Dynamic Mutual Calibration and View Planning for Cooperative Mobile Robots with Panoramic Virtual Stereo Vision[*]

Zhigang Zhu
Department of Computer Science, the City College
The City University of New York, New York, NY 10031

Deepak R. Karuppiah, Edward M. Riseman and Allen R. Hanson
Department of Computer Science, University of Massachusetts, Amherst, MA 01003

Short running head: **Panoramic Virtual Stereo**

Contact Information:

Prof. Zhigang Zhu
Department of Computer Science
The City College of New York /CUNY
Convent Avenue and 138th Street, New York, NY 10031
Tel: (212) 650 - 8799 Fax: (212) 650 - 6248
Email: zhu@cs.ccny.cuny.edu
URL: http://www-cs.engr.ccny.cuny.edu/~zhu/

[*] This work was supported by DARPA/ITO programs Mobile Autonomous Robot S/W (MARS) (Contract Number DOD DABT63-99-1-0004) and Software for Distributed Robotics (SDR) (Contract Number DOD DABT63-99-1-0022). It is also partially by the Army Research Office under grant number DAAD19-99-1-0016, by Air Force Research Lab (Award No. F33615-03-1-63-83) and by the CUNY GRTI program.

Abstract

This paper presents a panoramic virtual stereo vision approach to the problem of detecting and localizing multiple moving objects (e.g., humans) in an indoor scene. Two panoramic cameras, residing on different mobile platforms, compose a virtual stereo sensor with a flexible baseline. A novel "mutual calibration" algorithm is proposed where panoramic cameras on two cooperative moving platforms are dynamically calibrated by looking at each other. A detailed numerical analysis of the error characteristics of the panoramic virtual stereo vision (mutual calibration error, stereo matching error, and triangulation error) is given to derive rules for optimal view planning. Experimental results are discussed for detecting and localizing multiple humans in motion using two cooperative robot platforms.

List of Symbols:

χ : angular resolution of the panoramic image

D_k: target's distance to camera k *(k=1,2)*

ϕ_0, ϕ_1 *and* ϕ_2 : three interior angles of the triangle formed by two cameras and the target

θ_i : bearing angle of the target in image i (i=1,2)

β_{ij} : bearing angle of the camera i in image j

B: baseline length

R: radius of the cylindrical body of the mobile robot used for mutual calibration

α : angle subtended by the cylindrical body of the robot

W : width of the target

w_i : width of the target in the panoramic image i (i=1,2)

$T_i^{(k)}$: feature set of the blob i in camera k

∂X : error in estimating parameter X (X = B, $\partial\phi_1, \partial\phi_2$ β_{21}, θ_1, θ_2, α, w and D)

∂D_1^B : distance error due to the baseline error

∂D_1^ϕ : distance error due to angular error

∂D_1^+ : distance error when $D_1 > B$

∂D_1^0 : distance error when $D_1 = B$

∂D_1^- : distance error when $D_1 < B$

∂D_1^{fix} : distance error when the baseline is fixed

∂D_1^s : distance error in the size-ratio method

1 Introduction

Flexible, reconfigurable vision systems can provide an extremely rich sensing modality for sophisticated robot platforms. We propose a cooperative and adaptive approach to the problem of finding and protecting humans in emergency circumstances, for example, during a fire in an office building. Real-time processing is essential for the dynamic and unpredictable environments in our application domain, and it is important for visual sensing to rapidly focus attention on important activity in the environment. Any room or corridor should be searched quickly to detect people and fire. Field-of-view issues using standard optics are challenging since panning a camera takes time, and multiple targets/objectives may require saccades to attend to important visual cues. Using multiple conventional cameras covering different fields of view could be a solution, but the cost of hardware (cameras, frame grabbers and computers) and software (multiple stream data manipulation) would increase. Thus, we employ panoramic cameras to detect and track multiple objects (people) in motion, in a full 360-degree view, in real time.

We note that there is a fairly large body of work on detection and tracking of humans [1-5], motivated most recently by the DARPA VSAM effort [6]. On the other hand, different kinds of omni-directional (or panoramic) imaging sensors have been designed [7-12], and a systematic theoretical analysis of omni-directional sensors has been given by Baker & Nayar [7]. Omnidirectional vision has become quite popular with many vision approaches for robot navigation [10,13,14, 30, 31], 3D reconstruction [15-17, 29] and video surveillance [18-20, 28]. Research on multiple camera networks with panoramic cameras that are devoted to human and subject tracking and identification can be found in the literature [19-23, 30, 31]. The most related work is the realtime human tracking system by Sogo et al [21] using multiple omnidirectional cameras distributed in an indoor environment. The system detects people, measures bearing angles and determine their locations by triangulation. Generally there are two problems in such a system - (1) the correspondence problem among multiple targets, and (2) the measurement accuracy of target locations. The correspondence problem is more difficult in a panoramic stereo than a conventional stereo because both large baseline and low resolution make it hard to establish correspondences of the visual features. The second problem arises particularly when a target is (almost) aligned with the stereo pair. In order to solve these problems, Sogo et al [21] proposed a “N-ocular stereo” approach without visual features that only verifies the correspondences of multiple targets of binocular stereo by a third omnidirectional camera,. They showed that the uncertainty in estimating 3D locations was reduced by using the best estimations of pairs of four fixed panoramic cameras put in the vertices of a square region. However the error of localizing a target is still proportional to the square of the target’s distance from the cameras with fixed baseline distances; in their simulation, it increases 7-fold when a target moves 3.5 meters away from the cameras. Our work differs from theirs in that we deal with panoramic stereo vision on mobile platforms and thus study the issues of dynamic calibration and view planning. We propose a novel concept of mutual calibration and give a detailed error analysis of panoramic stereo that leads to dynamic stereo configurations with adaptive baselines and viewpoints for best depth estimation. The distinctive feature of our approach is the ability to compose cooperative sensing strategies across the distributed panoramic sensors of a robot team to synthesize optimal "virtual" stereo vision for human detection and tracking.

The idea of distributing sensors and cooperation across different robots stems from the requirements of potentially limited (sensor) resources for a large robot team and the need for mobile placement of sensor platforms given the limited resolution in panoramic sensors. Nevertheless, the advantages of cooperative vision arise from more than this compromise. Any fixed-baseline stereo vision system has limited depth resolution because of the physical constraints imposed by the separation of cameras, whereas a system that combines multiple views allows the planning system to take advantage of the current context and goals in selecting viewpoints. In this paper, we focus on the cooperative behavior involving cameras that are aware of each other, residing on different mobile platforms, to compose a virtual stereo sensor with a flexible baseline. In this model, the sensor geometry can be controlled to manage the precision of the resulting virtual sensor. The cooperative stereo vision strategy is particularly effective with a pair of mobile panoramic sensors that have the potential of almost always seeing each other. Once calibrated by "looking" at each other, they can view the environment to estimate the 3D structure of the scene.

The organization of the paper is as follows. After the introduction of two depth estimation methods using panoramic sensors in Section 2, we will mainly focus on the following critical issues of a panoramic virtual stereo system:

1) Dynamic “mutual calibration” between the two cameras on two separate mobile robots that forms the dynamic "virtual" stereo sensor with a full 360-degree view (Section 3);
2) A detailed numerical analysis of the error characteristics of the panoramic virtual stereo in order to derive the rules for optimal view planning of the moving sensor platforms (Sections 4 and 5); and
3) View planning by taking advantage of the current context and goals, based on a thorough error analysis of panoramic virtual stereo (Section 6).

Experimental systems and results for multiple human detection and localization will be given in Section 7, and conclusion and future work will be discussed in the last section.

2 Panoramic Virtual Stereo Geometry

In our experiments we use the panoramic annular lens (PAL) camera system [9] as it can capture its surroundings with a field of view (FOV) of 360-degrees horizontally and -15 ~ +20 degrees vertically (Fig. 1). In the application of human tracking and identification by a mobile robot, a vertical viewing angle that spans the horizon is preferred. After image un-warping, distortion rectification, and camera calibration [22, 24], we obtain cylindrical images generated by a panoramic "virtual" camera from the virtual viewpoint on the axis of the cylindrical image surface (Fig. 1c). Other omnidirectional sensors can also be applied; for example, the omnidirectional cameras proposed by Nayar's research group [7] have been used in our current experiments.

Panoramic virtual stereo vision is formed by two panoramic cameras residing on two separate (possibly mobile) platforms. Let's assume that both of them are subject to only planar motion on the floor and are at the same height above the floor. Suppose that in Fig. 2a, O_1 and O_2 are the viewpoints of the two cameras and they can be localized by each other in the panoramic images as P_{12} and P_{21} respectively. B is the baseline (i.e. distance O_1O_2) between them. The projection of a target T is represented by T_1 and T_2 in the two panoramic images. Then a triangle O_1O_2T can be formed. By defining an arbitrary starting orientation for each cylindrical image, three angles ϕ_1, ϕ_2 and ϕ_0 of the triangle can be calculated from the following four bearing angles: θ_1 and θ_2, the bearings of the target in image 1 and image 2 respectively, β_{12} and β_{21}, the bearing angles of camera 1 in image 2, and camera 2 in image 1 respectively. Therefore the distances from the two cameras to the target can be calculated by *triangulation* as

$$D_1 = B\frac{\sin\phi_2}{\sin\phi_0} = B\frac{\sin\phi_2}{\sin(\phi_1+\phi_2)}, \quad D_2 = B\frac{\sin\phi_1}{\sin\phi_0} = B\frac{\sin\phi_1}{\sin(\phi_1+\phi_2)} \qquad (1)$$

With stationary cameras, triangulation error, i.e. the error in estimating D_1(or D_2) varies with the target location – larger errors when the target is close to the baseline and smaller errors when better triangulation is possible. Here we show first that panoramic stereo can almost always estimate the distance of the target in the full 360° view. It is commonly known that the triangulation relation in Eq.(1) approaches singularity as the target moves towards the baseline O_1O_2. Fortunately, near colinearity of the sensors and the target can be easily verified, and even then the 3D location of the target can still be estimated by using the size-ratio of the target in two panoramic images

$$D_1 = B\frac{w_2\cos\phi_2}{w_1\cos\phi_1 + w_2\cos\phi_2}\cos\phi_1, \; D_2 = B\frac{w_1\cos\phi_1}{w_1\cos\phi_1 + w_2\cos\phi_2}\cos\phi_2 \qquad (2)$$

where w_1 and w_2 are the widths of the target in the panoramic image pair. Note that the cosines in the above equations only give signs since the angles are either 0° or 180°. As an example, if the target lies between O_1 and O_2 (Fig. 2b), the distances to them can be calculated as

$$D_1 = B\frac{w_2}{w_1 + w_2},\ D_2 = B\frac{w_1}{w_1 + w_2} \tag{3}$$

In the *size-ratio method*, since the two cameras view the target (e.g., a human) from exactly the opposite direction, the widths of the objects in the two images correspond *approximately* to the same width in 3D space (Fig. 2b), which makes the estimation plausible. As an alternative, we can also use the height information (in the same way as we use width) since the height of an object is more invariant. However, it is only applicable when the top and/or bottom of the figure are visible in both of the panoramic images and can be accurately localized. In contrast, the width information is easier to extract and more robust since we can integrate the results from different heights of the object. Realizing that the object and the robots may occlude (part of) each other when in a collinear alignment, we will use the width and height information adaptively.

