Dynamic Scene Reconstruction For 3D Virtual Guidance.

Alessandro Calbi[1], Lucio Marcenaro[2], and Carlo S. Regazzoni[1]

[1] DIBE, University of Genova, 16145 Genova, ITALY,
carlo@dibe.unige.it,
WWW home page: https://www.isip40.com
[2] TechnoAware S.r.l, Via Greto di Cornigliano 6, 16100 Genova, ITALY,
lucio.marcenaro@technoaware.com,
WWW home page: http://www.technoaware.com

Abstract. In this paper a system is presented able to reproduce the actions of multiple moving objects into a 3D model. A multi-camera system is used for automatically detect, track and classify the objects. Data fusion from multiple sensors allows to get a more precise estimation of the position of detected moving objects and to solve occlusions problem. These data are then used to automatically place and animate objects avatars in a 3D virtual model of the scene, thus allowing users connected to this system to receive a 3D guide into the monitored environment.

1 Introduction

Many algorithms have been studied during the last years for automatic 3D reconstruction [1, 2] from image analysis for application in different fields. In [3] a semi-automatic system is described that is based on a 3D reconstruction of a museum environment, obtained by a stereo vision sensor: proposed system is able to detect interesting events and to guide the users into the museum. A visualization system for ambient intelligence based on an augmented virtual environment that fuses dynamic imagery with 3D models in a real-time display to help observers comprehend multiple streams of temporal data and imagery from arbitrary views of the scene is presented in [4].

One of the fundamental task for an ambient intelligence application is automatic objects tracking and classification. Researchers developed many specific solutions [5] but no optimal algorithm exists to solve the tracking problem in all real situations. As the complexity of the scene increases and occlusions between static and non-static objects occur [6], performances of standard tracking and classification algorithms typically decrease. Multi camera systems have been often used for overcoming the occlusion problem. Collins at al. in [7] propose understanding algorithms to automatically detect people and vehicles, seamlessly track them using a network of cooperating active sensors, determine their three-dimensional locations with respect to a geospatial site model, and present this information to a human operator who observes the system through a graphical user interface.

In this paper a multi sensor system is described that is able to detect, track and classify multiple moving objects. A three-dimensional model of the observed area is automatically updated by the tracking system and dynamic avatars are maintained within the model. Such a system can be effectively used also for virtual guidance of users because it allows to reproduce the entire ambient evolution of the scene and to compute the path that users have to follow to reach their destination.

In section 2 processing modules for detecting and tracking moving objects are described; section 3 shows specific modules for multi-camera supervision, while section 4 deals with proposed 3D viewer and guidance module. Finally results are showed in section 5 and conclusions are drawn in section 6.

2 Detection and Tracking

In order to realistically reproduce a real environment and interactions between moving objects and to re-create them into a 3D model, the ambient intelligence system needs sophisticated modules for image analysis and object tracking and classification. Such processing modules allow the automatic comprehension of semantic contents of the image sequence. The primary objective of the system is the phase of *detection*, that is the automatic identification of the moving objects in the scene (entities perceived as different respect a reference background). Subsequently, the system has to evaluate and follow their position in time (*tracking*), being able to extract suitable information to describe the actions performed by the objects (*classification*) themselves. The last phase, therefore, will consist into recognize behaviors (see Figure 1).

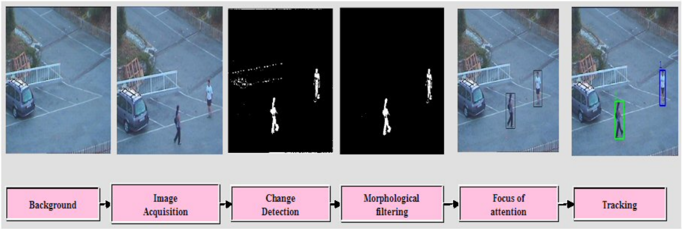


Fig. 1. Scheme of the principal modules of a tracking system.

The algorithms adopted by the system to pursue previous objectives can be subdivided into several logical modules, on the basis of the task they have to complete: in particular, it is possible to subdivide basic processing modules into three different main categories:

- *Low level modules* are responsible of extracting interesting data from acquired raw images (image acquisition, change detection, morphological filter, background updating, focus of attention);
- *Middle level modules* are able to get contextual information previously extracted from video sequence and to derive a semantic description of the observed world (blobs matching, feature extraction);
- *High level modules* are responsible to track objects features to keep the history of the temporal evolution of each blob; through classification algorithms [8] these modules are able to classify detected objects.

