3D ENVIRONMENT RECONSTRUCTION USING MULTIPLE MOVING STEREOVISION SENSORS

Sergiu Nedevschi*[1]*, Tiberiu Marita*[1]*, Radu Danescu*[1]*,
***Florin Oniga*[1]*, Dan Frentiu*[1]*, Ciprian Pocol*[1]**
[1] *Technical University of Cluj-Napoca, Computer Science Department*
E-mail:{Sergiu.Nedevschi;Marita.Tiberiu;Radu.Danescu}@cs.utcluj.ro

ABSTRACT: A method for 3D environment reconstruction based on stereovision sensors will be presented. The system is structured in a distributed fashion. Each sensor is composed from a pair of video cameras and an image-processing device, which is able to perform real-time stereo processing. The stereo-rig assembly is mounted on a computer driven pan-tilt unit, which allows two degrees of freedom in order to obtain a better coverage of the scene even with a reduced number of stereovision sensors. The output of one stereo sensor is a list of cuboids, describing the part of the environment that it sees. All the sensors must report the cuboids in the same coordinate system. The cuboids are communicated using a symbolic representation and a standard network communication protocol. Each sensor output is sent to a fusion computer, which assembles the complete description of the 3D environment. As possible employment of the system we can enumerate: warehouse activity planning, surveillance of harbors, parking lots, etc.
Keywords: stereovision, camera calibration, distributed computation, sensor fusion.

1. INTRODUCTION

Having a good 3D description of an environment is essential if we want to employ any kind of automated control over it. Stereovision is becoming more and more popular as a 3D measurement tool, having the advantage of being a passive method and also of providing a rich amount of 3D data.

Each stereovision sensor is limited by its own field of view. The field of view of each sensor can be considerably improved by placing the stereo-rig assembly on a pan tilt device, which allows two simultaneous rotations: tilt (pitch) and pan (yaw).

A complete description of a 3D scene is difficult to be achieved by only one sensor, due to the sensor's position, occlusions of the objects, etc. For that reason, data fusion from several sensors, placed strategically around the scene is a possible solution. Also, the accuracy of a sensor reading is not uniform in any point of its working range, and therefore by using multiple sensors and integrating their readings one can improve the uniformity of the reconstruction resolution over the whole working space [1].

For the sensor fusion algorithms there are two options: fusion of the 3D points reconstructed by basic stereovision or fusion of the high-level objects resulted from grouping of the 3D points at sensor level. Choosing one of the available approaches determines the structure and functionality of the whole system. The first approach needs a unique point grouping process, but exhibits a higher communication burden between the sensors and the fusion module. The second approach requires low communication bandwidth, but the point grouping must be performed locally by each sensor [1].

2. THE SENSORIAL SYSTEM ARCHITECTURE

The system consists of "*n*" Stereovision Sensors linked by TCP connection to the Sensor Fusion Module (SFM). The Stereovision Sensors must be placed around the space of interest in such a way that a good coverage of the scene is accomplished. This way, each sensor has a different view of the 3D scene, and issues as hidden object facets or object occlusions are easier to treat.

A Stereovision Sensor consists of a pair of cameras, mounted on a rig, linked to its host computer. The stereo-rig assembly is placed on a pan tilt device, which allows two simultaneous rotations: tilt (pitch) and pan (yaw) (fig. 1). The rotations of the pan-tilt device are fully controllable by computer interface with precisions up to 0.013 degrees [2] and therefore the instantaneous position of the stereovision sensor in the unique world coordinates system is known.