In panoramic virtual stereo where the viewpoint and baseline relation can change, it is interesting to find the best configuration for estimating the distance of a target. For this purpose, first a dynamic mutual calibration approach will be presented in Section 3. Then a detailed numerical analysis of the distance estimation error by the panoramic virtual stereo (with both the triangulation and size-ratio methods) will be given in Section 4, which will lead to useful results for view planning between the two mobile platforms with panoramic cameras.

3 Dynamic Mutual Calibration

In order to estimate the distance of a target, we need to first estimate the baseline and the orientation angles of the two panoramic cameras. In stereo vision, an *epipole* is defined as the projection of one camera's center in the other camera's image plane. In a stereo system with normal FOVs, epipoles are usually out of the FOVs in both cameras, therefore we must use a third target in the scene for stereo calibration. In contrast, the panoramic stereo has two "visible epipoles" because the two panoramic cameras can see each other. Here we propose a special dynamic calibration procedure called *mutual calibration* based on the visible epipole property in panoramic stereo. Mutual calibration neither needs to setup any additional calibration targets nor requires the use of a third object in the environment. Instead, each of the panoramic cameras can use the other as the calibration target. The advantage of "sensor as the target" in mutual calibration is that the geometric structures and the photometric properties of the sensors as well their platforms can be well designed and are known *a priori*.

Several practical approaches have been proposed for this purpose by using special structures, such as cylinders, vertical lines and rectangular planar surfaces [24]. The basic idea is to make the detection and calculation robust and fast. One of the approaches is to design the body of each robot as a cylinder with some vivid colors (e.g. white in the intensity images of our current implementation), which can be easily seen and extracted in the image of the other robot's camera (Fig. 3a). We assume that the rotation axis of each panoramic camera is coincident with the rotation axis of the cylindrical body of the corresponding robot, therefore the baseline between the two panoramic cameras can be estimated using the occluding boundary of either of the two cylinders, e.g., from the image of camera 2 we have

$$B = R / \sin(\frac{\alpha}{2}) \qquad (4)$$

where α is the angle between two occluding projection rays measured in the image of camera 2, and R is the radius of the 1st cylindrical body (Fig. 3b). The orientation angle (β_{12}) of the line O_2O_1 is simply the average of the bearings of two occluding boundary points P_1 and P_2. We can do the same in the image of camera 1.

Fig. 4 shows a calibration result. The cylindrical body of each robot (pointed at by an arrow in Fig. 4a and Fig. 4b) is detected and measured in the panoramic image of the other robot. In the experiment, the perimeter of the cylindrical image is 1000 pixels, so the angular resolution in degrees is 360/1000 = 0.36 ° per pixel. We define the *angular resolution* of the panoramic image as χ in radians for future use, which is 6.28 mrad/pixel in this experiment. The radius of the cylindrical body of each robot is designed as *R*=18.0 cm.

4 Error Analysis

Previous work (e.g., [21]) only gave the error distribution of the panoramic stereo with a fixed stereo geometry. Shum et al [25] studied the case of an omnidirectional camera moving within a circular region of the plane and concluded that it was the best to choose a pair of cameras that are vergent on a point with maximum vergence angle in order to accurately localize the point. In this paper we will discuss a more general case where the relations between the two panoramic cameras can change arbitrarily. Our task is to find the distance of a *given* target point from camera 1 by finding its correspondence in camera 2, so the localization error turns out to be the distance error. For this reason, we use a different error formulation: for a certain distance D_1 from camera 1 to the target, what is the error distribution of this distance with different locations of camera 2, which determines the configurations of baselines and angles of the panoramic stereo? Can we achieve a better distance estimation for distant targets with a larger baseline, which is also dynamically determined by the mutual calibration? Eq. (1) and Eq. (2) show that the accuracy of distance estimation depends on the accuracy in estimating the baseline and the bearing angles. Here we derive an analysis of the error of estimating distance D_1 from the first camera to the target. First, with the triangulation method, the estimated distance error can be computed by partial differentials of Eq. (1) as

$$\partial D_1 = \left|\frac{\sin\phi_2}{\sin(\phi_1+\phi_2)}\right|\partial B + B\left|\frac{\sin\phi_2\cos(\phi_1+\phi_2)}{\sin^2(\phi_1+\phi_2)}\right|\partial\phi_1 + B\left|\frac{\sin\phi_1}{\sin^2(\phi_1+\phi_2)}\right|\partial\phi_2$$

or

$$\partial D_1 = \frac{D_1}{B}\partial B + D_1\left|\cot(\phi_1+\phi_2)\right|\partial\phi_1 + \frac{D_2}{\sin(\phi_1+\phi_2)}\partial\phi_2 \tag{5}$$

where ∂B is the error in computing the baseline B, and $\partial\phi_1$ and $\partial\phi_2$ are the errors in estimating the angles ϕ_1 and ϕ_2 from the two panoramic images. Analyzing Eq. (5), we have found that the distance error comes from three separate error sources: mutual calibration error, stereo matching error and stereo triangulation error, which will be discussed below.

4.1 Calibration error

Dynamic mutual calibration estimates the baseline B, and the bearing angles β_{12} and β_{21} of the two cameras, all of which are subject to errors in localizing the calibration targets. The error in estimating the baseline by Eq. (4) can be derived as

$$\partial B = \frac{B\sqrt{B^2 - R^2}}{2R}\partial\alpha \le \frac{B^2}{2R}\partial\alpha \qquad (6)$$

where $R <<$ B, and $\partial\alpha$ is the error in estimating the angle α in an image. From Eq. (6) we can find that the baseline error ∂B is inversely proportional to the dimension of the cylindrical body for dynamic calibration given the same angle error $\partial\alpha$. On the other hand, given the radius R and the angle error, the baseline error is roughly proportional to the square of the baseline itself. The angle error ($\partial\alpha$) is determined by the errors in localizing the occluding boundaries of the second (or first) cylinder in the first (or second) panoramic image (Fig. 3). The errors in estimating the bearing angles β_{21} and β_{12} will introduce errors to the angles ϕ_1 and ϕ_2 of the stereo triangle (Fig. 2a). Since each bearing angle is the average of the orientations of the two occluding boundaries, their errors can be roughly modeled as the same as $\partial\alpha$, i.e. $\partial\beta_{21} = \partial\beta_{12} = \partial\alpha$. Note that these errors ($\partial B, \partial\beta_{21}, \partial\beta_{12}$) are derived from the specific mutual calibration method we are using in this paper. However, the following general relations hold: larger distance (baseline) between two cameras will introduce larger errors in estimating the baseline, but the errors in bearing angles are independent of the distance as long as the calibration targets can be detected.

4.2 Matching error

We want to find the distance of a *given* point T_1 in view 1 by finding its corresponding point T_2 in view 2. In this sense, there will be no error in providing the bearing angle θ_1 in view 1, i.e. $\partial\theta_1$=0, which implies that the error $\partial\phi_1$ is solely determined by the error of the angle β_{21} via calibration, i.e. $\partial\phi_1 = \partial\alpha + \partial\theta_1 = \partial\alpha$. However, the perspective view difference in O_1 and O_2 will introduce a stereo "matching error" (denoted as $\partial\theta_2$) in θ_2, the localization of T_1's matching point T_2, which could be a function of the location of the view point O_2 (related to O_1). Thus, $\partial\phi_2$ = $\partial\alpha +$ $\partial\theta_2$ is a (complicated) function of the viewpoint location and is generally larger than $\partial\phi_1$. Generally speaking, the "matching error" is determined by three aspects – visibility (the size of a target in the panoramic image), detectability (the contrast of the target with the background) and similarity (appearance differences between images of an object in two widely separated views). In the panoramic virtual stereo, the sizes and the appearances of a target can suffer from significant perspective distortion due to widely separated views. The matching error will be directly related to the primitives we are using for stereo matching.

4.3 Triangulation error and overall distance error

Now we want to find a numerical result of the following problem: for a certain distance D_1 from camera 1 to the target, what is the error distribution for different locations of camera 2, which determines configurations of baselines and angles of the panoramic stereo? Since it is hard to give a numerical function of the error $\partial\phi_2$ versus the location O_2, we will use the same measure error bounds for all the angles, i.e. $\partial\alpha = \partial\phi_1 = \partial\phi_2 \equiv \partial\phi$. We will re-examine this matching error qualitatively after we find the optimal baseline/viewpoints. We decompose the analysis into two steps. First, by fixing the baseline, we find the optimal angle ϕ_1 . It is equivalent to finding the optimal position of O_2 on a circle of origin O_1 and radius B (Fig. 5). Second, under the optimal angle configuration of all possible baselines, we find the optimal baseline B. An additional consideration is that a human has a size comparable to the robots, so the distances between a robot and the target should be at least greater than the dimension of the robot, *2R*. We have the following results by combining the error analysis in the triangulation method (Appendix 1) and in the size-ratio method (Appendix 2):

Case (1) . When $B \le D_1\text{-}2R$, the best estimation can be achieved when

$$B = 2\sqrt{D_1 R}\,,\quad \cos\phi_1 = \frac{3BD_1}{2D_1^2 + B^2} \tag{7}$$

and the error in the optimal configuration is

$$\partial D_1^+ = \partial D_1^B + \partial D_1^\phi = D_1\left(\frac{\sqrt{4D_1 R - R^2}}{2R} + \frac{\sqrt{(D_1 - 4R)(D_1 - R)}}{\sqrt{D_1 R}}\right)\partial\phi < 2D_1\sqrt{\frac{D_1}{R}}\partial\phi \tag{8}$$

where ∂D_1^B and ∂D_1^ϕ are the distance errors due to the baseline error and angular errors respectively. Note that in this case, the minimum error is achieved when $\phi_1 < 90^\circ$, $\phi_2 > 90^\circ$ and $\phi_0 < 90^\circ$ (see Appendix 1). For example, when R=0.18 m, D_1=4.0 m, $\partial\phi$ = 6.28 mrad (1 pixel), we have the best configuration of B=1.70 m and $\phi_1 = 54.2^\circ$, and the relative error is $\partial D_1 / D_1$=5.4%.

Case (2) . When $B \in (D_1\text{-}2R, \sqrt{D_1^2 + 4R^2}\,)$, the best estimation can be achieved when

$$B = D_1\,,\quad \cos\phi_1 = \frac{B^2 - 2R^2}{B^2} \tag{9}$$

and the error in the optimal configuration is

$$\partial D_1^0 = D_1\left(\frac{\sqrt{D_1^2 - R^2}}{2R} + \frac{3R}{\sqrt{D_1^2 - R^2}}\right)\partial\phi \tag{10}$$

Note that in this case, $\phi_1 < 90^\circ$ is the minimum angle by physical constraint of the minimum object distances, and $\phi_2 = \phi_0$.

Case (3) . When $B \geq \sqrt{D_1^2 + 4R^2}$, the best estimation can be achieved when

$$B = \sqrt{D_1^2 + 4R^2}\ \mathbf{,}\ \ \cos\phi_1 \ = \frac{D_1}{B} \tag{11}$$

and the error in the optimal configuration is

$$\partial D_1^- = D_1(\frac{\sqrt{D_1^2 + 3R^2}}{2R} + \frac{2R}{D_1})\partial\phi \tag{12}$$

Note that in this case, the minimum error is achieved when $\phi_1 < 90^\circ$, $\phi_2 < 90^\circ$ and $\phi_0 = 90^\circ$.

