2 - \gamma_{\min}^1$ is a constant parameter for a given image pair. More details can be found in [1].

2) Stereo Matching: After rectification, the stereo computation becomes traditional stereo matching problem. The state of the art of stereo matching is mature and in-depth [13]. Fig. 1 shows an example of spherical rectification and disparity map estimation for two images with quite different PTZ settings. In our study, the graph cuts (expansion-move [14]) based stereo matching algorithm is used which is available in http://vision.middlebury.edu/MRF/.

C. Precision of Depth

Before discussing the precision of depth estimation, we assume: 1) both the PTZ camera model and spherical stereo model fit the dual-PTZ-camera system well; 2) the uncertainties of acquired PTZ parameters are neglectable; 3) the adopted stereo matching algorithm could achieve pixel precision, so the uncertainty of image disparity $\delta(d_{\text{im}}) = 0.5$ (pixel). Then, according to (2), the uncertainty of real disparity $\delta(d_r) = 0.5\lambda$, and the uncertainty of depth is

$$\delta(D) \doteq \frac{D^2\lambda}{b}. \tag{3}$$

λ is a rectification parameter, which directly relates to the resolution of two source images. Generally speaking, the higher the zoom level, the smaller the λ. Equation (3) reveals that the uncertainty of depth is proportional to D^2, and raising the zoom level (so λ decreases) could effectively reduce the depth uncertainty.

D. PTZ Parameters Refinement for Single Camera

Many PTZ cameras can supply instantaneous PTZ parameters but always with errors. These errors are mainly caused by two reasons: one is mechanical clearance. Take the used SONY EVI D70 camera for example, the maximum of pan and tilt errors are about 0.2° and 0.15°, respectively. The other one is the asynchrony between image capturing and PTZ parameters acquiring, especially when camera is moving.

Let I_1 be the target image whose observed PTZ parameters are denoted by a 3×1 vector θ_1. We only have several calibration images from the same camera for this refinement. Denote x_1 as a feature point in I_1, we use the SIFT descriptor [15] to find the points matching x_1, which are denoted by $\{x_i\}$ $(i = 2, 3, \ldots, m)$. The calibration image which contains x_i is labeled by I_i whose acquired PTZ parameter vector is denoted by θ_i. Since each θ_i has some unknown uncertainties as well, we need some constraints to solve this problem.

Assume θ_i^0 is the groundtruth for image I_i, where $i = 1, 2, \ldots, m$, and $\Delta\theta_i (= \theta_i - \theta_i^0))$ is independent and identically distributed (i.i.d.) with a zero-mean Gaussian distribution. Let X_i and $\tilde{x}_{n,i}$ be the world coordinates and homogeneous normalized image coordinates of x_i, respectively, and X_0 be the ground-truth world coordinates corresponding to x_i. Our goal is to estimate X_0 by minimizing $E = \sum_i \|\Delta\theta_i\|^2$.

According to camera model in (1), we have

$$X_i = \kappa_i M(\theta_i)\tilde{x}_{n,i}, \text{ and } X_0 = \kappa_i' M(\theta_i^0)\tilde{x}_{n,i} \tag{4}$$

where $M(\theta) = R(p,t)^{-1}Z(z)^{-1}$, κ_i and κ_i' are the two scaling factors. Since $\|\Delta\theta_i\|$ is very small, we approximately deem that $\kappa_i \doteq \kappa_i'$. We compute the 1-order Taylor expansion of X_i at θ_i^0

$$X_i \doteq \kappa_i M(\theta_i^0)\tilde{x}_{n,i} + \left(\frac{\partial\left(\kappa_i M(\theta_i)\tilde{x}_{n,i}\right)}{\partial\theta_i}\right)^T \Delta\theta_i \doteq X_0 + P_i\Delta\theta_i. \tag{5}$$

P_i is a 3×3 matrix and always reversible, so

$$\Delta\theta_i = P_i^{-1}(X_i - X_0). \tag{6}$$

The minimization objective can be rewritten as $E = \sum_i (X_i - X_0)^T B_i (X_i - X_0)$, where $B_i = \left(P_i^{-1}\right)^T P_i^{-1}$. We set the partial derivative of E with respect to X_0 to be 0, i.e.,

$$\frac{\partial E}{\partial X_0} = -2\sum_i B_i (X_i - X_0) = 0. \tag{7}$$

Then, $X_0 = \left(\sum_i B_i\right)^{-1} \sum_i (B_i X_i)$, which can be regarded as a weighted average of X_i. According to (6), $\Delta\theta_1$ is computable. In order to improve the precision and robustness of $\Delta\theta_1$, for all the feature points $x_1^j \in I_1$, we follow the same procedure to estimate X_0^j. We use the weighted least-squared method to calculate $\Delta\theta_1$. Denote m_j as the number of points matched with x_1^j (including itself), the larger the m_j, the higher the weight. Finally, the refined θ_1^0 can be obtained.