3 Multi camera modules

In order to increase the area covered by a single sensor and to manage the situation of occlusion between the blobs a multi camera approach is adopted. The structure of a multi camera system is based on three steps [9]: Data alignment, Data association, State Estimation.

Data Alignment is needed in order to make the data comparable: dealing with video cameras, this step issues are related to:

- Temporal alignment: the sensors are synchronized to compare features referring to the same instant using a NTP (Network Time Protocol) server.
- Spatial alignment: through a joint cameras calibration procedure it's possible to obtain the correspondences between each image plane and the absolute *world coordinates*, exploiting geometric and optic features of each sensor. The calibration procedure is based on the Tsai algorithm described in [10].

Data Association consists on the *m-ary* decision process among the objects in the fields of view of the used cameras. Many different features are extracted to let the system autonomously adapt the association to different situation occurring in the scene. The use of different features has the advantage to extract in every instant and among the others the better discriminating feature, which will be responsible of the greater separation among the classes we are trying to distinguish in the decision process.

In order to be able to manage such different data and obtain a coherent representation, we define independent similarity functions connected to the information obtaining from each feature; each function provides an autonomous similarity coefficient yielding continuous values distributed between 0 and 1. The feature functions are based on measures in the map reference system (in term of position and speed of the blobs) and in the image plane (valuating the shape factor and chromatic characteristics). The results provided by each function leads to define an *Object Similarity Coefficient* (OSC), calculated as the mean value of the previous coefficients. To apply the criterion and choose the correct associations we seek for the highest values for each object in a camera field of view compared to all the objects in all the cameras image planes with a field of view overlapped with the first.

Once data are aligned and objects associated, the *state estimation* phase performs the actual redundant information exploitation: when the single cameras positioning data are available, they can be fused simply through the use of mean values. But when objects are not well separated in the image plane, a little more care must be put in the estimation phase.

In our system we consider 3 cases:

- if the objects to associate are well separated in both the fields of view, we use the position mean value;
- if the objects result occluded in the field of view of one of the sensors, we use the position computed by the other;
- if both the fields of view present occlusions, we apply the location data related to the objects' couple with the *strongest* OSC value in the association phase.

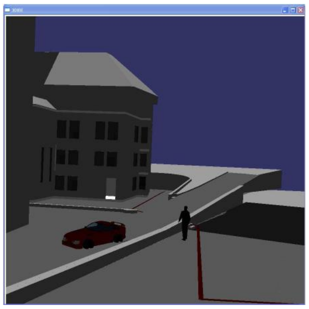


Fig. 2. 3D vision of the scene.

4 3D Viewer and Guide

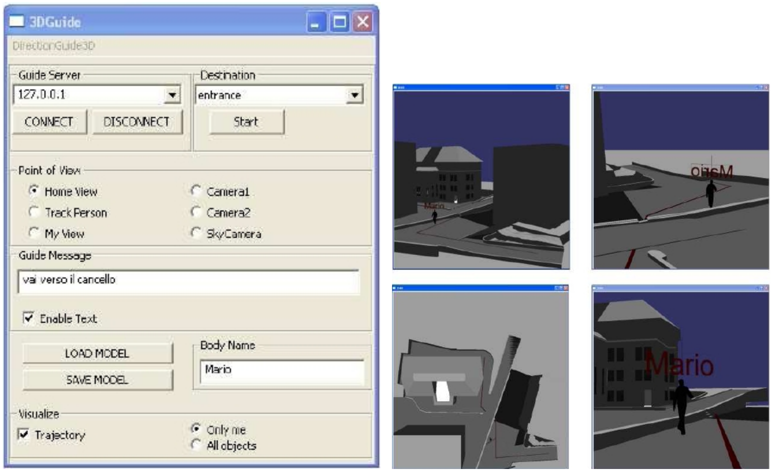
A visual 3D modality has been developed for the appliance of localization, detection and navigation. The idea is to provide to the users entering into the monitored areas a virtual ambient that reproduces in details the rooms and the ambient interested by the system of ambient intelligence. This ambient has the aim to represent in real time isolated zones of the areas where, for instance, the system detects the presence of other users. This result is performed following the next steps:

1. Creation of a three-dimensional map of the ambient of interest;
2. Importation of the model into a 3D ambient engine;
3. Real time acquisition of the spatial coordinates of the objects and them classification;
4. Implementation of virtual cameras;
5. Equipment the system of the required intelligence to evaluate the minimum path from the current position to the destination.