Case (4). In the case of colinearity of sensors and the target, triangulation is invalid. However we can use the size-ratio method. A similar error analysis (Appendix 2) shows that if the target lies between the two cameras, minimum error is obtained when the second camera *O₂* moves as close as possible to the target, i.e. $D_2 = 2R$, or

$B = D_1 + 2R,$

and the minimum error can be expressed by

$$\partial D_1^s = D_1(\frac{\sqrt{(D_1 + R)(D_1 + 3R)}}{2R} + \frac{2R}{W})\partial w \tag{13}$$

We always have $\partial D_1^s > \partial D_1^-$ given that $B > D_1$, , $\partial w = \partial\phi$ and $W << D_1$. Similar results can be obtained when the target lies in one side of both sensors. It can be also proved that we always have $\partial D_1^- < \partial D_1^0$, which means that it is better to set the baseline slightly greater than the distance D_1 when they have to be approximately equal. (In addition, the equality condition cannot be satisfied before we have an accurate estimation of D_1). By some tedious mathematical comparison of Eq. (8) and Eq. (12) under different D_1, we arrive at the following observation:

Conclusion 1. If the distance from camera 1 (the main camera) to the target is greater than 11.5 times the radius of the robot, i.e. $D_1 > 11.5R$, we have $\partial D_1^+ < \partial D_1^-$, which means that the best configuration is $B = 2\sqrt{D_1 R}$, $\cos\phi_1 = \dfrac{3BD_1}{2D_1^2 + B^2}$ *(Eq. (7)). Otherwise, we have* $\partial D_1^+ >= \partial D_1^-$ *i.e. the best configuration is* $B = \sqrt{D_1^2 + 4R^2}$, $\cos\phi_1 = \dfrac{D_1}{B}$ *(Eq. (11)).*

It is also interesting to compare the panoramic virtual stereo with a fixed baseline stereo. Assume that in a fixed baseline stereo system on a robot, the two cameras are mounted as far apart as possible. For a robot with cylindrical body of radius *R*, the maximum stereo baseline in that case would be *B=2R*. Let us assume that there is no error in stereo camera calibration (i.e. *B* is accurate). Since we always have $B < D_1$ in fixed-baseline stereo, we can use Eq. (19) in Appendix 1 to estimate the distance error in the *best* case, i.e.

$$\partial D_1^{fix}\big|_{B=2R} \approx \frac{D_1^2}{R}\partial\phi \qquad (14)$$

Comparing Eq. (14) with Eq. (8), we have the following conclusion:

Conclusion 2. The flexible baseline triangulation method is almost always more accurate than a fixed baseline stereo. The error in fixed baseline stereo is proportional to D_1^2*, but the error in flexible baseline stereo is proportional to* $D_1^{1.5}$*. The error ratio is*

$$\partial D_1^+\big|_{B=2\sqrt{D_1 R}} : \partial D_1^{fix}\big|_{B=2R} = 2\sqrt{\frac{R}{D_1}} \qquad (15)$$

and $\partial D_1^+\big|_{B=2\sqrt{D_1 R}} < \partial D_1^{fix}\big|_{B=2R}$ *when $D_1 > 4R$, which is almost always true.*

The above error analysis results can be used in the optimal view planning. Though the exact number 11.5R in Conclusion 1 is deduced from the calibration method we are using, the guidelines apply to general cases. The distance error map under different viewpoints of camera O_2 is given in Fig. 6 for D_1= 34R = 6 m to verify the above conclusion. Minimum error is $\partial D_1 / D_1 \partial\phi$ =11.2 when *B*=220cm, $\phi_1 = 62.1°$ (We have two such symmetric locations for O_2).

The upper bound of the relative error is $\partial D_1 / D_1$=7.0% when $\partial \phi$ is equivalent to 1 pixel. The selection of optimal viewing angle and baseline for different distances is shown in Fig. 7. Note that parameters in Fig. 7 are slightly different from those in Fig. 6 because the curves in Fig. 7 are drawn using Eq. (8) and Eq. (12) with some approximation and practical consideration. The error analysis can also be used in the integration of the results from more than two such stationary sensors.

5 Matching Primitives and Matching Error Revisited

Since our primary goal is to detect and to track moving targets (humans) in 3D space, the primitives of the panoramic virtual stereo are image blobs of human subjects that have already been extracted from the two panoramic images. A fast moving object extraction and tracking algorithm using motion detection and background subtraction with a stationary panoramic camera has been developed [24]. Fig. 8 depicts the results of our multiple human detection and tracking procedure. Multiple moving objects (4 people) were detected in real-time while moving around in the scene in an unconstrained manner; the panoramic sensor was stationary. Each of the four people was completely extracted from the complex background, as depicted by the bounding rectangle, direction and distance of each object. The dynamic track, represented as a small circle and icon (elliptic head and body) for the last 30 frames of each person is shown in Fig. 8b in different colors. The frame rate for multiple object detection and tracking was about 5 Hz in a Pentium 300MHz PC for 1080*162 panoramic images, and thus can be 15-20 Hz with current standard CPUs.

We have realized that the bearing of the centroid of an entire blob is subject to the effects of the positions of arms and legs, and the errors in body extraction. We have found that the bearing of the head of a human is more accurate than the entire blob of the human subject for three reasons: (1) it is usually visible in the panoramic images; (2) it is almost symmetric from all directions of the robot's viewpoints; and (3) it is easy to extract from the background (see Fig. 8 and Fig. 9). The quasi-symmetry property of a head makes it more suitable for matching across two widely separated views. The idea to select invariant features for stereo matching can be further extended by extracting different parts of a human blob for partial match between two views.

The head part of a blob is extracted by using the knowledge that it is the topmost part of the blob and it has roughly a fixed height-width ratio (e.g., 3:2) in a panoramic image. Here the exact height of the head segment is not critical since we only use the bearing angle of the head for triangulation. Fig. 9 shows the extracted human blobs and heads from a pair of panoramic images. Bearing of the head is more suitable for building up correspondence between a pair of human blobs from two widely separated views because of the aforementioned reasons. Notice that the centroid of each head region gives correct bearing of the head even if the size and view differences are large between two images of the same human subject. The estimated height is not accurate and not consistent across the corresponding image pair. For example, the second human subject in the images shows that the bearing of the head is more accurate than the entire blob, which is an inaccurate detection of the human body: the left side is "underestimated" due to the similarity between the shirt and the door, and the right side is "overestimated" due to its shadow.

From each panoramic image, a set of objects (blobs) is extracted, which is *annotated* by the following parameters

$$\mathbf{T}^{(k)} = \{T_i^{(k)} = (I_i^{(k)}, \theta_i^{(k)}, w_i^{(k)}, h_i^{(k)}), i = 1, ..., N_k\} \qquad (16)$$

where *k* (1 or 2) is the number of cameras, $I_i^{(k)}, \theta_i^{(k)}, w_i^{(k)}, h_i^{(k)}$ are the photometric feature, bearing angle of the head of the target *i* in camera *k*, the width of the image blob, and the vertical coordinate of the top of the blob (indicating the height of the human).

The best triangulation configuration is derived when all the angular errors ($\partial\alpha, \partial\phi_1, \partial\phi_2$) are treated as the same, and are assumed to be independent to the view configuration of the panoramic stereo. However, as we discussed in Section 4.2, the error $\partial\phi_2$ should be a function of the position of O_2 (given the locations of O_1 and *T).* A quantitative result can be derived in the same manner as above if the function is known or can be approximated; but here we only give a qualitative analysis. The error map in Fig. 6 shows that there is a relatively large region (black part of the minimum-error curve) with errors that are less than twice the minimum error. The large errors only occur when angle ϕ_0 is very close to 0° and 180°. Therefore a tradeoff can be made between the matching error (resulting from widely separated views) and the triangulation error (resulting from small baseline). The target appears similar from both the cameras at the best triangulation configuration (in the typical case when $D_1 > 11.5R$), since the distances from the two cameras to the target are comparable ($D_2 = \sqrt{D_1^2 - 4D_1R}$). In addition, it is interesting to note that larger view difference can give a better measurement of the dimension of the 3D object (person), which is similar to the volume intersection method (Fig. 10).

6 Cooperative Strategies in the Real System

In our panoramic virtual stereo vision approach, we face the same problems as in traditional *motion stereo*: dynamic calibration, feature detection, and matching. In our scenario, we are also dealing with moving objects before 3D matching, which seems to add more difficulty. Fortunately, the following cooperative strategies can be explored between two robots (and their panoramic sensors) to ease these problems: a “monitor-explore” working mode, mutual awareness, information sharing and view planning.

6.1 Monitor-explore mode

In the two-robot scenario of human searching, one of the robots is assigned as the "monitor" and the other as the “explorer”. The role of the monitor is to monitor the movements in the environment, including the motion of the explorer. One of the reasons that we have a monitor is that it is advantageous for it to be stationary while detecting and extracting moving objects. On the other hand, the role of the explorer is to follow a moving object of interest and/or find a better viewpoint for constructing the virtual stereo geometry with the camera on the monitor. However, the motion of the explorer introduces complications in detecting and extracting moving objects, so we assume that the explorer remains stationary in the beginning of an operational sequence in order to initialize moving objects to be tracked. Then a tracking mechanism that can handle ego-motion of the robot continues to track objects of interest. Such a tracking procedure may integrate the motion, texture and other cues, which need future work. We also expect that the explorer will remain stationary in an advantageous location after it has found a good viewpoint for 3D estimation. The role of the monitor and the explorer can and will be exchanged during mission execution. The exchange of roles as well as the motion of the explorer may be determined by evaluating expected gain in triangulation accuracy. For example, when the expected improvement is not significant, the robots may just remain in their current state.

6.2 Mutual awareness and information sharing

Mutual awareness of the two robots is important for their dynamic calibration of relative orientations and the distance between the two panoramic cameras. In the current implementation, we have designed a cylindrical body with known radius and color, so it is easy for the cooperating robots to detect each other. It is interesting to note that while the motion of the explorer increases the difficulty of tracking other moving objects by itself, tracking information from the monitor is quite useful. It is also possible to use more complicated but known *natural appearances* and geometrical models of a pair of robots to implement the mutual awareness and dynamic mutual calibration.

The two panoramic imaging sensors have almost identical geometric and photometric properties. Thus it is possible to share information between them about the targets as well as the robots in the scene. For example, when some number of moving objects are detected and extracted by the stationary monitor, it can pass the information of the number of objects and their geometric and photometric features of each object to the explorer that may be in motion, thereby increasing robustness of tracking by the moving explorer. Information sharing is especially useful for the mutual detection of "cooperative" calibration targets since models of the robots are already known *a priori*. In our simplified case, the cylindrical bodies of both robots always have the same appearances from any viewing angle. Therefore, whenever the monitor has detected the cylindrical body of the moving explorer, it can estimate the bearing and distance to the explorer. On receiving this information from the monitor, the explorer can try to search for the cylindrical body of the monitor in its image by predicting its size and color under the current configuration and illumination conditions.