III. Multiresolution Depth Estimation

In scene understanding, generally speaking, depth image of the scene with a higher resolution is always more useful. However, to obtain a depth map with uniformly high resolution always needs huge computation [16] and complicate operation such as camera movement control, especially for large visual scope. So, in some applications, to obtain multiresolution depth image could be helpful. For example, regions with large depth variations should have high depth resolution, but for regions with smooth depth change, using high-resolution depth estimation might be less necessary and inconvenient.

In our study, we use a master-slave mode: one camera firstly serves as a panorama with large visual field. We call it *master* and denote it by "Cam-1." The other camera, which is called *slave* and denoted by "Cam-2," is completely controlled by the master. Our system deals with this problem: when a region of interest is selected in master-camera view, the goal is to estimate the depth map of this region with as high precision as possible.

One way to achieve high-resolution depth is to enlarge the baseline [16]. However, since the dual-PTZ-camera system has fixed baseline,

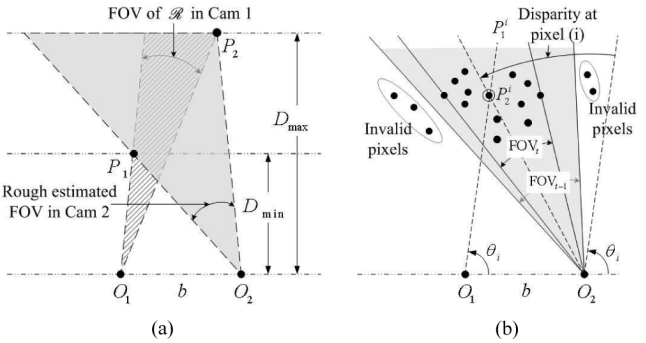


Fig. 2. Sketch map of FOV estimation for Cam-2. (a) Rough estimation by using the maximum and minimum depth. (b) Precise estimation by using estimated depth at all pixels.

we use another way, i.e., increasing the zoom levels. For a nonautomatic system, two cameras can be manually controlled to maximize the zoom levels with maintaining the selected region entirely visible, and then, the estimated depth map meets the requirement. However, for an automatic application, we need an active depth estimation framework.

We propose a coarse-to-fine active depth estimation approach to deal with this problem. Denote I_t^i ($i = 1, 2$) as the images with timestamp t from two cameras; $\mathcal{R}$ as the target region in the scene, and $\mathcal{R}^i$ is the corresponding region in I_t^i. Since $\mathcal{R}$ is selected in Cam-1, $\mathcal{R}^1$ is known already. According to the camera model, the exact PTZ parameters can be calculated to ensure $\mathcal{R}$ will be entirely in FOV with a maximal image resolution. However, the slave (Cam-2) has no information about whether $\mathcal{R}$ is in its FOV. If the depth range of $\mathcal{R}$ is known, $\mathcal{R}^2$ can be calculated by triangulation. If this depth information is accurate enough, Cam-2 could also focus on $\mathcal{R}$ with the highest resolution as Cam-1. Then, a precise depth map can be estimated from the two high-zoom images. We use a rough depth range of $\mathcal{R}$, e.g., $[2m, \infty)$, to start the coarse-to-fine procedure. The synoptic procedure is shown as follows.

1. $t = 0$. Choose $\mathcal{R}$ in panoramic view of Cam-1 $\left(I_t^1\right)$.
2. Estimate $\mathcal{R}^2$ from the rough depth range and $\mathcal{R}^1$ [Fig. 2(a)].
3. $t = t + 1$. Calculate new PTZ parameters (PTZ_t^i, $i = 1, 2$) for two cameras, and move them to these new PTZ locations.
4. For image pair I_t^1 and I_t^2, refine the corresponding PTZ parameters, and estimate the depth map of target region.
5. If stop conditions match, return; otherwise, recalculate $\mathcal{R}^2$ by the estimated depth information and goto step 3.

We mainly discuss the following problems in this procedure.

1) Determining $\mathcal{R}^2$: If only a rough depth range is available, we use the maximum and minimum depth to determine $\mathcal{R}^2$, see Fig. 2(a). If a depth (disparity) map of $\mathcal{R}$ has been estimated, we use it to calculate $\mathcal{R}^2$ in rectified image of Cam-2 first. For each pixel in rectified image of $\mathcal{R}^1$, we find the corresponding location in rectified image of I_t^2 according to its disparity, and remove this point if it is out of the FOV. Then, we find the minimal FOV as $\mathcal{R}^2$ which should ensure all pixels be visible. A sketch map is shown in Fig. 2(b). Finally, we increase by 10% on the size of $\mathcal{R}^2$ as allowance.