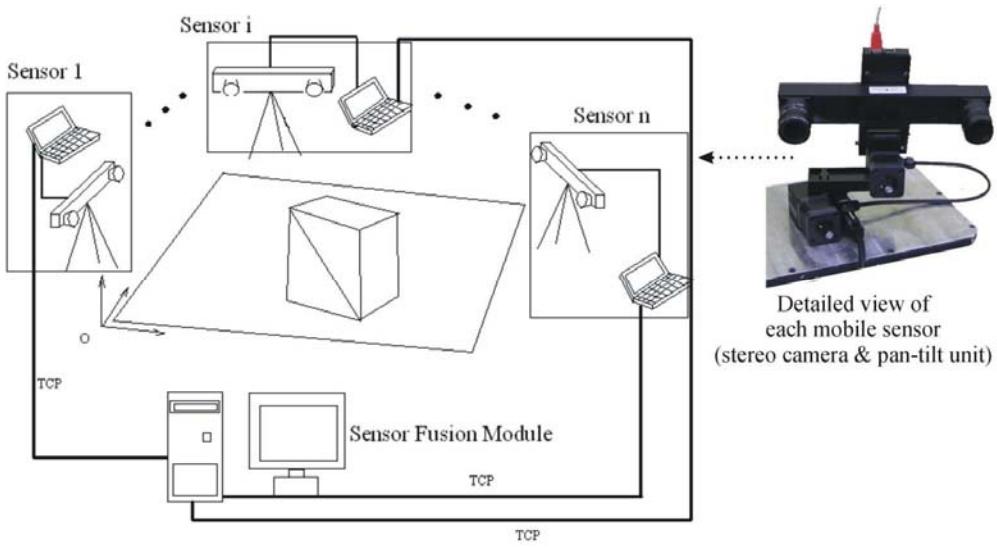

Fig. 1. The architecture of the sensorial system.

The 3D stereo reconstruction is performed on each host computer in processing cycles completed for each synchronously acquired image pairs. The reconstructed 3D points, expressed in the unique world coordinates system (fig. 1) and grouped into 3D objects (cuboids) represent the sensor's output.

Calibration of the Stereovision Sensor is required to completely determine the geometry of the cameras. In order to provide the results relative to the same coordinates system, each resting position (ZERO position) of the pan-tilt unit must be calibrated relative to the unique 3D world coordinates system.

The SFM has processing and communication responsibilities. The processing responsibilities consist of fusion of the sensorial objects. The communication responsibilities consists of handling the connections with the Stereovision Sensors and with the clients, which can be a viewing application, another SFM, etc. The SFM acts like the synchronization master for the sensor array, ensuring that all sensors capture the scene at the same time. It is also responsible for delivering the information to client applications. The result is delivered in the form of 3D cuboids expressed in the unique world coordinates system.

3. STEREOVISION SENSORS CALIBRATION

In order to reconstruct and measure the 3D environment using stereo cameras, the cameras must be calibrated. The calibration process estimates the camera's intrinsic parameters (which are related to its internal optical and geometrical characteristics) and the extrinsic ones (which are related to the 3D position and orientation of the camera relative to a global world coordinate system).

The intrinsic parameters of each camera are calibrated individually. The estimated parameters are the focal length and the principal point coordinates and the lens distortions. The parameters are estimated by minimizing the projection error from multiple views of a set of control points placed on a coplanar calibration object with known geometry [3].

For the benefit of the point fusion algorithm the calibration of the extrinsic parameters must be performed in the same world coordinate system (a unique coordinates system belonging to the scene), and must be very precise. If the precision requirements are not met, the set of points from different sensors will have different meaning, and their fusion will be erroneous. Since the stereo rig is mounted on the pan-tilt unit the extrinsic parameters calibration procedure have to estimate the initial resting (ZERO) position of the stereo-rig. During the pan-tilt movements the instantaneous absolute values of the extrinsic parameters are updated by knowing the rotational offsets programmed by the pan-tilt unit. The "ZERO position" values of the extrinsic parameters of the cameras are estimated by minimizing against the extrinsic parameters the projection error for a set of 3D control points with measured coordinates in a world reference system. For the specific setup of the current application having multiple stereovision sensors, each stereo pair of cameras is calibrated using a set of control points measured in a unique world coordinates system - the coordinate system of the scene (fig. 2) [1,3]. The procedure estimates the absolute extrinsic parameters of each camera in the unique world coordinate system for the initial (ZERO/resting) position of the pan-tilt unit. A pair of translation vectors $\mathbf{T}_K^{Lo}$ and $\mathbf{T}_K^{Ro}$ and a pair of rotation vectors/matrices $\mathbf{r}_K^{Lo}/\mathbf{R}_K^{Lo}$ and $\mathbf{r}_K^{Ro}/\mathbf{R}_K^{Ro}$ are obtained for each camera pair "k".

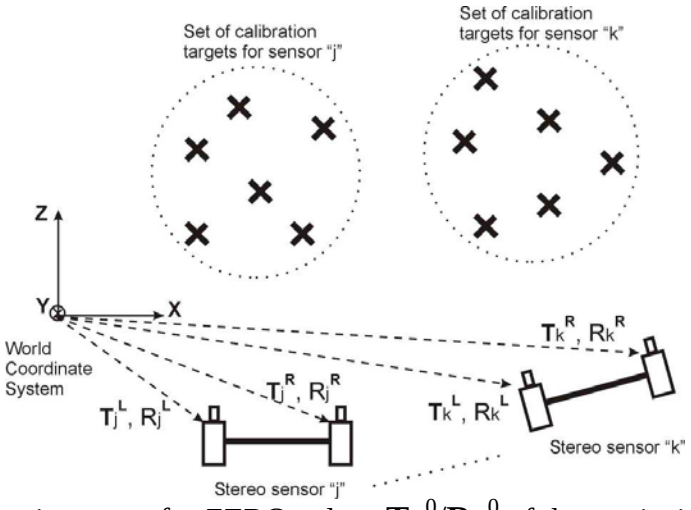

Fig. 2. Calibration setup for ZERO values $\mathbf{T}_K^0/\mathbf{R}_K^0$ of the extrinsic parameters

