

Scale Invariant Features for Camera-Planning in a Mobile Trinocular Active Vision System

Aly A. Farag and Alaa E. Abdel-Hakim

Computer Vision and Image Processing Laboratory

University of Louisville, Louisville, KY 40292

E-mail: {farag, alaa}@cvip.uofl.edu

<http://www.cvip.uofl.edu>

ABSTRACT

In this paper, we present a camera-planning approach for a mobile trinocular active vision system. At a stationary version of this system, the sensor planning module calculates the generalized cameras' parameters (i.e., translation distance from the center, zoom, focus and vergence) using deterministic geometric specifications of both the sensors and the objects in their field of view. Some of these geometric parameters are difficult to be predetermined for the mobile system operation. In this paper, a new camera-planning approach, based on processing the content of the captured images, is proposed. The approach uses a combination of a closed-form solution for the translation between the three cameras, the vergence angle of the cameras as well as zoom and focus setting with the results of the correspondences between the acquired images and a predefined target object(s) obtained using the SIFT algorithm. We demonstrate the accuracy of the new approach using practical experiments.

1. INTRODUCTION

A trinocular vision system for 3D reconstruction (CardEye) has been developed by our research team [1]. Instead of using just two cameras as in the known stereo vision systems, CardEye uses three cameras to improve the recovery process and make it more robust. The sensor planning in CardEye aims to determine generalized camera parameters such as position, orientation and optical settings such that the object features are within the field of view and in focus. As a stationary 3D reconstruction system, it is supplied with some information about the object to be scanned and the working conditions. Specifically, the radius of the virtual sphere that contains the object and the distance between the center of that sphere and the cameras should be known before starting the sensor planning process. Then, sensor planning is performed using those parameters combined with the geometrical specifications of the system. Thus, the generated parameters are accurately geometrically calculated.

This work is supported by US Army under grant DABT60-02-P-0063.

In this paper, the stationary trinocular active vision system is developed to be mounted on a mobile robot. The function of the stationary system is extended to not only reconstruct 3D objects in well-known positions, but also to fetch specific target object(s) in the robot's navigation environment, then reconstruct the 3D model. The mobility nature of the proposed system makes the dynamic feeding with the distance between the cameras and the center of the target object extremely difficult or impossible in many cases. Hence, the conventional geometrical sensor planning approach for the stationary system fails. Therefore, in this paper, we present a new camera-planning approach for the mobile system based on detection of a target object in the robot's navigation environment with utilizing the geometric specifications of the system. The proposed approach discards the distance between the center of the virtual sphere containing the object and the cameras as an input parameter.

First, we use the Scale Invariant Feature Transform (SIFT) [2] for detecting the target object. Then, the cameras' parameters are determined such that the number of the correspondent SIFT features of the target object in the three images is maximum. The SIFT approach transforms an image into a large collection of local feature vectors (descriptors). Those SIFT descriptors are robust to image translation, scaling, rotation and partial occlusion and partially invariant to illumination and affine projection. Therefore, the SIFT approach is extremely adequate for the proposed mobile system.

After detecting the target object, or part of it, in one or more image of the three cameras, the geometric information of the system is employed in conjunction with the SIFT results for the system sensor planning with the purpose of maximizing the number of features in the field of view. The main contribution, in this paper, is in using SIFT, as an invariant feature-based approach, in planning the cameras to search and get the best view of a target object for the consequent 3D reconstruction process.

In the next two subsections, we present an overview for the related work in both of the sensor planning and invariant features research fields. In section 2, we give a

detailed description and analysis of the stationary version of our active trinocular vision system. In section 3, we explain our proposed approach and illustrate how the user-input parameters, exist in the stationary system, are discarded. Then we conclude with experimental results.

1.1. Related Work

1.1.1 Sensor Planning

A number of different vision planning systems have been developed in the past years that use a prior information about the observed object and applied sensors to automatically generate sensor parameters that satisfy different vision constraints [3]. The difference among those techniques is in the approach used to determine sensor parameter values. The following techniques are mostly applicable to vision systems that observe known objects in known positions. For example, visual inspection systems, surveillance, monitoring systems, or accurate 3D model reconstruction systems [4].

Several systems use a generate and test approach [3], where sensor positions and settings are chosen and tested to meet the requirements of the task. For active vision systems, a single sensor configuration may not always result in a sufficiently informative view. Therefore, other methods take a synthesis approach, [5-7]. In those sensor planning techniques, the requirements are characterized analytically and sensor parameter values are directly determined from an analytical relationships that satisfy the predefined constraints.

