IMAGE CONTENT-BASED ACTIVE SENSOR PLANNING FOR 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 sensor planning approach for a mobile trinocular active vision system. At the stationary state (i.e., no motion) the sensor planning system calculates the generalized cameras' parameters (i.e., translational 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 sensor 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 settings 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 can be performed using those parameters combined with the geometrical specifications of the system. Thus, the generated parameters are accurately geometrically calculated.

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 system 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 sensor 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 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 image or more 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.

1.1. Related work

A number of different vision planning systems have been developed in the past years that use prior information about the observed object and applied sensors to automatically generate sensor parameters that satisfy different vision constraints [3]. The difference between 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, surveillance, monitoring systems, or accurate 3D model reconstruction systems [1]. Several systems use a generate-and-test

0-7803-8554-3/04/$20.00 ©2004 IEEE.

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 these sensor planning techniques, the requirements are characterized analytically and sensor parameter values are directly determined from an analytical relationships that satisfy the predefined constraints.

2. THE 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, 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.

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 1 describes the geometry of the vision module of CardEye 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 ($l = \sqrt{t^2 + d^2}$).

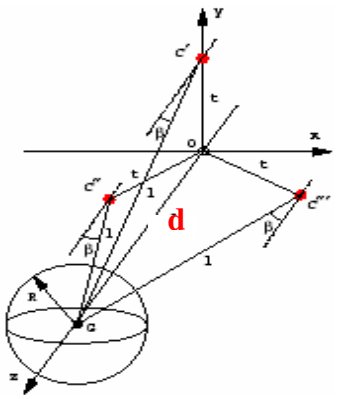


Figure 1: The trinocular active vision module's system geometry. The target for this system is a sphere with radius R.

3. SENSOR PLANNING IN CARDEYE

For sensor planning, 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 [4]. For the quality of the 3D reconstruction, both of the contradicting constraints should be maximized.

In Figure 2, the overlap angle and angular disparity curves are plotted for various object distances in the range of 1.2m to 10m, which is estimation for the system's working space. The goal is to maximize these two constraints. The intersections of the corresponding functions provide a possible selection of t and β for the optimum. The detailed derivation of these curves is shown in [4].

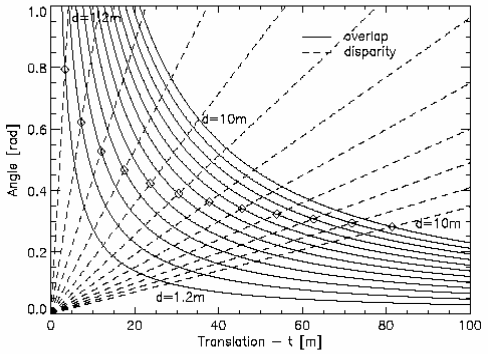


Figure 2: Overlap and disparity curves as object distance changes (1.2m-10m).

4. 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 points of interest 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. In addition, efforts have been directed to make these features 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.

Mikolajczyk and Schmid [8] 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 SIFT as scale-invariant regions based on local extrema in scale-space constructed 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 in a way such that local geometric deformations are allowed by representing blurred image gradients in multiple orientation planes with 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. Mikolajczyk and Schmid [12] have made a performance evaluation for SIFT [2] versus other invariant feature descriptors. They concluded that SIFT is the best with respect to the scale, rotation, and illumination changes.

5. IMAGE-CONTENT BASED SENSOR 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. Obviously, this is impossible from the practical point of view.

Our goal here is to discard the distance parameter as an input to the sensor planning process. Recalling one of the main objectives of the sensor planning process, is to maximize the number of correspondences among the acquired images (overlap). 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, as much as possible, 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. (1).

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

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 center of the virtual sphere containing the object *C* lies at the minimum possible distance (d=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. This procedure is a "generate and test [3]" like approach.

5.1 Limitations and constraints

In this section, we explore a limitation on the proposed sensor planning approach. This limitation 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. (2)

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

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 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. The procedure that performs these actions is in the navigation process and beyond the scope of this paper.

6. EXPERIMENTAL RESULTS

In this section, 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 3. It is placed at 2.5 m from the cameras. According to the SIFT feature extractor, it has 336 points of interests. The objective function of Eq. (1) is evaluated at different system parameters' values that correspond to different assumed virtual distances of the object. Figure 4 shows the values of the objective function at different values of the target object's virtual distance. It is clear that the value of the objective function is minimum, i.e. the number of matches is maximum at d=2.5m, which is the same as the ground truth. The minimum value of *f* equals 0.4419 not zero. This means that some parts of the target object don't appear in the captured images. This can be seen in Figure 5, which shows the acquired images of with system parameters corresponding to distances of 7.0, 4.0, 2.5 m, respectively.

Figure 3: The target object

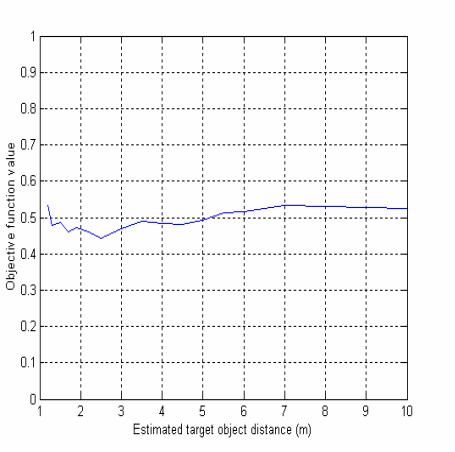
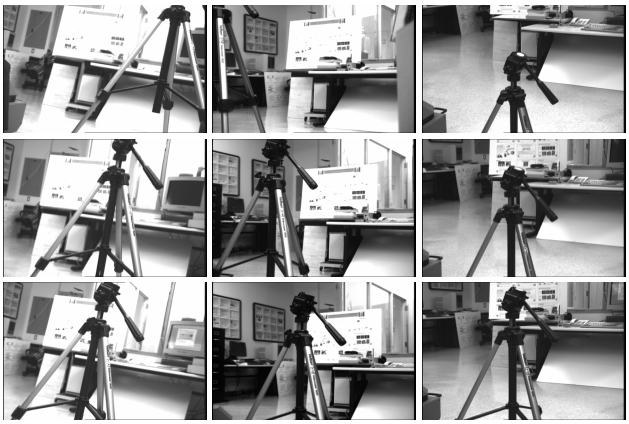


Figure 4: Objective values vs the estimated distance



Camera 1 Camera 2 Camera3

Figure 5: Samples of the captured images
Top: d=7.0m Middle: d=4.0m Bottom: d=2.5m

7. 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. The ultimate goal for the proposed algorithm was to maximize the effectiveness of the 3D reconstruction from one frame. 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.

8. 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, may be used.

9. 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, pp. 398 -403, October 2000.

[2] D. G. Lowe, "Object recognition from local scale-invariant features", *ICCV'99*, 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 1EEE 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'03*, pp. 257-263, 2003.