2) Choosing Z_t^1 and Z_t^2: In step 3, we make three rules to determine the zoom parameter of Cam-2: 1) zoom value should be no smaller than previous one; 2) it should be as large as possible and ensure $\mathcal{R}$ will be visible after the camera moves to new PTZ position. For Cam-i($i = 1, 2$), we denote $Z_{\max}^i$ as the maximal zoom level to keep $\mathcal{R}^i$ entirely visible. Since image scale ratio has fixed relation with zoom level, we precalculate a lookup table to save this information. According to the size of the bounding box of $\mathcal{R}^i$ and current zoom level, $Z_{\max}^i$ can be obtained by querying this table. 3) If we need more images for PTZ refinement, we may use a smooth factor ($s \in [0, 1]$) to slow down the growth rate of zoom level, i.e., $Z_t^2 = sZ_{\max}^2 + (1-s)Z_{t-1}^2$. So, I_t^2 and I_{t-1}^2 will have large overlapping FOV for PTZ refinement. However, the disadvantage is that smaller s might lead to more iterations.

For Cam-1, since the exact $\mathcal{R}^1$ is known, $Z_{\max}^1$ is the ideal optimal value. However, it is unnecessary to make Cam-1 move to $Z_{\max}^1$ in the first iteration, because: 1) the precision of depth estimation is decided by $\min\left\{Z_t^1, Z_t^2\right\}$, too large Z_t^1 will not help improving the precision; 2) the slower increase of Z_t^1 is helpful for PTZ refinement. We also use a smooth factor 0.5 to slow down the growth rate, i.e., $Z_t^1 = 0.5Z_{\max}^1 + 0.5Z_{t-1}^1$. However, Z_t^1 should not become a bottleneck of the precision of depth estimation, when $Z_t^1 < Z_t^2$, we set $Z_t^1 = \min\left\{Z_{\max}^1, Z_t^2\right\}$.

3) Stop Conditions: We use the uncertainty of depth to represent the precision of estimated depth map. Since $\delta(D) \propto \lambda$ (3), if we want $\delta(D) < \epsilon_D$, we only need to have λ less than a threshold $Th(\lambda)$, i.e.,

$$\lambda < \frac{\epsilon_D b}{\bar{D}^2} = Th(\lambda)$$

where b is the baseline, and $\bar{D}$ is the average estimated depth of $\mathcal{R}$. In each iteration, λ will be computed in spherical rectification, so we design two stop conditions for iteration: 1) all pixels in $\mathcal{R}$ have valid depth values, and the precision meets the demand, i.e., current rectification parameter $\lambda < Th(\lambda)$; 2) λ does not decrease. If one of the two conditions matches, the iteration will stop.

IV. Mosaic of Depth Map

Since many prevalent cameras have the conflict between FOV and image resolution, to obtain high-resolution depth information always sacrifices the scope of FOV. In our study, we proposed an approach to deal with the depth map mosaicing problem by fusing several depth maps with different resolution and visual angles, so that the final mosaiced depth image will cover large visual scope and maintain those high-resolution depth information. This technique could be very helpful for large scene understanding.

There are mainly two ways to get a panoramic depth map. One is to mosaic image for each camera first, and then do stereo matching with the mosaiced image pair to estimate a panoramic depth map [7]. The other way is to firstly calculate depth map of each image pair, and then mosaic all the depth maps to construct the panoramic depth map. Compared with the former one, this way has the following advantages: 1) multiresolution depth information can be maintained; 2) the depth value at each pixel is calculated by fusing several depth values, so the robustness and reliability of panoramic depth map will be improved; 3) it allows dynamically updating by adding new image pairs, and this mechanism is useful for scene understanding. Although this way needs more computation, for scene understanding, the quality of depth information is much more important than the time consumed, so we choose the second way in our study.

Assume there are N image pairs with estimated depth map, and we denote I_D^i as the the depth map of the ith sample. In order to get a mosaic depth map, we need a reference panoramic image coordinate system. For simplicity, we choose the image coordinate system of one reference image of one sample whose pan and tilt parameters are closest to the center of total pan and tilt ranges of all samples, and choose the average zoom level of all samples as that of the panoramic depth map. Each depth map is able to be mapped into the panorama by using the corresponding PTZ parameters.