6.3 View planning for a pair of robots and a single target

View planning is applied whenever there are difficulties in object detection and 3D estimation by the virtual stereo system. In our case, we define the view planning as the process of adjusting the viewpoint of the exploring camera so that the best view angle and baseline can be achieved for the monitoring camera to estimate the distance to the target of interest. Occlusion of the human or the robot may occur when an object (either a human or a robot) is between the observing camera and the target, the configuration when triangulation is invalid (we use the size-ratio method in that situation for an initial estimate). The error analysis in Section 4 provides guidelines for "best" viewing planning as follows:

(1) Observation rule - This rule is applied when the two robots "observe" the target from a distance. *If the initial estimated distance from viewpoint O_1 to the target, D_1, is greater than 11.5R, the explorer should move as close as possible to an optimal position that satisfies the minimum distance error conditions, i.e., baseline constraint* $B = 2\sqrt{RD_1}$ *and the viewing angle constraint* $\cos\phi_1 = \frac{3BD_1}{2D_1^2 + B^2}$.

(2). Approaching rule - This rule is applied when both the two robots are close to the target and the explorer is trying to "approach" the target. *If the estimated distance is smaller than 11.5R, the explorer should approach to the target to satisfy the baseline constraint* $B = \sqrt{D_1^2 + 4R^2}$ *and the viewing angle constraint* $\cos\phi_1 = \frac{D_1}{B}$.

(3). Mutual-awareness rule - *When two panoramic cameras are aware the existence of each other, the maximum distance of the baseline is $B = 2R/w\chi$, given the angular resolution of the* *panoramic image, χ, the size of the cylindrical robot body, R, and minimum number of detectable pixels of the robots, w.*

For example, assume that w=10 pixels is the minimum detectable width, then the maximum baseline is B=2.8 m given R=0.18 m and χ = 6.28 mrad/pixel. This constraint on the baseline still allows the optimal configuration of the panoramic virtual stereo to provide an effective estimation of the distance of a target 10 meters away (Fig. 7).

***(4). Navigation rule** – The view planning strategy should also consider the cost of moving in finding a navigable path to the selected position. This cost is also a function of distance, smoothness of the path and time to travel.*

Note that the explorer is always trying to find a best position in the presence of a target's motion. It is a more difficult problem than localizing a stationary target. For a stationary target, as the robot is moving, the system could collect many stereo estimates (of the target) along the way and integrating them to form a much higher quality estimate of the actual position. However, for a continuously moving object, the integration (if possible) requires formation of the dynamic track of the moving object. This integration could use our derived error model with a Kalman filter methodology.

6.4 View planning for multiple robots and multiple objects

These strategies can be extended to more than two cooperative robots, and in fact more than two robots will make the work much easier. For example, we can keep two of the three robots in a team stationary so that they can easily detect the moving objects in the scene, including the third robot in motion. Thus the locations of all the moving objects can be estimated from the pair of stationary panoramic cameras. Then, for a target of interest, we can find (dynamically) the best viewpoint for the third robot in order to estimate the target's distance from either of the two stationary robots. By using the knowledge of the (dynamic) locations of the target, other moving objects and the three robots, a navigable path for the third robot can be planned to the desirable goal. These measurements can also facilitate the detection of the target and the two stationary robots by the mobile robot, for example, by tracking the objects with visual features inherited from the other two robots. Thus, the stereo triangulation relation can be constructed between the moving and the stationary platforms.

On the other hand, the view planning rules for a single moving object can also be extended to deal with multiple moving objects. There are three interesting cases.

(1). In general, N+1 robots can construct optimal configurations for N moving objects (N >2), i.e. a main robot can cooperate with each of the N robots for the detection and localization of each of the N objects (Fig. 11a). However, this method is inefficient and needs to move the N robots.

(2). As a special case (Fig. 11b), two moving robots with panoramic cameras (O_1 and O_2) can construct optimal configurations for estimating the distances of two moving objects($T^{(1)}$ and $T^{(2)}$), by the alignment of the two viewpoints of the cameras to mirror each other.

(3). As an approximation method, two moving robots with panoramic cameras can construct near optimal configurations for estimating the distances of multiple moving objects. This can be done by clustering the targets into two groups, and the two cameras then configure two best triangulations for the centers of the two groups (Fig. 11b). It should be apparent that more than two robots usually can do a better job in view planning.

7 Experimental System and Results

In our experimental system, we mounted one panoramic annual lens (PAL) camera on an RWI ATRV-Jr. robot (the explorer), and the other PAL camera on a tripod (the monitor)(Fig. 3a). Two Matrox-Meteor frame grabbers, each connected to a PAL camera were installed on the ATRV-JR and a desktop PC respectively, at the time both had 333M Hz PII processors. The communication between two platforms is through sockets over an Ethernet link (wireless Ethernet communication will be used in the future system). The 3D moving object detection and estimation programs run separately on the two machines at about 5 Hz. Only camera and object parameter data (i.e., baseline, bearing angles, sizes, and photometric features in Eq. (16)) were transmitted between two platforms so the delay in communication can be ignored at the current processing rate (5Hz). In the implementation of examples shown in this paper, we assume that the most recent results from both platforms correspond to the events at same time instant. Synchronized image capture is currently using network time protocol (NTP) with temporal interpolation [26], and it not an issue with higher speed CPUs.

Fig. 12 shows the result from an experiment to evaluate the panoramic stereo's performance of tracking a single person walking along a known rectangular path when the two cameras were stationary. Each red dot (dark in B/W print) represents a location of the person. The dense clusters of dots show the six locations where the person made turns during the walking. We used two methods to localize the moving subject – the triangulation method when a good triangle of the target and the two cameras can be formed, and the size-ratio method when the target was near the locations of colinearity. The theoretical error bounds superimposed on the 2D map in Fig. 12 (d), were computed assuming that all the angular errors in Eq. (5) and Eq. (6) were equivalent to 1 pixel. The target (T) position where the theoretical best triangulation on this track can be expected is shown in the figure (Fig. 12(b) as well as (d)), which is consistent with the real experimental results. Even if the localization errors in images are larger than the assumed 1 pixel error in our previous analysis, the average error of the estimated track is ±10 cm, which is comparable to the theoretical error bounds. The bad localizations occurred when the extraction of the human blobs is not correct. For example, in the beginning and the end of the walk where the size-ratio method is used, large errors occurred not because the size-ratio method is not stable computationally, but because the detection of the human figure is not correct when it is small and occluded by one sensor platform.

Fig. 13 shows the results of detecting and tracking two people who walked from the opposite directions along the same known rectangular path. In this example, a simple "greedy" match algorithm [22, 27] was used where the similarities of the matching primitives in intensity and the consistencies in 3D measurements are calculated. In the 2D map of the room (center of each picture in Fig. 13), the red (which is darker in B/W print) dot sequence shows the path of one person, and the green (which is lighter in B/W print) dot sequence shows that of the other. The proposed 3D match, localization and tracking algorithms produced rather good results with consistent 3D localization for both people. The average localization error is about 20 cm. There are about 5% mis-matches in this set of experiment, which happened in two places. One place is when the shadow of a person was projected on the wall and was detected and mis-matched by the system. The second place of error is when the two people met. Further improvements and experiments on stereo match, view planning and evaluation are needed.

8 Concluding Remarks

This paper has presented a panoramic virtual stereo approach for two (or more) cooperative mobile platforms. There are three main contributions in our approach: (1) a simple but effective dynamic mutual calibration between two panoramic sensors; (2) a thorough error analysis for the panoramic virtual stereo vision system; and (3) viewing planning based on optimal stereo configurations. The integration of omni-directional vision with mutual awareness and dynamic calibration strategies allows intelligent cooperation between visual agents, which provides an effective way to solve problems of limited resources, view planning, occlusions and motion detection of mobile robot platforms. Experiments have shown that this approach is encouraging. At the system level, the panoramic virtual stereo is one of the important modules to localize multiple moving human subjects in the distributed sensor network architecture proposed in [26]. In particular, the error modeling and the view planning strategy developed in this paper are applied. Interesting future work include the following:

(1) ***Improvement of the mutual calibration accuracy*** - By integrating a panoramic camera with a pan/tilt/zoom camera, the system can increase the capability in both viewing angle and image resolution to detect the cooperating robots as well as the targets. Robust and accurate dynamic mutual calibration is one of the key issues in cooperative stereo vision.

(2) ***Improvement of 3D matching*** - By using the image contours of objects and more sophisticated features (color, texture, etc), more accurate results can be expected. This is another significant factor that affects the robustness and accuracy of 3D estimation.

(3) ***Tracking of 3D moving objects*** - More sophisticated algorithms for tracking moving objects should be incorporated, in the presence of occlusion, and by moving cameras as well as stationary cameras.

References

1. Haritaoglu, I., D. Harwood and L. Davis, W4S: a real-time system for detection and tracking people in 2.5D, In *Proceedings of ECCV*, 1998.
2. Lipton, A. J., H. Fujiyoshi, R. S. Patil, Moving target classification and tracking from real-time video, In *Proceedings of DARPA Image Understanding Workshop*, volume 1, pages 129- 136, November 1998.
3. Papageorogiou, C., T. Evgeniou, and T. Poggio, A trainable object detection system, In *Proceedings of DARPA Image Understanding Workshop*, volume 2, pages 1019-1024, November 1998.
4. Pentland, A., A. Azarbayjani, N. Oliver and M. Brand, Real-time 3-D tracking and classification of human behavior, In *Proceedings of DARPA Image Understanding Workshop*, volume 1, pages 193-200, May 1997.
5. Brill, F. Z., T. J. Olson and C. Tserng, Event recognition and reliability improvements for the autonomous video surveillance systems, In *Proceedings of DARPA Image Understanding Workshop*, volume 1, pages 267- 284, November 1998.
6. *DARPA Image Understanding Workshop Proceedings*, VSAM- Video Surveillance and Monitoring Session, Monterey, November 1998

7. Baker, S. and S. K. Nayar, A theory of catadioptric image formation, In *Proceedings of the 6th International Conference on Computer Vision*, IEEE, India, January 1998.

8. Nalwa, V., A true omnidirectional viewer, *Technical Report*, Bell Lab, Holmdel, NJ, Feb, 1996

9. Greguss, P., Panoramic imaging block for three-dimensional space, *U.S. Patent* 4,566,763 (28 Jan, 1986)

10. Yagi, Y., S. Kawato, Panoramic scene analysis with conic projection, In *Proceedings IROS*, 1990

11. Yamazawa, K, Y. Yagi and M. Yachida, Omnidirectional imaging with hyperboloidal projections, In *Proceedings IROS*, 1993.

12. Powell, I., Panoramic lens, *Applied Optics*, vol. 33, no 31, Nov 1994:7356-7361

13. Zhu, Z, S. Yang, G. Xu, X. Lin, D. Shi, Fast road classification and orientation estimation using omni-view images and neural networks, *IEEE Trans Image Processing*, Vol 7, No 8, August 1998: pp. 182-1197.

14. Hong J, Tan X, Pinette B, R. Weiss and E.M. Riseman, Image-based homing, In *Proc Int. Conf Robotics and Automation*, April 1991, pp 620-625

15. Ishiguro, H., M. Yamamoto and S. Tsuji, Omni-directional Stereo, *IEEE Trans. PAMI*, Vol. 14, No.2, 1992: 257-262

16. Konolige, K. G., R. C. Bolles, Extra set of eyes, In *Proceedings of DARPA Image Understanding Workshop*, volume 1, pages 25- 32, November 1998.