1.1.2 Local Invariant Features Descriptors

Local invariant features descriptors are description vectors which contains some keys that describe a local image region in a manner invariant to spatial transformation and other distortion factors. To be efficient in features matching, the descriptors should be distinctive and at the same time robust to changes in viewing conditions as well as to errors of the point detector. Local photometric descriptors computed at interest points are distinctive, robust to occlusion and do not require segmentation. Many of the recent work has been concentrated on how to make these descriptors invariant to image transformation like scaling, translation and rotation. Also, on making them robust with respect to the changes in the gray level, illumination, brightness ...etc. The main common idea of the approaches based on these descriptors is to

construct invariant "image regions" which are used as support regions to compute invariant descriptors.

There are several researches have been made in this area. For example, Mikolajczyk and Schmid [8] have developed affine invariant interest points with associated affine invariant regions. Tuytelaars and Van Gool [9] construct two types of affine invariant regions, one based on the combination of interest points and edges and the other based on image intensities. In [10], by computing Gaussian derivatives, steerable filters and differential invariants are used for obtaining the local descriptors. Complex filters are proposed in [11] by computing a kernel for each pixel in the image weighted by a Gaussian function.

In [2], Lowe has proposed scale-invariant regions based on local extrema in scale-space built with difference-of-Gaussian (DoG) filters. Specifically, features are detected through a staged filtering approach that identifies stable points in scale space. Image keys are created that allow for local geometric deformations by representing blurred image gradients in multiple orientation planes and at multiple scales. Then, the keys are used as input to a nearest-neighbor indexing method that identifies candidate object matches. Final verification of each match is achieved by finding a low-residual least-squares solution for the unknown model parameters. Lowe has called this method for image feature generation "Scale Invariant Feature Transform" (SIFT).

Mikolajczyk and Schmid [12] have made a performance evaluation for SIFT descriptors [2], steerable filters [10], differential invariants [13], complex filters [11], moment invariants [14] and cross-correlation for different types of interest points [2,15,16].

The evaluation criterion they have used in [12] for two images representing the same scene the detection rate is as shown in Eq(1).

$$P_{correct} = \frac{\text{Number of correct matches}}{\text{Number of possible matches}} \quad (1)$$

They have tested the previously mentioned descriptors according to this evaluation criterion and with respect to various changes in features. They have concluded that the SIFT and steerable filters are better in the case of rotation and scaling. SIFT is the best with affine transforms and steerable filters with illumination changes. Therefore, SIFT is the most suitable algorithm for our system. Also, SIFT has been used in robot localization [17], which is a relevant application to our work.

2. STATIONARY SYSTEM DESCRIPTION

The proposed system is a robot controlled, mobile, trinocular multi sensor, active vision system. Our research team has developed and implemented the stationary version of this system, CardEye (Figure 1), as a flexible and precise tool to mimic the functionality of the human vision system [1]. For sake of improvement of the 3D reconstruction process, CardEye, unlike the known stereo systems, has three cameras. The use of three cameras makes the 3D recovery more robust than just two cameras. CardEye has the basic mechanical properties of active vision platforms - pan, tilt, focus, zoom, aperture, vergence and baseline. To reduce the system complexity and redundancy, the mechanical properties were assigned to the system as a whole and not to each camera. As a consequence, the three cameras are coupled together to perform the same motion, to fixate to a point, or to change the baseline while a robot, carrying the mobile system, moves. Active lenses add the zoom and focus properties to the system.

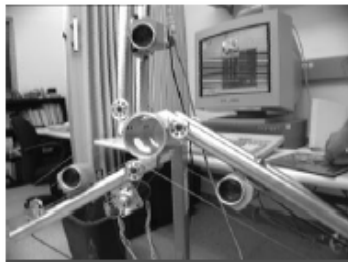


Figure 1: The stationary CardEye trinocular active vision system

The vision module of this active vision system, Figure 2, contains three Hitachi KP-M1 CCD cameras (e.g. c_1, c_2, c_3) with H10x11E Fujinon zoom lenses. The cameras are placed at equal distances from each other and an active lighting device that is the integration of two laser devices, a range finder and a pattern generator is mounted at the center of the trinocular head (O).