Before fusing the depth maps, we should address the ***disparity drift*** problem, i.e., depth values from different depth maps might not be comparable, because the rectified image might have horizontal displacement error caused by the uncertainties in PTZ parameters, and consequently, the calculated disparity could be different from the real value. The PTZ refinement can only relieve this problem to some extent. We

need another technique to solve it. Because the uncertainties of the acquired PTZ parameters are very small, the disparity drifts of a given image pair can be approximately regarded as constant. Denote ρ_i as its disparity drift of the ith sample. For convenience, we deal with the disparity map instead of depth map. We first map the disparity map into the panorama using the refined PTZ parameters, and denote $d^i(x)$ as the calculated disparity at pixel x in the panorama. Then the real disparity should be $d_r^i(x) = d^i(x) + \rho_i$, and the corresponding depth $D(x) = b/d_r^i(x)$. We assume that the stereo matching result is reliable, so one way to estimate $\rho(= [\rho_1, \rho_2, \ldots, \rho_N]^T)$ is to minimize the sum of disparity variance at all pixels

$$E = \sum_x \left(\frac{1}{N(x)} \sum_{i \in \mathrm{Ind}(x)} \left| d_r^i(x) - \bar{d}_r(x) \right|^2 \right) \tag{8}$$

where $\mathrm{Ind}(x)$ is the indexes set of samples which contribute valid and reliable disparity values at pixel x; $N(x)$ is the total number of elements in $\mathrm{Ind}(x)$; and $\bar{d}_r(x)$ is the average disparity at x among samples in $\mathrm{Ind}(x)$. Considering outliers, we remove those samples whose $d_r^i(x)$ is distant to $\bar{d}_r(x)$. This problem can be solved in an iterative and analytic way with an initial value $\rho = 0_{N\times 1}$.

There are three points we should note the following. 1) As E is invariant when adding a constant number on each element of ρ, the estimated ρ_i has only relative meaning. In order to make the solution unique, we fix one of ρ_i in advance, e.g., $\rho_1 = 0$. Although the absolute disparity drift cannot be estimated, the depth comparability problem among different depth maps is handled. 2) Since the corresponding displacement in rectified image is $\delta(d_{\mathrm{im}}) = \rho_i/\lambda$, the higher the zoom level is, the greater $\delta(d_{\mathrm{im}})$ will be. Under the maximal zoom level, $\delta(d_{\mathrm{im}})$ is always less than 20 pixels. If the calculated ρ_i dissatisfies this constraint, we set $\rho_i = 0$. 3) When we want to calculate the depth from disparity, we need to find the global disparity drift ρ_0, so the real absolute disparity drift $\rho_i^r = \rho_i + \rho_0$. If we have a reference point with known ground-truth depth, ρ_0 can be calculated; otherwise, we assume $E[\rho_i^r] = 0$, so $\rho_0 = -1/N \sum_{i=1}^N \rho_i$.

1) Main Procedure:

1. *PTZ parameters refinement*. We use all images from the one camera to refine their PTZ parameters. The refined parameters will be used for stereo rectification and constructing the panorama.
2. *Stereo computation*. For each sample, compute the rectification parameter λ_i and disparity map I_d^i.
3. *Disparity drift estimation*. We project I_d^i into the panorama with the refined PTZ parameters, and estimate the disparity drift ρ_i.
4. *Depth value fusion*. Calculate the ith depth map $\left(I_D^i\right)$ from I_d^i and ρ_i. Considering that the depth value with smaller λ should be more reliable and accurate, we use a weighted averaging method to fuse the depth value for each pixel. For pixel x, if the ith sample contributes valid depth information $(D_i(x))$, we set a flag $\eta_i(x) = 1$; otherwise, $\eta_i(x) = 0$. Let $\lambda_{\min}(x) = \min_i \{\lambda_i | \eta_i(x) = 1\}$ be the minimal λ at pixel x, then the weight of $D_i(x)$ is defined as $w_i(x) = \lambda_{\min}(x)\eta_i(x)/\lambda_i$. The final depth $D(x) = \sum_i D_i(x) w_i(x) / \sum_i w_i(x)$.