17. Kawasaki H, Ikeuchi K. Sakauchi M, Spatio-temporal analysis of omni image, In *CVPR'00*, pp 577-584, 2000.

18. Boult, T., E., R. Micheals, X. Gao, P. Lewis, C. Power, W. Yin, A. Erkan, Frame-Rate omnidirectional surveillance and tracking of camouflaged and occluded targets, In *Proceedings of the Second IEEE Workshop on Visual Surveillance*, June 1999: 48-58.

19. Ng, K. C., H. Ishiguro, M. Trivedi and T. Sogo, Monitoring dynamically changing environments by ubiquitous vision system, In *Proceedings of the Second IEEE Workshop on Visual Surveillance*, June 1999: 67-73

20. D. Gutchess and A.K. Jain, Automatic Surveillance Using Omnidirectional and Active Cameras, In *Proceedings of the 4th Asian Conf. Computer Vision*, Taipei, Jan. 2000

21. Sogo, T., H. Ishiguro, M. M. Trivedi. N-ocular stereo for real-time human tracking. *Panoramic Vision: Sensors, Theory and Applications*, (R. Benosman and S. B. Kang, eds.), Springer Verlag, 2000.

22. Zhu, Z., K. D. Rajasekar, E. Riseman, A. Hanson. Panoramic Virtual Stereo Vision of Cooperative Mobile Robots for localizing 3D Moving Objects. In *Proceedings of IEEE Workshop on Omnidirectional Vision – OMNIVIS'00*, Hilton Head Island, 29-36, JUNE 2000.

23. Trivedi, M., K. Huang, I. Mikic. Intelligent Environments and Active Camera Networks. *IEEE Systems, Man and Cybernetics*, October 2000.

24. Zhu, Z., E. M. Riseman, A. R. Hanson, Geometrical modeling and real-time vision applications of panoramic annular lens (PAL) camera, *Technical Report TR #99-11*, Computer Science Department, University of Massachusetts Amherst, February, 1999.

25. Shum, H.-Y., A. Kalai, S. M. Seitz Omnivergent stereo, in *Proceedings of the IEEE Seventh International Conference on Computer Vision*, September 1999: pp 22 - 29

26. Karuppiah, D. R., Z. Zhu, P. Shenoy, and E. M. Riseman, A fault-tolerant distributed vision system architecture for object tracking in a smart room, *IEEE Second International Workshop on Computer Vision Systems*, Vancouver, Canada, July 2001. B. Schiele and G. Sagerer (Eds.), Springer Lecture Notes in Computer Science 2095, pp 201-219.

27. Zhu, Z., D. R. Karuppiah, E. M. Riseman and A. R. Hanson, Adaptive Panoramic Stereo Vision for Human Tracking and Localization with Cooperative Mobile Robots. Accepted by *Robotics and Automation Magazine*, special issue on panoramic robots.

28. Cielniak, G., M. Miladinovic, D. Hammarin, L. Göransson, A. Lilienthal and T. Duckett, Appearance-based tracking of persons with an omnidirectional vision sensor. *IEEE Workshop on Omnidirectional Vision (in conjunction with CVPR), June 2003*

29. Shakernia, O., R. Vidal and S. Sastry, Structure from small baseline motion with central panoramic cameras. *IEEE Workshop on Omnidirectional Vision (in conjunction with CVPR), June 2003*

30. Adorni, G., S. Cagnoni, M. Mordonini and A. Sgorbissa, Omnidirectional stereo systems for robot navigation. *IEEE Workshop on Omnidirectional Vision (in conjunction with CVPR), June 2003*

31. Menegatti, E., A. Scarpa, D. Massarin, E. Ros and E. Pagello, Omnidirectional distributed vision system for a team of heterogeneous robots. *IEEE Workshop on Omnidirectional Vision (in conjunction with CVPR), June 2003*

Appendix 1. Best baseline and viewpoint when $B < D_1$

In the first step, we are trying to find the minimum value of the error due to the second and third terms of Eq. (5), i.e.

$$\partial D_1^{\phi} = D_1 \left|\cot(\phi_1 + \phi_2)\right| \partial\phi_1 + \frac{D_2}{\sin(\phi_1 + \phi_2)} \partial\phi_2 \qquad (17)$$

It is equivalent to find the optimal position of O_2 on a circle of origin O_1 and radius B. We first consider the case where $B < D_1$. In this case, $(\phi_1 + \phi_2) > 90°$, so Eq. (17) can be re-written as a function of ϕ_1 by using the sine and cosine laws

$$\partial D_1^{\phi} = \frac{B^2 + 2D_1^2 - 3BD_1 \cos\phi_1}{B\sin\phi_1}\partial\phi \tag{18}$$

where we assume that the same measure errors in angles, i.e. $\partial\phi_1 = \partial\phi_2 = \partial\phi$. By some mathematical deductions, we can find that the minimum error can be achieved when $\cos\phi_1 = \frac{3BD_1}{2D_1^2 + B^2}$. The minimum error under the *best configuration* is

$$\partial D_1^{\phi}\big|_{\min} = \frac{\sqrt{(D_1^2 - B^2)(4D_1^2 - B^2)}}{B}\partial\phi < \frac{2D_1^2}{B}\partial\phi \tag{19}$$

The error in Eq. (18) increases from the minimum value to ∞ when the angle ϕ_1 changes from the optimal value to 0° and 180° respectively (Fig. 6a). Note that in this case, the minimum error is achieved when $\phi_1 < 90^\circ$, $\phi_2 > 90^\circ$ and $\phi_0 < 90^\circ$. Here we compare this result with three special cases (Fig. 14):

(1). *Max-vergent configuration*: Two rays O_1T and O_2T have the maximum vergent angle given the fixed baseline B. In this case $\phi_2 = 90^\circ$, the distance error due to angular errors is

$$\partial D_1^{\phi}\Big|_{\phi_2 = 90^\circ} = \frac{2D_1\sqrt{D_1^2 - B^2}}{B}\partial\phi > \partial D_1^{\phi}\big|_{\min} \tag{20}$$

(2). *Symmetric configuration*: Two rays O_1T and O_2T have the same length given the fixed baseline B. In this case $\phi_1 = \phi_2$, the distance error due to angular errors is

$$\partial D_1^{\phi}\Big|_{\phi_1 = \phi_2} = \frac{D_1\sqrt{4D_1^2 - B^2}}{B}\partial\phi > \partial D_1^{\phi}\big|_{\min} \tag{21}$$

(3). *Mirror configuration*: The ray O1T is perpendicular to the baseline B. In this case $\phi_1 = 90^\circ$, which seems to be a “mirror” of case (1), however, the distance error is larger

$$\partial D_1^{\phi}\Big|_{\phi_1=90^\circ} = \frac{2D_1^2 + B^2}{B}\partial\phi > \frac{2D_1^2}{B}\partial\phi \qquad (22)$$

By a simple comparison we have

$$\partial D_1^{\phi}\Big|_{\phi_1=90^\circ} > \frac{2D_1^2}{B}\partial\phi > \partial D_1^{\phi}\Big|_{\phi_1=\phi_2} > \partial D_1^{\phi}\Big|_{\phi_2=90^\circ} > \partial D_1^{\phi}\Big|_{\min} \qquad (23)$$

which implies three conclusions: (1). Given the fixed baseline, the distance error in the max-vergent configuration or the symmetric configuration is slightly larger than the best configuration. (2). The errors in all the three configurations (max-vergent, symmetric and the best) are smaller than $\frac{2D_1^2}{B}\partial\phi$, which is smaller than the mirror configuration error. (3). Given the fixed baseline, the max- vergent configuration is the closest to the best configuration.

In the second step, we will find the optimal baseline in the case of optimal angle. Inserting Eq. (6) and Eq. (19) into Eq. (5) and assuming that the angle error $\partial\alpha$ in Eq. (6) also equals to $\partial\phi$, we have

$$\partial D_1 = (\frac{D_1\sqrt{B^2 - R^2}}{2R} + \frac{\sqrt{(D_1^2 - B^2)(4D_1^2 - B^2)}}{B})\partial\phi < D_1(\frac{B}{2R} + \frac{2D_1}{B})\partial\phi \qquad (24)$$

It is intuitive that the larger is the baseline, the better the triangulation will be (term 2 in Eq.(24), however the estimated error in the baseline is also larger (term 1). The minimum value can be achieved when $B \approx 2\sqrt{D_1 R}$, which means that 1) more accurate baseline estimation can be obtained given a larger cooperative robotic target (i.e. R), hence the optimal baseline for estimating distance *D₁ can* be larger, and 2) the farther the target is, the larger the baseline *should* be.

Assuming that the human object has a size comparable to the robots, the distances between a robot and the target should be at least greater than the dimension of the robot, *2R*. So Eq.(19) is only valid when $D_2 \geq 2R$, hence we should have $D_1 \geq B+2R$. Similarly, we can find the optimal solutions when $B = D_1$ and $B > D_1$.

Appendix 2 Comparison between triangulation and size-ratio approach

The error for the size-ratio method can be calculated in a similar way, For example, the distance error for Eq. (3) is

$$\partial D_1 = \frac{D_1}{B}\partial B + \frac{D_1}{w_1 + w_2}\partial w_1 + \frac{B - D_1}{w_1 + w_2}\partial w_2 \qquad (25)$$

Using the mutual calibration error Eq. (6) we have

$$\partial D_1 = D_1 (\frac{\sqrt{B^2 - R^2}}{2R} + \frac{B - D_1}{W})\partial w \qquad (26)$$

where W is the width of the target, and we have $w_1 + w_2 = W(\frac{1}{D_1} + \frac{1}{D2})$. We assume that $\partial w_1 = \partial w_2 = \partial \alpha = \partial w$, where w is measured in radians. Obviously, we have $B>D_1$, $D_1>2R$ and $D_2>2R$. Eq. (26) implies that a larger target means better distance estimation. The minimum error is obtained when the baseline is as large as possible ($B = D_1+ 2R$), i.e., the second camera O_2 moves as close as possible to the target ($D_2= 2R$). So the minimum error can be expressed by

$$\partial D_1^s = D_1 (\frac{\sqrt{(D_1 + R)(D_1 + 3R)}}{2R} + \frac{2R}{W})\partial w \qquad (27)$$

We always have $\partial D_1^s > \partial D_1^-$ given that $B>D_{1,}$, $\partial w = \partial \phi$ and $W << D_1$.

Figure Captions

Fig. 1. Panoramic Annular Lens and images. (a) PAL camera (b) An orignal PAL image (768*576) (c) Cylindrical panoramic image

Fig. 2. Two 3D estimation methods. (a) Panoramic triangulation (top view) (b) Panoramic size-ratio method (top view)

Fig. 3. Finding the orientation and the distance using a cylinder. (a) Setup (b) Geometry (top view)

Fig. 4. Dynamic calibration by cylinders (which are pointed by arrows)

Fig. 5. Best view angles and baselines

Fig. 6. Error map for distance D_1 when camera O_2 is in different locations of the map by fixing camera O_1 and the target T (D_1 = 34R = $6m$, R = 18 cm). The labels in the two axes are distances (in meters); the black-white curve shows where the minimum errors can be achieved for viewpoint O_2 on circles with different radii around O_1 *(see explanation in the text)*; the error value ($\partial D_1 / D_1 \partial \phi$) is encoded in intensity: see the corresponding bar.

Fig. 7. Best baselines and angles vs. distance curves (The numbers in the parentheses are given when R = 0.18m)

Fig. 8. Tracking multiple moving objects. (a) Cylindrical images with bounding rectangles around moving objects superimposed, (b) Object tracks, each track is for the last 32 frames

Fig. 9. Head extraction and bearing estimation. The large rectangle around each human subject is the bounding rectangle of the corresponding blob, and the small rectangle inside indicates the centroid of the head.