4. THE 3D STEREO RECONSTRUNCTION

In order to perform the 3D stereo reconstruction we have first to obtain the instantaneous absolute extrinsic parameters of each stereovision sensor.

For every absolute yaw and pitch angle rotations of the pan-tilt unit, relative to its ZERO position, at the moment of time "*t*": $\delta\mathbf{r}_K^t = [\delta r_X^t, \delta r_Y^t, 0]$, a pair of rotation and translation vectors (derived from the assembly's cinematic model) are obtained:

$$\begin{cases} \delta\mathbf{r}_K^t = [\delta r_X^t, \delta r_Y^t, 0] \\ \delta\mathbf{T}_K^t = [\delta T_X^t, \delta T_Y^t, \delta T_Z^t] \end{cases} \tag{1}$$

The instantaneous absolute extrinsic parameters of each stereovision sensor are updated for each camera using the following equations:

$$\begin{cases} \mathbf{T}_K^t = \mathbf{T}_K^0 + \delta\mathbf{T}_K^t = \mathbf{T}_K^0 + [\delta T_X^t, \delta T_Y^t, \delta T_Z^t] \\ \mathbf{r}_C^t = \mathbf{r}_K^0 + \delta\mathbf{r}_K^t = \mathbf{r}_K^0 + \delta\mathbf{r}_K^t = \mathbf{r}_K^0 + [\delta r_X^t, \delta r_Y^t, 0] \\ \mathbf{R}_K^t = Rodrigues(\mathbf{r}_K^t) \end{cases} \tag{2}$$

where *Rodrigues()* is a set of functions for conversions from a rotation vector to its corresponding rotation matrix and vice versa [4].

The stereo reconstruction algorithm used is mainly based on the classical stereovision principles available in the existing literature [5]: find pairs of left-right correspondent points and map them into the unique 3D world coordinate using the stereo system geometry determined by calibration.

Constraints, concerning real-time response of the system and high confidence of the reconstructed points, were used in the stereo correlation process [1]. After this step of finding correspondences, each left-right pair of points is mapped into a unique 3D point [3].

The result of reconstruction is a set of 3D points that must be clustered into objects. The grouping is performed mainly based on the local density of the points and the vicinity criteria: a local group of points must be dense enough to be considered as candidate and two points are considered to be in the same group if they are close to each other. Both these criteria are adapted to the fact that the density of reconstructed points per object decreases with the distance (due to the perspective projection) and their positioning error increases with the same distance. When we are dealing with known objects shapes (ex. surveillance of a warehouse: containers have parallelepiped shape), additional shape-constraints can be imposed to have a better grouping. For each cluster of points, the circumscribing cuboid is built as specified in the environmental model [1]. For each vertex of a cuboid the confidence factor is evaluated based on the density of neighboring 3D points. The orientation of each object is also inferred.

5. SENSOR FUSION

The main simplification of the fusion problem comes from the fact that the cuboids are defined in the same coordinate system, and therefore no geometrical transformations are necessary in order to compare their position. The fusion algorithm [1] can be summarized as follows. Two objects detected by two different sensors are joined if they occupy the same 3D space (the distance between their centers of mass must be smaller then their maximum "radius"). If the joining condition is satisfied, the objects are fused, by combining the corresponding vertices of the two objects into a resulting vertex. The combination of the object's vertices can be made using two approaches: as a weighted sum, using the confidence level as the weight, or we can take as valid coordinate the coordinate with the highest confidence. For the first variant, the motivation is that each observation adds some information, and should not be disregarded. For the second variant the reason is that we presume that we distribute the vision sensors in such a manner that they cover the scene as best as possible, and thus each point of one object is best seen by one of the sensors (and this is expressed by a high degree of confidence of that object point reconstructed by that particular sensor), and therefore its observation is accurate enough, and there is no need to add other information, which could be in fact measurement noise.