2) Dynamic Updating: When a new image pair is captured to update the panoramic depth map, we first refine the PTZ parameters for each image, and then utilize stereo vision approach to calculate λ_{new} and new disparity map I_d^{new}. In order to estimate the disparity drift (ρ_{new}), we project I_d^{new} into the panoramic coordinate system, and find the region where both I_d^{new} and the panoramic depth map have valid value. ρ_{new} is the average of difference between current panoramic disparities (which can be calculated from depth by the equation $d_r = b/D$) and new estimated disparities in this region. For pixel x, denote $D(x)$ as the current panoramic depth value at x, and $D^{\mathrm{new}}(x)$ as the new estimated

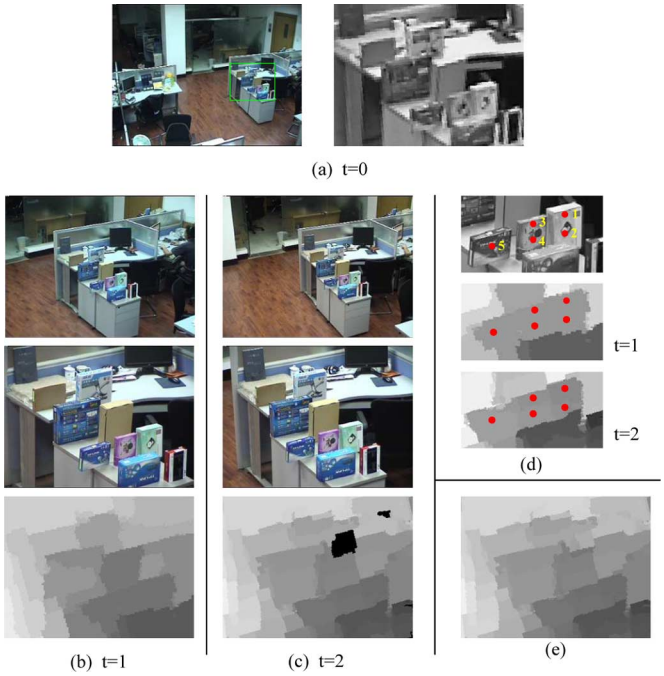


Fig. 3. Experiment of active depth estimation. (a) The reference view $(t = 0)$ and the scaled target region. (b) and (c) are the first and second iteration ($t = 1$, 2), respectively. In (b) and (c), from top to bottom, the three images are master camera image (I_t^1), slave camera image (I_t^2) and estimated depth map of target region, whose gray level linearly reveals the magnitude of depth value. (d) A detail comparison between two depth maps in a local part. (e) Final depth map.

one. We use an updating factor $\alpha(x)$ to update the new depth value into the panorama, where $\alpha(x) = 0.5\lambda_{\mathrm{new}}/\lambda_{\min}(x)$. If $\lambda_{\mathrm{new}} < 2\lambda_{\min}(x)$, we have

$$D_{\mathrm{updated}}(x) = (1 - \alpha(x))D^{\mathrm{new}}(x) + \alpha(x)D(x)$$

otherwise, the depth at x will not change, i.e., $D_{\mathrm{updated}}(x) = D(x)$, which means the resolution of new calculated depth at x is too low.

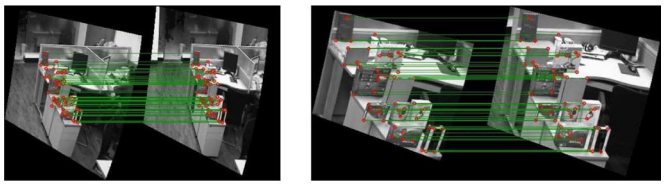
V. Experimental Results

We utilize two Sony EVI D70 cameras to compose the dual-PTZ-camera system with baseline $b = 0.78$ m. The size of captured images is 320×240.

A. High-Resolution Depth Estimation

We select a target region with stationary objects in the reference view of the master camera, and use the coarse-to-fine framework to automatically estimate the depth map of this selected region.

In this experiment, we only use two iterations to illustrate the algorithm, and the result is shown in Fig. 3. In the initialization [Fig. 3(a)], the target region is selected in master camera with zoom level $Z_0^1 = 0$, and the maximum zoom level of master camera $Z_{\max}^1 = 10.5$. In the first iteration (Fig. 3(b)), we use the given rough depth range $[2m, \infty)$ to calculate the PTZ parameters of slave camera, and $Z_1^2 = 4$ (the zoom lookup table outputs the greatest integral zoom level less than or equal to the optimal value). According to the zoom selection rule $Z_1^1 = 5$. In the second iteration [Fig. 3(c)], new PTZ parameters of slave camera are calculated by previous estimated depth information. The new zoom levels of two cameras are $Z_2^1 = Z_2^2 = 9$. Because only two iterations are used, we set the zoom of master camera to be the maximum, i.e., $Z_2^1 = \left\lfloor Z_{\max}^1 \right\rfloor = 10$.



(a) (b)

Fig. 4. Testing points with matching lines in two rectified image pairs of the two iterations: (a) $t = 1$ and (b) $t = 2$.