Fig. 10. Viewing differences and distance/dimension estimation . (a) small viewpoint difference (b) large viewpoint difference

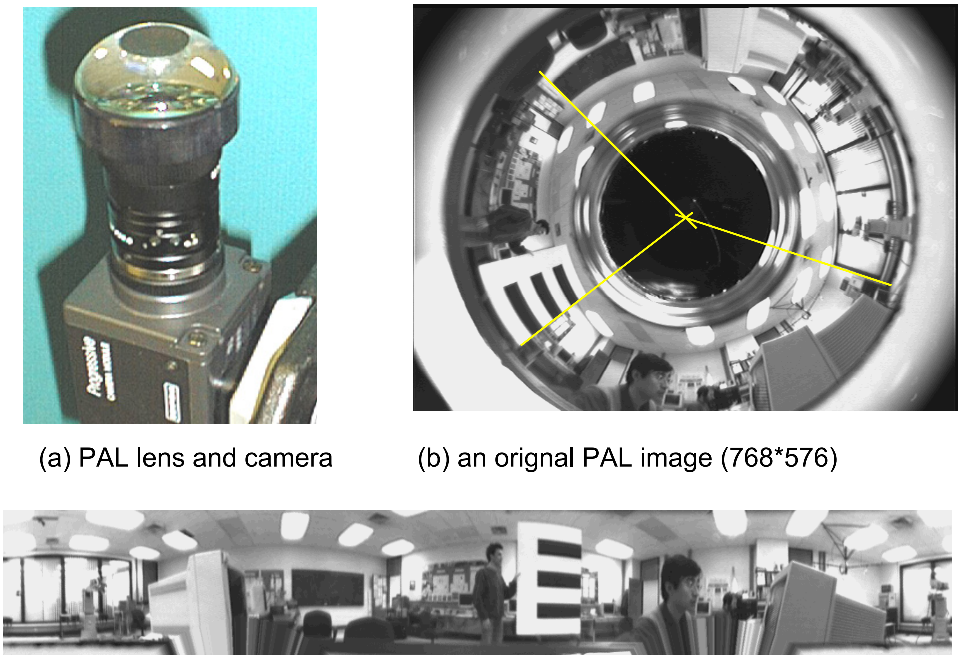
Fig. 11. View planning for multiple robots and multiple objects. (a). N objects, N+1 robots; (b) 2 (groups of) objects, 2 robots

Fig. 12. Panoramic stereo tracking result. The images in (a) and (c) are the panoramic image pair from two panoramic cameras. Each image is actually the corresponding background with the superimposed blob images and their annotations of the blobs. In (c) the real localization results are plotted in the top view of the room where each grid is 50x50 cm^2. Each small circle (red in color version) represents a location of

the person in walking. In (d) The theoretical distance error bounds from camera 1 are shown for the same track. The real estimates for best triangulation results validate the theoretical analysis.

Fig. 13. Panoramic stereo tracking two people. The four pictures show localizing and tracking results (1) before they met, (2) when they met, (3) after they departed and (4) when they arrived at their goals, out of 214*2 localization results. Each picture of (1) to (4) has the same layout as in Fig. 12. Each small circle in darker tone (red in color version) or in lighter tone (green in color version) represents a location of the corresponding person marked by a bounding rectangle with the same tone (color) in images.

Fig. 14. Best viewpoints given the baseline distance



(a) PAL lens and camera

(b) an orignal PAL image (768*576)

(c). Cylindrical panoramic image

Fig. 1. Panoramic Annular Lens and image

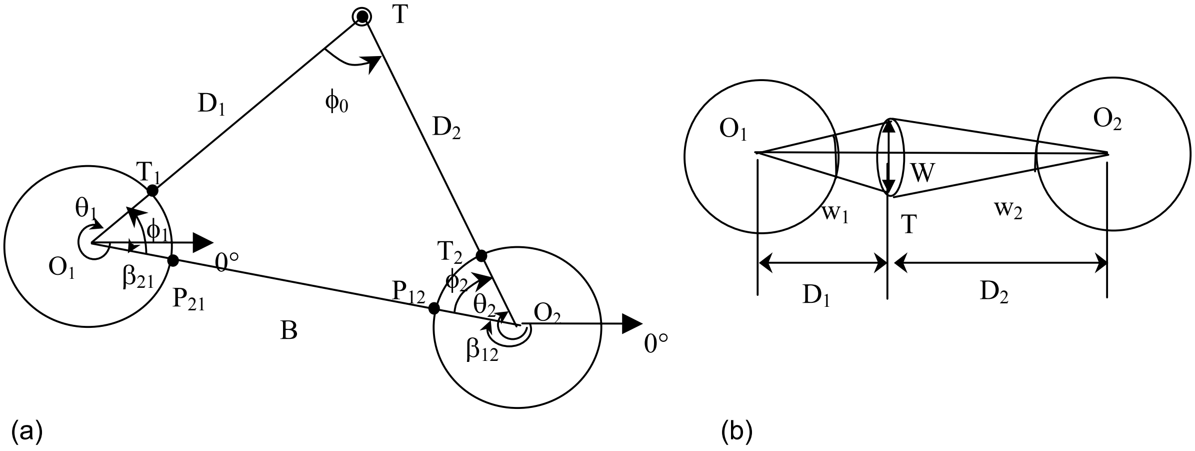
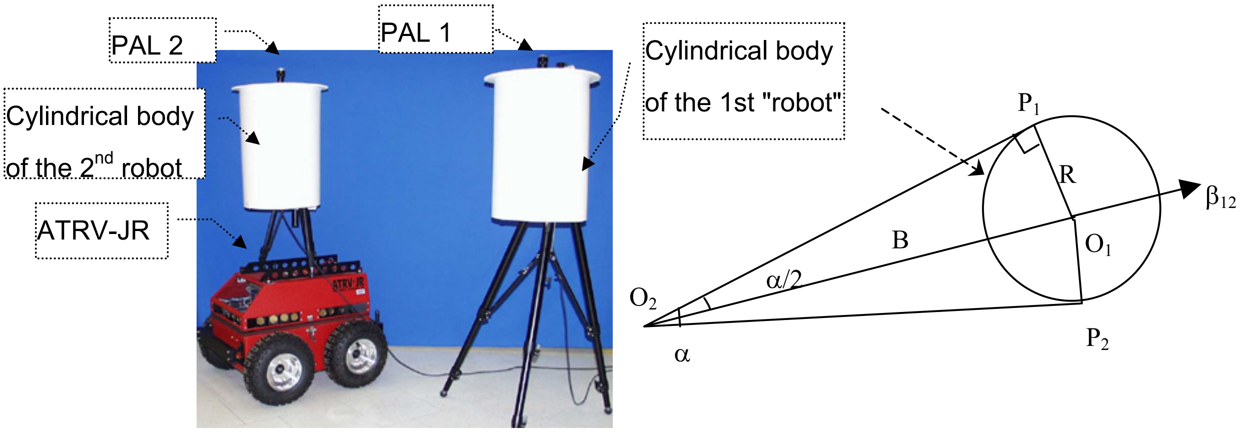


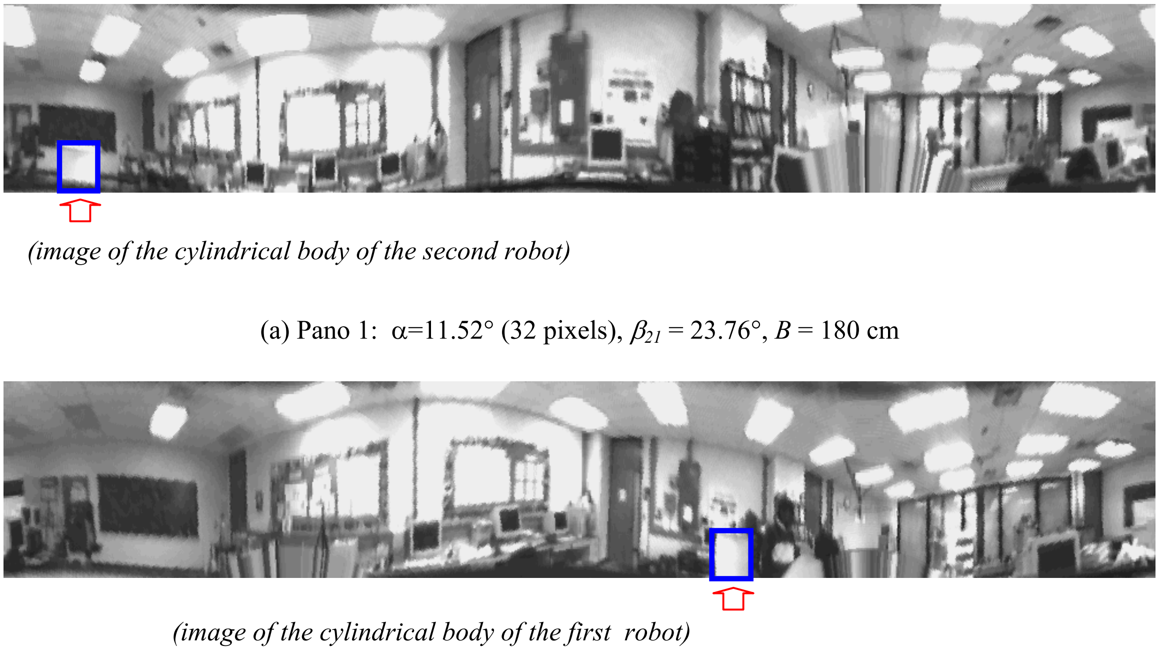
Fig. 2. Two 3D estimation methods. (a) Panoramic triangulation (top view) (b) Panoramic size-ratio method (top view)



(a) Setup

(b) Geometry (top view)