The cameras can translate (t) along their mounts to change the baseline distance. At the same time, the cameras can rotate towards each other to fixate to a point in space. This is known as the vergence property. Figure 3 describes the geometry of the vision module in more detail. The object is inside a sphere that has a radius R . By adjusting the vergence angle ($\beta = \tan^{-1}(t/d)$), all cameras can fixate on the same point in 3D space. The system's fixation point is the center C of the sphere. The center of the sphere is at distance d from the origin along the z axis. This distance is called object

distance. The distance from the cameras' optical center to point C is l , and it can be easily calculated ($l = \sqrt{t^2 + d^2}$).

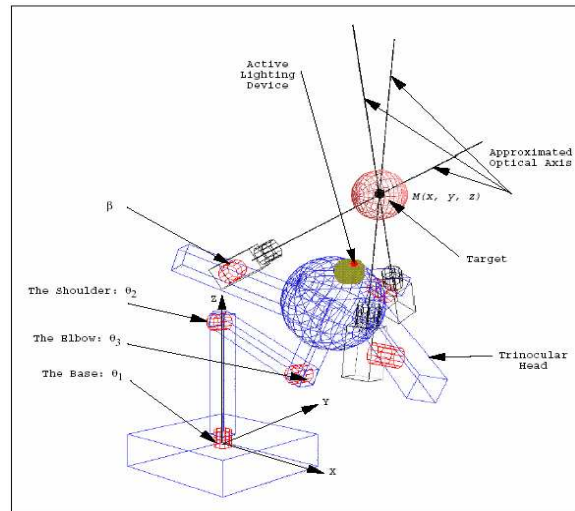


Figure 2: The trinocular active vision module

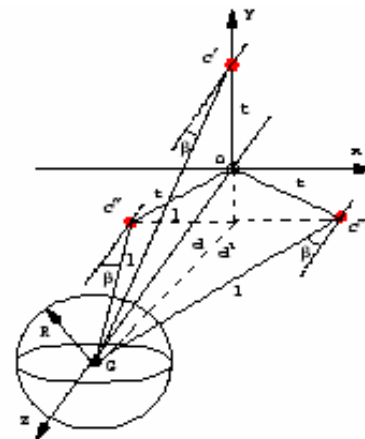


Figure 3: The trinocular active vision module's system geometry. The target for this system is a sphere with radius R .

In order for the sensor planning system to generate satisfactory sensor locations, two main constraints must be satisfied to maximize the effectiveness of 3D reconstruction from three 2D images. These constraints are the *overlap* and *disparity* constraints. In this section, these two system constraints will be discussed in details[1].

2.1.1 Overlap Constraint

In order to understand the overlap constraint, visibility of sphere from a camera's viewpoint has to be determined [18]. When two cameras observe the same sphere, two spherical caps are created. The intersection between these two spherical caps is called the *overlap surface area*. From the our system's geometry, the three cameras are at distance t from the center and equal distances from each other ($\sqrt{3}t$). Therefore, it is sufficient to investigate one of the three equal overlap surface areas between any two cameras as shown in Figure 4.

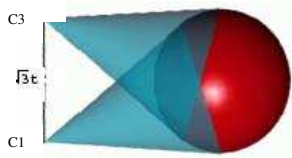


Figure 4: Overlap surface area of two cameras created by the intersection of two spherical caps

From the standpoint of 3D reconstruction, its desirable to maximize the overlap surface area between the two cameras so that the number of corresponding points between the two 2D images is relatively high. To determine the exact size of this overlap surface as a function of t seems to be an extremely complicated problem. From Figure 5, which obtained by projecting the geometry to the plane created by $c_1, c_3,$ and C , it can be seen that the larger δ , the larger the area, and the goal of maximum overlap surface area is achieved

The following is a derivation for the overlap angle δ : Projecting β , to the c_1c_3C plane angle σ is obtained:

$$\sigma = \tan^{-1}\left(\frac{\sqrt{3}t}{2d'}\right) \quad (2)$$

$$\text{where } d' = \sqrt{d^2 - \left(\frac{t}{2}\right)^2} \quad (3)$$

$$\gamma = \frac{\pi}{2} - \sigma \quad (4)$$

$$\delta = 2\gamma \quad (5)$$

Therefore, the overlap surface area which is directly related to the angle, δ , in radians is given as:

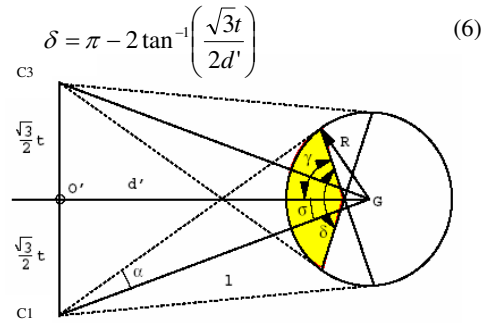


Figure 5: Relationship between δ and overlap surface area. The larger δ , the larger the overlap surface area.