6. RESULTS

For algorithm testing we have used two stereovision sensors. Since only one pan-tilt unit was available for experiments, one stereo sensor was placed on fixed tripod and the second one on the mobile pan-tilt device. The two stereo sensors were calibrated using the method described in the calibration section, using a common coordinates system. For the mobile sensor, the ZERO position was calibrated. Then pan-tilt movements were generated by the moving sensor's host computer. For each new position of the mobile stereovision sensor the 3D scene was reconstructed and the detected objects represented as cuboids.

The perspective views of the scene for each stereovision sensor at the moment of time *"t"* are presented in the left side of figures 3.a and 3.b. The reconstruction results for each stereovision sensor is presented as a bird-eye view of the scene in the right part of the same images, and as white cuboids projected on the original perspective image.

The sensor results were sent to the fusion unit, which integrated the data into a fused scene description. The function used to combine the objects was the one that selects the coordinate of the higher confidence. The weighted average function was also tested, but the results seem of lower quality. The results are displayed in fig. 3.c as a bird-eye view. The final result corresponds to the aim of the algorithm: combining together the scene description of different sensors and refining the measurements of each sensor against each other, in the case where the same object is viewed by more than one sensor.

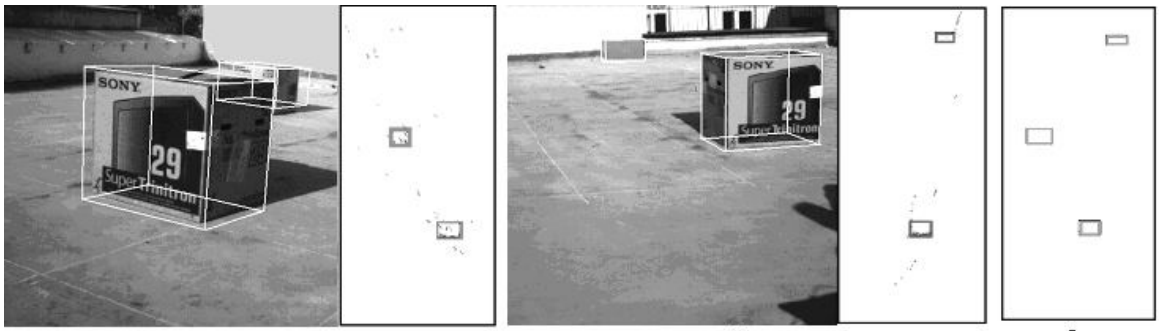

Fig. 3. Stereo reconstruction and object fusion results

7. CONCLUSIONS

A method for extracting the 3D scene description from multiple moving stereovision sensors has been presented. The stereovision sensors are able to perform real-time image pair processing and extract 3D points of the environment, points that they subsequently group into high-level cuboids. The communication of the cuboids to a fusion system is performed using a minimum bandwidth. Fusion is performed at cuboid level, and a complete description of the scene is obtained. The fused description has the advantage of increasing global field of view by uniting the fields of view of each moving sensor, and the advantage of refining the description of individual objects, if they are viewed by more than one sensor.

A point-level fusion approach is to be considered as an alternative approach, and the results compared to the current method, both in terms of reconstruction accuracy and overall time performance.

As further extension of the presented method would be the reconstruction of the 3D environment when the stereovision sensors are mounted on mobile robots. In such a case each sensor would have different poses relative to the world coordinates system, which can be static or also moving. The sensors movements would have more degrees of freedom as in the case of the pan-tilt unit and more complex tracking and fusion algorithms must be employed.

8. REFERENCES

[1] S. Nedevschi, R. Danescu**,** D. Frentiu, T. Marita, F. Oniga, C. Pocol, **Real-Time Extraction of 3D Dynamic Environment Description Using Multiple Stereovision Sensors**, Proceedings of International Conference on CCCT 2003, Orlando, Florida, 29 July – 1 August, 2003, Volume 3, pp. 520-524

[2] **Direct Perception Web Site**, http://www.dperception.com, 2003.

[3] S. Nedevschi, T. Marita, R. Danescu, F. Oniga, D. Frentiu, C. Pocol, **Camera Calibration Error Analysis in Stereo Measurements**, *microCAD International Scientific Conference, March 2003, Miskolc, Hungary*, pp. 51-56.

[4] Jean-Yves Bouguet, **Camera Calibration Toolbox for Matlab**, MRL - Intel Corp., http://www.vision.caltech.edu/bouguetj/calib_doc/, 2003.

[5] Trucco E., Verri A, **Introductory techniques for 3D Computer Vision**, New Jersey: Prentice Hall, 1998.