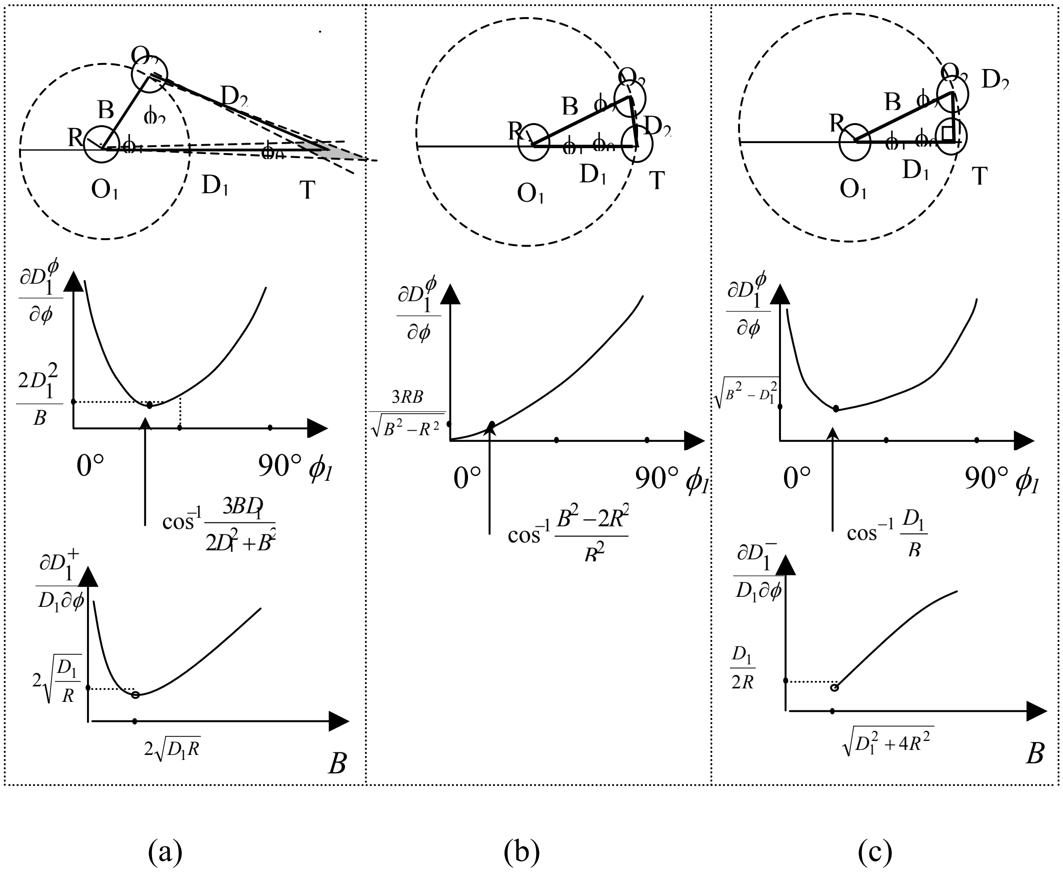
Fig. 3. Finding the orientation and the distance using a cylinder


(image of the cylindrical body of the second robot)
(image of the cylindrical body of the first robot)

(a) Pano 1: α=11.52° (32 pixels), β_{21} = 23.76°, B = 180 cm

(b) Pano 2: α=11.52° (32 pixels), β_{21} = 227.88°, B = 180 cm

Fig. 4. Dynamic calibration by cylinders (which are pointed by arrows)



O2
B
ϕ2
D2
R
ϕ1
ϕ0
O1
D1
T
$\frac{\partial D_1^\phi}{\partial \phi}$
$\frac{2D_1^2}{B}$
0°
90° ϕ_1
$\cos^{-1}\frac{3BD_1}{2D_1^2+B^2}$
$\frac{\partial D_1^+}{D_1 \partial \phi}$
$2\sqrt{\frac{D_1}{R}}$
$2\sqrt{D_1 R}$
B
$\frac{3RB}{\sqrt{B^2-R^2}}$
$\cos^{-1}\frac{B^2-2R^2}{R^2}$
$\sqrt{B^2-D_1^2}$
$\cos^{-1}\frac{D_1}{B}$
$\frac{\partial D_1^-}{D_1 \partial \phi}$
$\frac{D_1}{2R}$
$\sqrt{D_1^2+4R^2}$
(a)
(b)
(c)

Fig. 5. Best view angles and baselines. (a). $B \le D_1\text{-}2R$; (b). $B \in (D_1\text{-}2R,\ \sqrt{D_1^2+4R^2}\ )$; (c). $B \ge \sqrt{D_1^2+4R^2}$

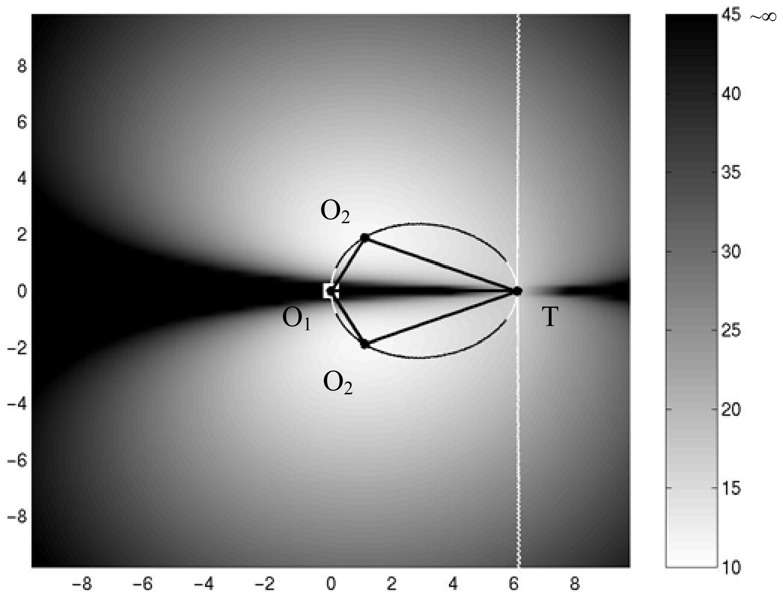


Fig. 6. Error map for distance D_1 when camera O_2 is in different locations of the map by fixing camera O_1 and the target T (D_1 = 34R = $6m$, R = 18 cm). The labels in the two axes are distances (in meters); the black-white curve shows where the minimum errors can be achieved for viewpoint O_2 on circles with different radii around O_1 *(see explanation in the text)*; the error value ($\partial D_1 / D_1 \partial \phi$) is encoded in intensity: see the corresponding bar.

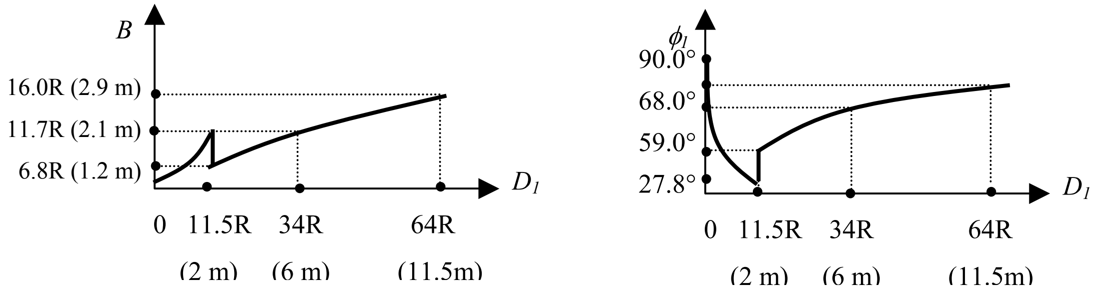
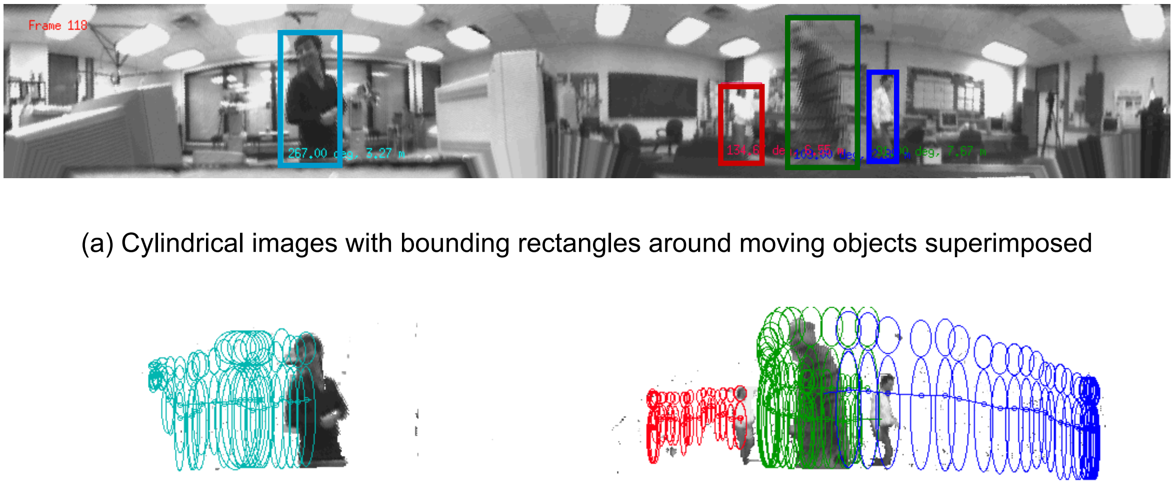


Fig. 7. Best baselines and angles vs. distance curves (The numbers in the parentheses are given when R = 0.18m)



(a) Cylindrical images with bounding rectangles around moving objects superimposed

(b) Object tracks, each track is for the last 32 frames

Fig. 8. Tracking multiple moving objects

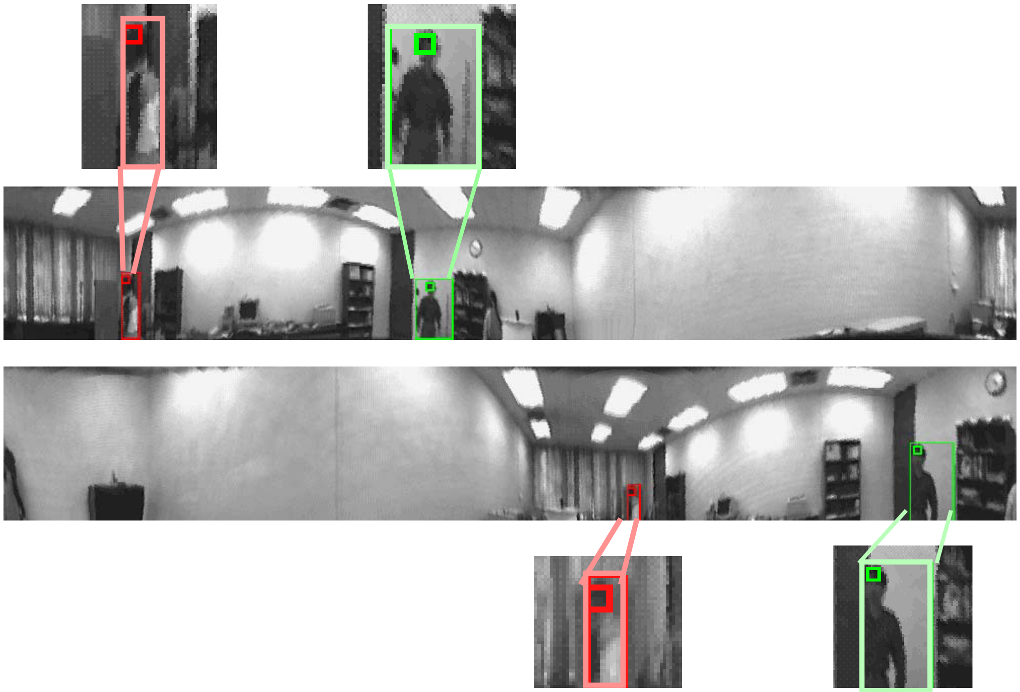
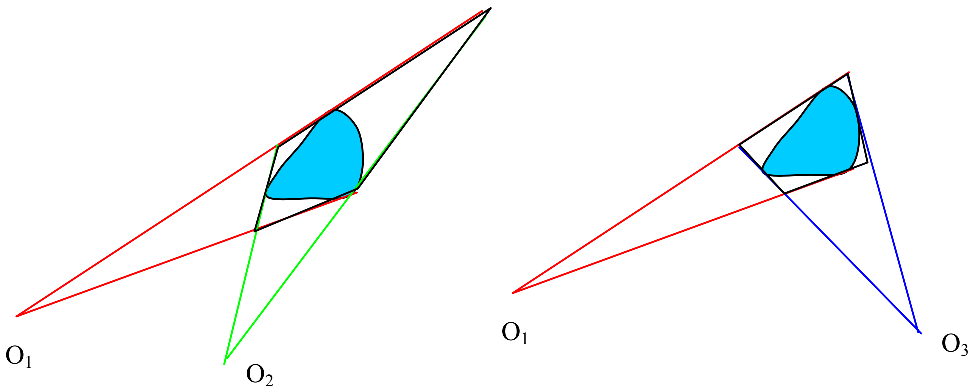
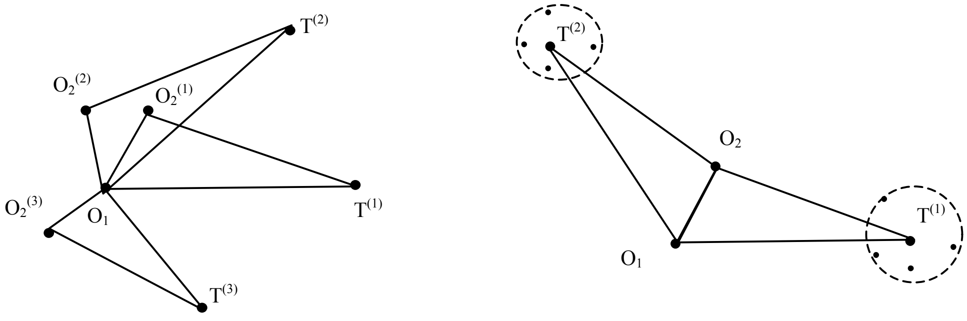


Fig. 9. Head extraction and bearing estimation. The large rectangle around each human subject is the bounding rectangle of the corresponding blob, and the small rectangle inside indicates the centroid of the head.