2.1.2 Disparity Constraint

The disparity constraint has an important effect on the planned locations of the sensors. Side-to-side differences in the positions of similar images in the two eyes are called *horizontal disparities* and can produce a compelling sensation of three-dimensionality [19, 20]. The total angular disparity of the two cameras is defined as $\eta = 2\phi$ [19] and can be seen in Figure 6.

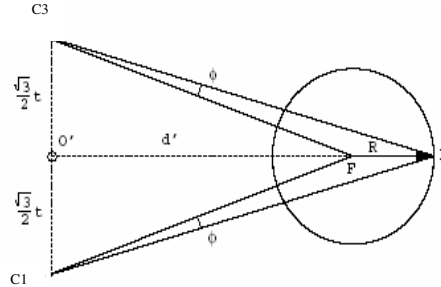


Figure 6: Total angular disparity, η . The value for η is defined as 2ϕ .

The following is a derivation for η :

$$\phi = \tan^{-1}\left(\frac{\sqrt{3}t}{2d'}\right) - \tan^{-1}\left(\frac{\sqrt{3}t}{2(d'+R)}\right) \quad (7)$$

For small angles, the tangent of an angle is equal to the angle itself in radians

$$\phi \approx \frac{\sqrt{3}t}{2d'} - \frac{\sqrt{3}t}{2(d'+R)} \quad (8)$$

Since $d'R$ is relatively small by comparison with d^2 the angular disparity, η in radians is:

$$\eta \approx \frac{\sqrt{3}Rt}{d^2} \tag{9}$$

where t is the distance between the cameras and the origin of the system, d' is the projection of the object distance, and R is the radius of the sphere circumscribing the object.

2.2 Analysis of System Constraints

In this subsection, a complete system analysis and initial result are presented to help solve the system vision sensor planning problem. The goal for the sensor planning system is to maximize the effectiveness of the 3D reconstruction algorithm from one frame. For effective reconstruction, the frames must display adequate depth information and have a fairly large overlap area. These two constraints are in conflict. While the overlap constraint is trying to move the cameras as close as possible, the disparity constraint is moving the cameras away from each other. The following analysis will show that the intersections of the corresponding functions provide a satisfactory selection for the optimum.

The parameters of d (object distance) and R (estimated size of object) are inputs given for the sensor planning of the stationary version of the system. The only independent variable is t , the parameter for which the system first has to solve. The proposed solution uses an analytical approach to characterize the equations by carefully changing one parameter at a time. Throughout this analysis the size of the object is set to a constant $R=0.2m$.

2.2.1 Effect of object distance

In Figure 7, the overlap angle and angular disparity curves are plotted for various object distances in the range of 1.2m to 10m, which is an initial estimation for the system's working space. The lenses' minimum focus distance is 1.2m, and the system will be used indoors. A step size for the object distance, d , is selected using the formula $d = 1.2 + 0.8i$, where $i = 0, 1, \dots, 10$. The goal is to maximize these two constraints, but according to the functions, Eq(6) and Eq(9), the overlap angle is decreasing as t increases and has a maximum at $t = 0$, which means that the three cameras are at the same location, i.e., at the origin. On the other hand, the angular disparity function is increasing with t . The intersections of the corresponding functions provide a possible selection for the optimum.

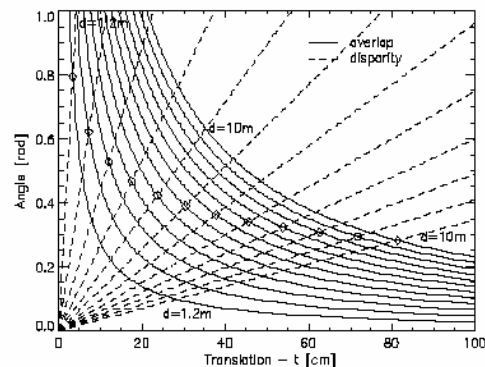


Figure 7: Overlap and disparity curves as object distance changes (1.2m-10m) [1].