TABLE I
COMPARISON OF DEPTH ESTIMATION BETWEEN THE TWO ITERATIONS

	Iter-1	Iter-2
$\mathscr{E}$ (m)	0.053	0.016
Standard deviation (m)	0.071	0.021
λ	0.0033	0.0012
$\delta(D)$ (m)	0.113	0.041
Disparity drift (ρ)	-0.0031	-0.0030

In order to compare the depth distinguishability between the two estimated depth maps in two iterations, we choose five sample points with similar but increasing depths (the average neighboring discrepancy is about 4 cm), which is shown in Fig. 3(d). In order to reveal different depth value in figure, we increase the contrast of depth image. In Iter-1, all points have the same depth values, and in Iter-2, the five different depth values are detected.

We also make a quantitative comparison. We choose tens of points with known ground-truth depths, and ensure the observations of each point in the four rectified images (of I_1^i and I_2^i, $i = 1, 2$) are well matched in pixel precision, so that no stereo matching error will affect this comparison. In this experiment, 47 points are selected (see Fig. 4) whose average depth $D_0 = 5.18$ m.

We randomly choose $M = 10$ points to calculate the disparity drift (ρ) of each iteration. Denote d_i as the calculated disparity at the ith point, and d_i^0 as the ground-truth disparity ($= b/D_i^0$, where D_i^0 is the corresponding ground-truth depth), then

$$\rho = \arg\min_{\rho} \sum_{i=1}^{M} \left| d_i + \rho - d_i^0 \right|^2 \doteq \frac{1}{M} \sum_{i=1}^{M} \left(d_i^0 - d_i \right).$$

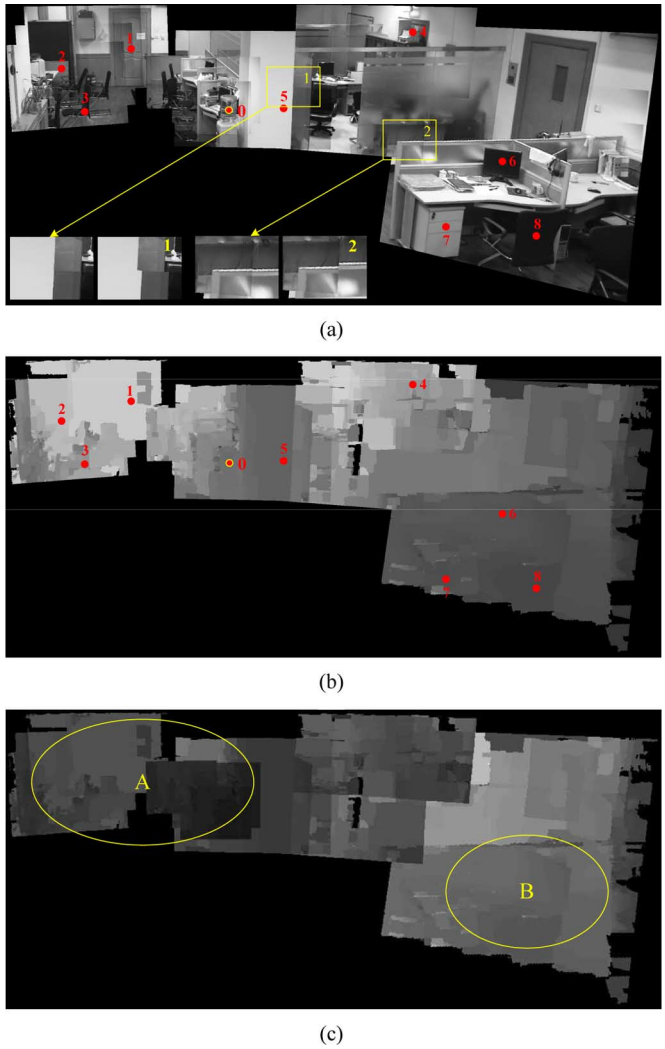
So the estimated depth $D_i = b/(d_i + \rho)$. For the rest 37 points, we denote $\mathcal{E}$ as the average absolute difference between estimated depth and the groundtruth. The depth uncertainty is $\delta(D) = D_0^2 \lambda / b$, where D_0 is the average depth ($= 5.18$ m). The result is shown in Table I.

From Table I, we can see the following. 1) Iter-2 has much smaller $\mathcal{E}$, so it has better depth precision than Iter-1. 2) When zoom level rises, λ decreases, and $\delta(D)$ decreases as well. So, Iter-2 has higher depth distinguishability. Generally speaking, later iterations always have better distinguishability and precision of depth than previous iterations. 3) The PTZ refinement is effective for the large overlapping region, so the disparity drifts in two iterations are both very small and similar.