(a) small viewpoint difference (b) large viewpoint difference

Fig. 10. Viewing differences and distance/dimension estimation



(a). N objects, N+1 robots (b) 2 (groups of) objects, 2 robots

Fig. 11. View planning for multiple robots and multiple objects

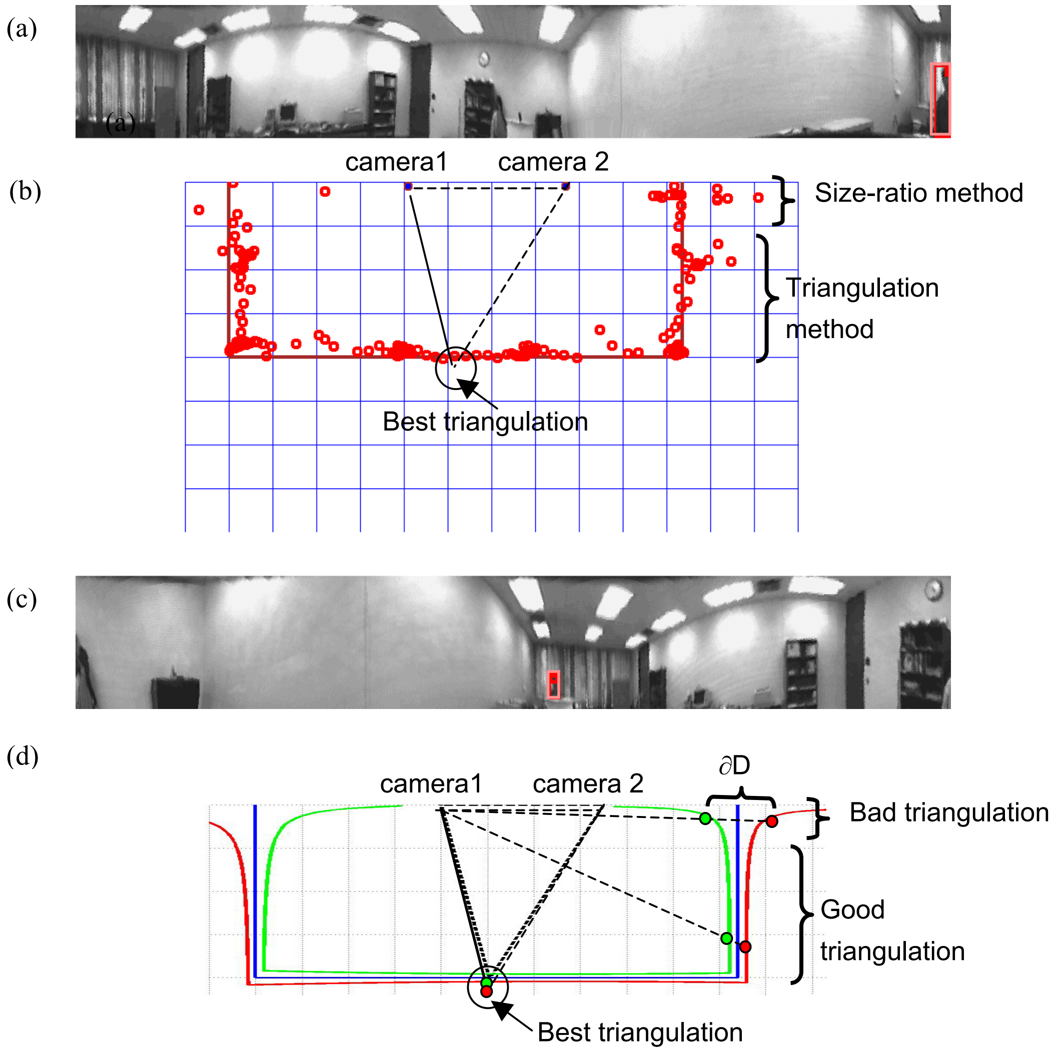
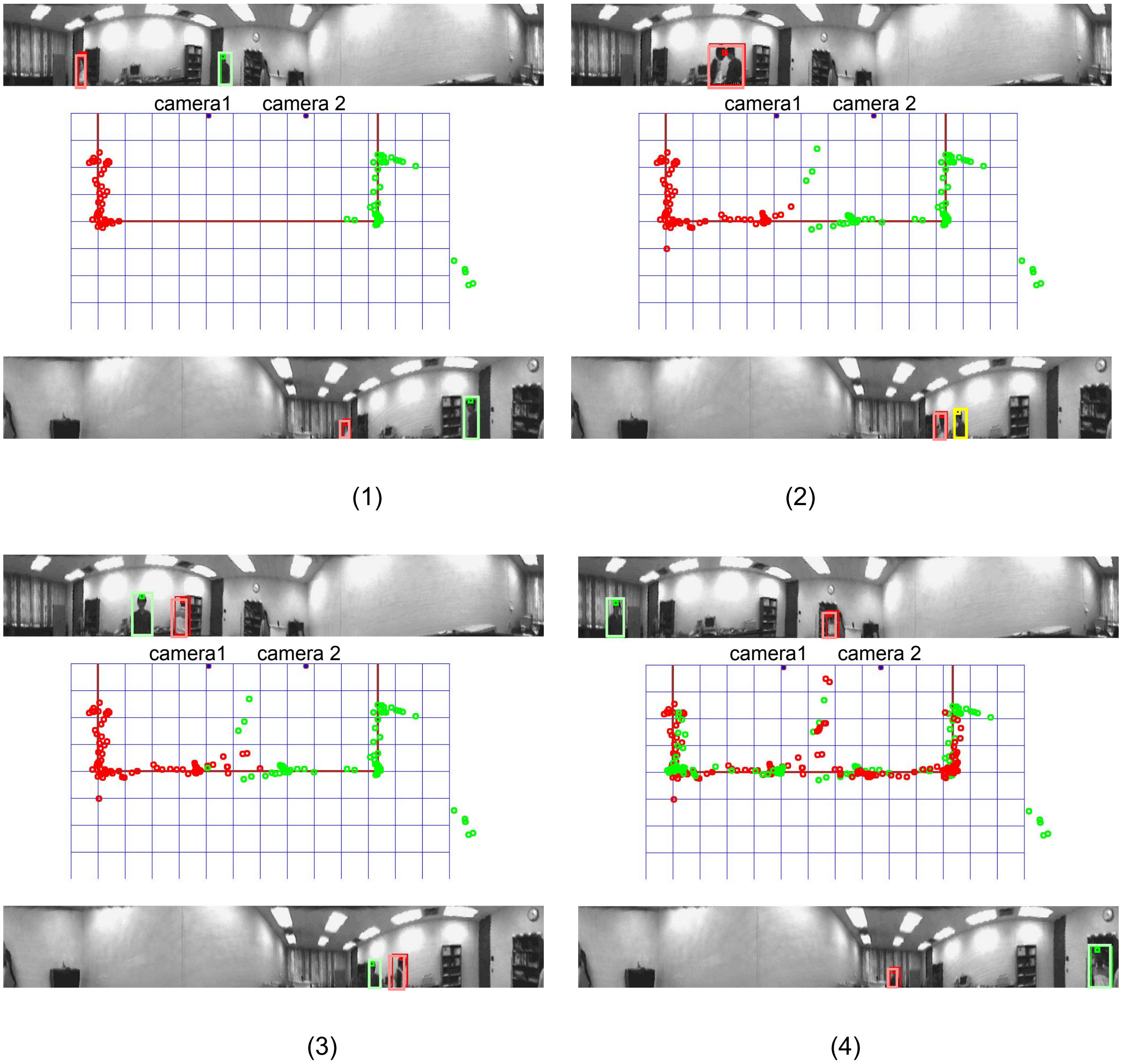


Fig. 12. Panoramic stereo tracking result. The images in (a) and (c) are the panoramic image pair from two panoramic cameras. Each image is actually the corresponding background with the superimposed blob images and their annotations of the blobs. In (c) the real localization results are plotted in the top view of the room where each grid is 50x50 cm^2. Each small circle (red in color version) represents a location of the person in walking. In (d) The theoretical distance error bounds from camera 1 are shown for the same track. The real estimates for best triangulation results validate the theoretical analysis.



(1) (2)

(3) (4)

Fig. 13. Panoramic stereo tracking two people. The four pictures show localizing and tracking results (1) before they met, (2) when they met, (3) after they departed and (4) when they arrived at their goals, out of 214*2 localization results. Each picture of (1) to (4) has the same layout as in Fig. 12. Each small circle in darker tone (red in color version) or in lighter tone (green in color version) represents a location of the corresponding person marked by a bounding rectangle with the same tone (color) in images.

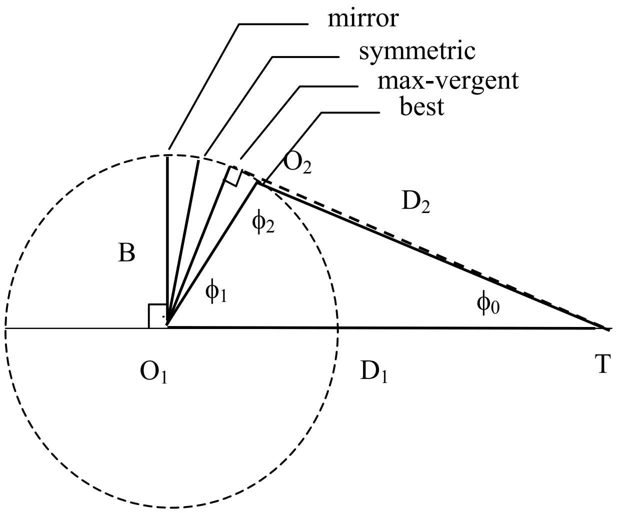


Fig. 14. Best viewpoints given the baseline distance