2.3 Object Size Constraint

The analysis and optimization in the previous section was performed for a fixed size ($R = 0.2m$) object. Additional analysis is needed to investigate the effect if the size of the object changes. When the size of an object is changed, the field of view angle of the sensor must be changed as well to ensure that the object is within the field of view of the sensor. For every CCD camera, there is a physical limit for this field of view angle which corresponds to the minimum dimension of the active sensor's image array. This limit was determined to be $\alpha = 14.4^\circ$ at the widest angle setting of the active lens attached to the CCD camera of the vision system. Using this angle, the maximum radius of an object at given distance is calculated and presented in Table 1.

The vision system is designed to handle objects with sizes in the range of $0.2m \leq R \leq 1.0m$. As the object size is increased, the workspace in which to place this object decreases. On the other hand, when the workspace decreases, the overlap and disparity constraints are improved. Since the range for the object distance is shorter, the recalculated data for the overlap angle is a subset of the initial data presented before. This still results in a high value and even less difference in the overlap angle.

Table 1: Illustration of closest possible placement of various size objects [1]

$D [m]$	$R_{max} [m]$	Size range for $R [m]$
1.200	0.329	0.2-0.3
1.925	0.528	0.3-0.5
2.650	0.727	0.5-0.7
3.375	0.926	0.7-0.9
4.100	1.125	0.9-1.0

Figure 8 shows the final relations between t, R and d according to following formulas:

Case 1: $0.2m \leq R \leq 0.3m, 1.200m \leq d \leq 7.0m$

$$t = 0.005622d^2 + 0.04068d + 0.04125 [m]$$

Case 2: $0.3m < R \leq 0.5m, 1.925m \leq d \leq 7.0m$

$$t = 0.005812d^2 + 0.04702d - 0.01307 [m]$$

Case 3: $0.5m < R \leq 0.7m, 2.650m \leq d \leq 7.0m$

$$t = 0.006205d^2 + 0.05530d - 0.09068 [m]$$

Case 4: $0.7m < R \leq 0.9m, 3.375m \leq d \leq 7.0m$

$$t = 0.006882d^2 + 0.06668d - 0.20372 [m]$$

Case 5: $0.9m < R \leq 1.0m, 4.100m \leq d \leq 7.0m$

$$t = 0.007990d^2 + 0.08380d - 0.37802 [m]$$

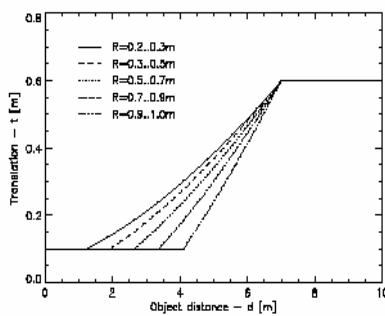


Figure 8: Final results of active sensor planning for the translation parameter t .

Using this solution for t , parameters such as vergence angle β , and field of view angle α can be easily solved according to first equations. These parameters are applied to the system directly using electrical stepper motors through a controlling device.

3. SIFT FOR CAMERA-PLANNING

As mentioned before, the geometrical sensor planning approach fails with the mobile trinocular system. The main reason for this failure is the difficulty of feeding the system with the parameter d , the distance between the camera and the center of the target object. This difficulty comes from the dynamic behavior of the system, which lead to continuous changes in that parameter. This case needs a continuous human intervention to continuously modify the distance values. Clearly, this is impossible from the practical point of view.

Our goal here is to discard the distance parameter from the sensor planning process. Recalling one of the main objectives of the sensor planning process, is to maximize the number of correspondences in the acquired images. On the other hand, one of the goals of the mobile system navigation is to find a specific

object(s) in the navigation environment. The combination of these two goals together creates the motivation to look at the contents of the acquired images as the direct goal of the sensor planning process.

Thus, the main goal of the proposed sensor planning approach can be rewritten as the determination of the system parameters that maximizes the number of correspondences between a target object image and the acquired images.