B. Depth Map Mosaicing

In this experiment, we generated two panoramic depth maps for indoor and outdoor scene. For the indoor experiment, we use 14 image pairs captured manually and for the outdoor, we use 16 by automatic servoing.

Fig. 5 shows the indoor experimental result. The mosaic of gray level image, which is shown in Fig. 5(a), is used as a reference for the panoramic depth map, and it also reveals the effect of PTZ refinement



(a)

(b)

(c)

Fig. 5. Indoor experiment of depth map mosaicing. (a) The mosaic of gray level image by using the refined PTZ parameters for reference camera. Tow comparisons between mosaiced images with refined (left) and original (right) PTZ parameters are shown in the bottom left. (b) The mosaiced depth map whose gray level linearly reveals the magnitude of depth value. Gray level 0 indicates invalid value. (c) The relative uncertainty map (smaller gray level indicates smaller relative uncertainty).

(see the comparison in two local regions). The mosaic depth map is shown in Fig. 5(b), and the gray level linearly reveals the magnitude of depth value.

We also generate a relative uncertainty map of the mosaiced depth map to show the distribution of reliability of panoramic depth map. The relative uncertainty at each pixel is $\delta(D)/D$, where $\delta(D)$ is calculated by (3). We linearly magnify these values to improve the visual effect in Fig. 5(c). The darker intensity in image indicates more reliable. Consider region A and B. If the two regions are captured with the same image resolution, the relative uncertainty should be proportional to its depth. Because region A has larger depth than region B, the average relative uncertainty of A will be larger than that of B. In this experiment, as region A is captured with a much higher zoom level than B, the average relative uncertainty of A is contrarily much smaller than that of B. So multiresolution depth information is maintained in the mosaic depth map.

In order to quantitatively test the accuracy of depth estimation, we manually choose several points and compare the calculated depth values with the measured groundtruth. Note that as we mentioned

TABLE II
COMPARISON BETWEEN ESTIMATED DEPTH AND GROUNDTRUTH AT THE LOCATIONS LABELED IN FIG. 5(B)

Point Index	Estimated depth	groundtruth depth
1	14.71	14.91
2	13.80	14.12
3	10.58	10.62
4	11.00	10.83
5	7.36	7.60
6	6.38	6.50
7	6.03	6.11
8	5.47	5.55

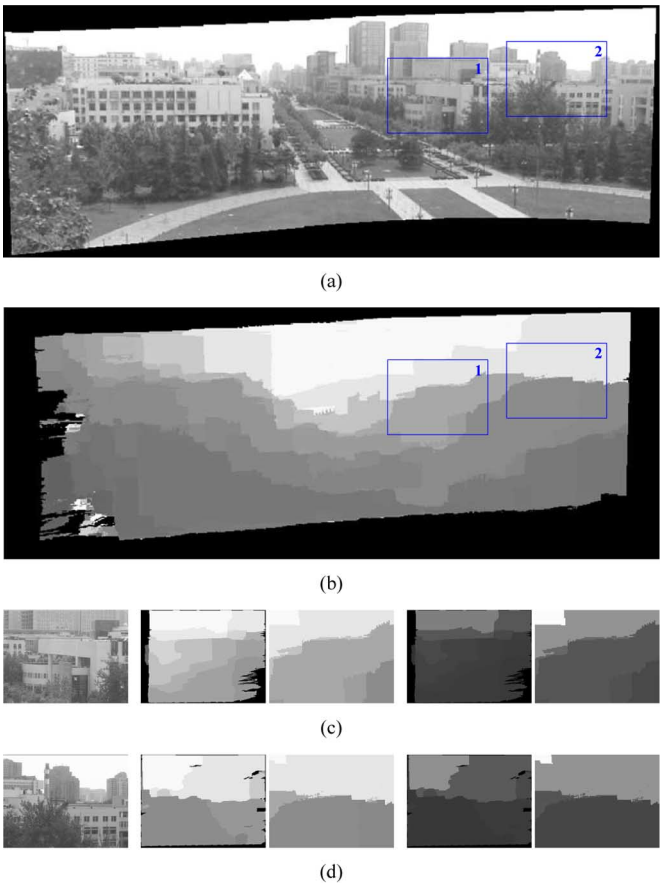


Fig. 6. Outdoor experiment of depth map mosaic: (a) the mosaic of gray level image by using the refined PTZ parameters for reference camera; (b) the mosaic of depth map whose gray level $g = 50 + 200(1 - e^{-0.004D})$, where D is the magnitude of depth value. Zero indicates invalid depth value. (c), (d) Two examples of depth map updating. From left to right, the five images are new image from reference camera, new estimated and original depth map, new and original relative uncertainty map (smaller gray level indicates smaller relative uncertainty).

before, the estimated disparity drift only deals with the relative comparability between depth maps. So we use a reference point in the scene with known depth to estimate the absolute disparity drift, i.e the point labeled "0" in Fig. 5(a). Actually, this global disparity drift is very small, and the equivalent depth discrepancy is about 0.1 m. The testing points are labeled from "1" to "8" in Fig. 5(a). The groundtruth and calculated depths of these points are shown in Table II.

Fig. 6 shows the outdoor experimental result. We also use a reference point in the scene with known depth (about 170 m) to calculate the absolute disparity drift for the panoramic disparity map, then final depth map can be obtained [Fig. 6(b)].

TABLE III
SOME MID-RESULTS OF EXPERIMENTS IN FIGS. 5 AND 6

	Indoor Exp.		Outdoor Exp.	
$\max\|\Delta Pan\|(^\circ)$ of two cameras	0.136	0.167	0.061	0.056
$\max\|\Delta Tilt\|(^\circ)$ of two cameras	0.120	0.110	0.081	0.086
$\text{mean}\|\rho_i\|$	0.0038		0.0023	

We use two sets of new data with higher zoom levels to update the panoramic depth map. The FOV of new reference images are illustrated by two rectangles in Fig. 6(a). In Fig. 6(c) and (d), from left to right, the five images are new image from reference camera, new estimated and original depth map, new and original relative uncertainty map. The relative uncertainty is also defined as $\delta(D)/D$. In this experiment, the relative uncertainties of two new estimated depth maps are about 37% and 45% of each original one, respectively. So, the depth distinguishability increases after updating.

From all above experimental results, we can see that the mosaic procedure works well and the final depth map could well reflect the depth distribution of the panoramic view.

Finally, we list the result of PTZ refinement and disparity drift estimation of the experiments in Figs. 5 and 6, see Table III. Note that the global disparity drift estimated via reference point is considered in ρ_i, and for "Exp. in Fig. 6," , the two updating samples are not counted. For the indoor experiment, the discrepancy of depth caused by the disparity drift at $D_0 = 10$ m is about $D_0^2|\rho|/b = 0.49$ m; for the outdoor, it is about 118 m when $D_0 = 200$ m. These numbers tell the importance of disparity drift estimation.

REFERENCES

[1] D. Wan and J. Zhou, "Stereo vision using two ptz cameras," *Comput. Vis. Image Understand*, vol. 112, no. 2, pp. 184–194, 2008.
[2] A. Bandopadhay, J. Alomoinos, and I. Weiss, "Active vision," *Int. J. Comput. Vis.*, vol. 2, pp. 353–366, 1988.
[3] H. Truong, S. Abdallah, S. Rougeaux, and A. Zelinsky, "A novel mechanism for stereo active vision," presented at the Australian Conf. Robotics and Automation, 2000.
[4] A. Zelinsky, A. Dankers, and N. Barnes, "Active vision—rectification and depth mapping," presented at the Australian Conf. Robotics and Automation, 2004.
[5] W. N. Klarquist and A. C. Bovik, "Fovea: A foveated vergent active stereo system for dynamic three-dimensional scene recovery," in *Proc. ICRA*, 1998, pp. 3259–3266.
[6] B. Scassellati, A Binocular, Foveated Active Vision System, 1998, Tech. Rep.
[7] S.-C. Park and S.-W. Lee, "Fast distance computation with a stereo head-eye system," *Biol. Motivated Comput. Vis.*, pp. 434–443, 2000.
[8] M. Li and J.-M. Lavest, "Some aspects of zoom lens camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 11, pp. 1105–1110, Nov. 1996.
[9] S. Sinha and M. Pollefeys, "Towards calibrating a pan-tilt-zoom cameras network," presented at the OMNIVIS, ECCV Conf. Workshop CD-Rom Proceedings, 2004.
[10] R. G. Willson, Modeling and Calibration of Automated Zoom Lenses, CMU-RI-TR, 1994, Tech. Rep.
[11] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 993–1008, Aug. 2003.
[12] D. Wan, J. Zhou, and D. Zhang, "A spherical rectification for dual-ptz-camera system," in *Proc. ICASSP*, 2007, vol. 1, pp. 777–780.
[13] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1-3, pp. 7–42, 2002.
[14] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[16] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys, "Variable baseline/resolution stereo," *CVPR*, pp. 1–8, 2008.