The invariant features of the target object(s) are extracted and the SIFT descriptors are built for each of those features. The process of extracting the features of the target object(s) is performed off-line. During the navigation of the robot carrying the system, features of the acquired images are extracted and a SIFT matching is performed to those features of the target object. To make sure that the target object appears in each of the acquired three images, the normalized Euclidian distance between the ideal matching case and the present case is evaluated, as an objective function to be minimized, according to Eq(10).

$$f = \frac{1}{\sqrt{3N}} \sqrt{(N-n_1)^2 + (N-n_2)^2 + (N-n_3)^2} \quad (10)$$

where

f : the normalized Euclidian distance

N : The number of features in the target object

n_1, n_2, n_3 : The number of matched features between the target object and the first, second and third image, respectively.

The minimum value of f is zero. It is reached when all the features of the target object appear in each of the three images. So, it is the ideal case. The worst case is for $f=1.0$. This means that no part of the object appears in the field of view.

During navigation, when f goes below a certain threshold, this means that a part of the target object appeared in field of view. Then, the sensor planning process is triggered.

At the beginning, the system assumes that the of the object lies at the minimum possible distance (1.2m) from the cameras. Then, the system parameters are adjusted and f is reevaluated. Similarly, f is evaluated at different system parameters settings that correspond to different d 's. The optimum system parameters are obtained at the minimum f or at a certain stopping threshold.

3.1 Limitations and constraints

In this section, we explore some limitation and constraints on the proposed sensor planning approach. The first constraint that should be taken into account is the object size, R . The target object size is supposed to be known. In the case that the object size is less than the closest distances specified in Table (1), it is set to the closest value in those ranges. This case occurs when a big object placed at a relatively small distance, then some parts of that target object will be missed.

The second limitation of the proposed approach is the case that the center of the target object is far away from the z-axis. In this case, a part of the target object appears in just one or two images. To overcome this case, three secondary objective functions are estimated as shown in Eq(11)

$$f_i = \frac{1}{\sqrt{3N}} \sqrt{(N - n_i)^2} \quad (11)$$

where

f_i : The normalized Euclidian distance for the i^{th} image

n_i : The number of matches between the i^{th} image and the target object

When one or more of f_i 's go below a sub threshold, some navigation actions are to be taken by the carrying robot to get the center of the target object as close as possible to the z-axis. For example, turning right or left to get the object in the field of view of the three cameras.

4. EXPERIMENTAL RESULTS

In this subsection, sample experimental results for the proposed sensor planning algorithm developed for the mobile trinocular active vision system are presented. The system is tested on a target object shown in Figure 9. It is placed at 2.5 m from the cameras. According to the SIFT feature extractor, it has 728 points of interests. The objective function of Eq(10) is evaluated at different system parameters' values that correspond to different virtual object distances. Figure 10 shows the values of the objective function at different values of virtual target object distance. It is clear that the value of the objective function is minimum, i.e. the number of matches is maximum, at $d=2.5$ m, which is the same as the ground truth. The minimum value of f equals 0.37 not zero. This means that some parts of the target object don't appear in the captured images. This can be seen in Figure 11, which shows the acquired images of with system parameters corresponding distances of 7.0, 4.0, 2.5 m, respectively.



Figure 9: The target object

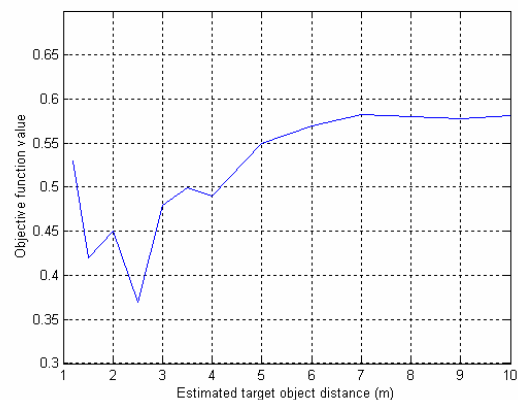


Figure 10: Objective values vs the estimated distance

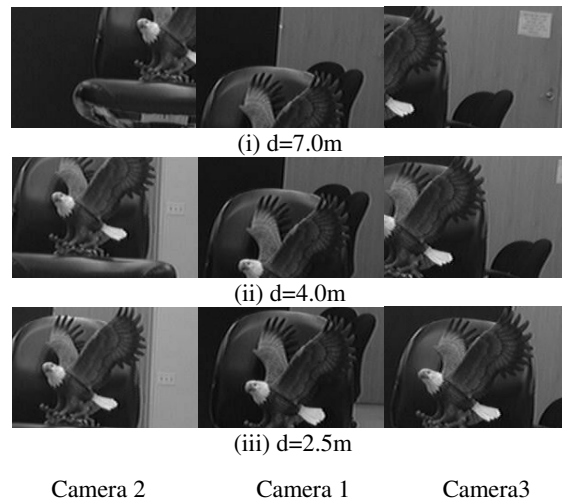


Figure 11: Samples of the captured images

5. CONCLUSION

In this paper, an algorithm has been presented to solve the sensor planning problem for a mobile trinocular,

active vision system. This algorithm used a combination of a closed-form solution for the translation between the three cameras, the vergence angle of the cameras as well as zoom and focus setting with the results of the correspondences between the acquired images and a predefined target object(s) obtained using the SIFT algorithm. Using the proposed approach, two goals are achieved, the first goal is to detect the target object (s) in the navigation field. The second goal is setting the cameras in the best possible position with respect to the target by maximizing the number of correspondences between the target object and the acquired images. After applying sensor planning, the results showed that the images captured by the three cameras displayed adequate depth information and had a fairly large overlap area. Therefore, the goals of the sensor planning were satisfied.

6. FUTURE WORK

In future work, the process of sensor planning can be speeded up by utilizing the contents of the captured images to guide the system to the optimum settings. For example, possible modifications include the use of adaptive learning techniques for estimating the system parameters.

7. REFERENCES

- [1] Elsayed Hemayed, Moumen Ahmed and Aly Farag, "CardEye: A 3D Trinocular Active Vision System," Proc. 3rd IEEE Conference on Intelligent Transportation Systems (ITSC'2000), Dearborn, Michigan, pages: 398-403, October 2000.
- [2] D. G. Lowe, "Object recognition from local scale-invariant features", *ICCV*, pp. 1150-1157, 1999.
- [3] K. Tarabanis, R. Y. Tsai and P. K. Allen, "A survey of sensor planning in computer vision," *IEEE Transactions on Robotics and Automation*, Vol. 11, No. 1, 86-104, February 1995.
- [4] Peter Lehel, Elsayed E. Hemayed, Aly A. Farag, "Sensor Planning for a Trinocular Active Vision System", IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'99), Fort Collins, Colorado, pp. 306-312, June 1999.
- [5] D. J. Cook, P. Gmytrasiewicz, and L. B. Holder, "Decision-theoretic cooperative sensor planning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 10, 1013-23, October 1996.
- [6] I. Stamos and P. K. Allen, "Interactive sensor planning," *Proceedings of the 1998 IEEE International Conference on Computer Vision and Pattern Recognition*, 489-494, 1998.
- [7] E. Trucco, M. Umasuthan, A. M. Wallace and V. Roberto, "Model-based planning of optimal sensor placements for inspection," *IEEE Transactions on Robotics and Automation*, Vol. 13, No. 2, 182-193, April 1997.
- [8] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector", *ECCV*, pp. 128-142, 2002.
- [9] T. Tuytelaars and L. Van Gool, "Wide baseline stereo matching based on local, affinely invariant regions", *BMVC*, pp. 412-425, 2000.
- [10] W. Freeman and E. Adelson, "The design and use of steerable filters", *PAMI*, 13(9):891-906, 1991.
- [11] F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets", *ECCV*, pp. 414-431, 2002.
- [12] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors", *CVPR*, pp. 257-263, 2003.
- [13] J. Koenderink and A. van Doorn, "Representation of local geometry in the visual system", *Biological Cybernetics*, 55:367-375, 1987.
- [14] L. Van Gool, T. Moons, and D. Ungureanu, "Affine / photometric invariants for planar intensity patterns", *ECCV*, pp. 642-651, 1996.
- [15] C. Harris and M. Stephens, "A combined corner and edge detector", *Alvey Vision Conference*, pp. 147-151, 1988.
- [16] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points", *ICCV*, pp. 525-531, 2001.
- [17] S. Se, D. Lowe and J. Little, "Local and global localization for mobile robots using visual landmarks", *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pp. 414-420, 2001.
- [18] W. F. Taylor, *The Geometry of Computer Graphics*. Wadsworth & Brooks, 1992.
- [19] I. P. Howard and B. J. Rogers, *Binocular Vision and Stereopsis*. Oxford University Press, 1995.
- [20] G. Xu and Z. Zhang, *Epipolar Geometry in Stereo, Motion and Object Recognition*. Kluwer Academic Publishers, 1996.