Data provided by the modules of tracking and classification have to be filtered as a consequence of the error's propagation inherent to the image processing. Therefore, these data are stabilized by using Kalman [11] and median filters [12]. Once the position, speed and class of each blob are filtered, it's possible to send them to a server machine which task is to acquire and elaborate data from the sensors and to forward them to the 3D maker. At this time, the system is able to represent the virtual model: figure 2 shows the virtual 3D representation of the scene.

In figure 3(a) is represented the 3DGuide interface where, under the buttons of connection, selection of the destination and of the virtual camera, is presented the textual guidance message. One of the most relevant benefit of the 3D virtual viewer is surely the possibility to change the point of view (figure 3(b)): in this manner, the users can use the 3D model itself placing virtual cameras into the model selecting the best point of view to reach the destination or to see other users moving into the environment.

The three-dimensional vectorial model has been generated by using AutoDesk AutoCAD and 3D Studio Max [13] software using precise measurements of the environment; afterward, this model is imported into a 3D graphical engine. We adopted open-source libraries: the library OpenSceneGraph [14] provides the rules to build the model;



3DGuide

DirectionGuide3D

Guide Server: 127.0.0.1

Destination: entrance

Point of View: ● Home View ○ Track Person ○ My View ○ Camera1 ○ Camera2 ○ SkyCamera

Guide Message: vai verso il cancello

☑ Enable Text

Body Name: Mario

Visualize: ☑ Trajectory ● Only me ○ All objects

(a) 3D guide interface (b) Virtual point of view

Fig. 3. 3DGuide interface and Virtual generated points of view of the scene.

with the library Cal3D [15] each object can perform a lot of movements as walk, run, turn, stand, etc.; eventually the last library used by the proposed system, ReplicantBody [16], allows to animate the human model by integrating [14] and [15].

5 Results

In this section we present the most significant results related to the multi-camera data fusion process and the effects of tracking errors to 3D virtual reality rendering. In first instance, we evaluate the goodness of the strategy of data association examining sequences with 4 mutually occluding moving objects, in the form of association rate confusion matrices: in the principal diagonal cells the rate of correct association is reported while the crossing values in the other cells define the wrong associations. Some associations were discarded defining the data belonging to the NC class.

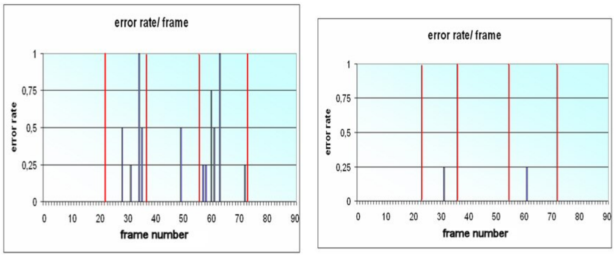
(a) $OSC \equiv f(P)$

	1i	**2i**	**3i**	**4i**	**NC**
1j	90.0	3.3	0.0	0.0	6.7
2j	3.3	87.7	0.0	0.0	9.0
3j	0.0	0.0	85.5	4.5	10.0
4j	0.0	0.0	3.3	91.1	5.6
NC	6.7	9.0	11.2	4.5	

(b) $OSC \equiv f(P, V, S, C)$

	1i	**2i**	**3i**	**4i**	**NC**
1j	97.7	0.0	0.0	0.0	2.3
2j	0.0	97.7	0.0	0.0	9.0
3j	0.0	0.0	93.3	1.1	5.6
4j	0.0	1.1	0.0	98.9	0.0
NC	2.3	1.2	6.7	0.0	

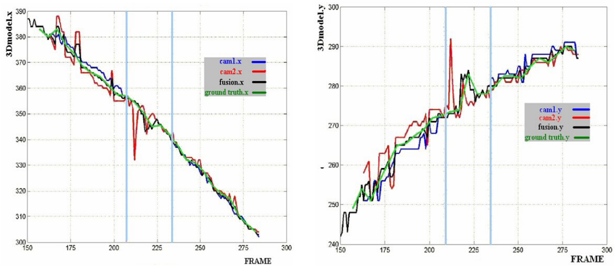
Table 1. Association rate confusion matrix: 2 cameras, 4 objects sequences.



(a) $OSC \equiv f(P)$.

(b) $OSC \equiv f(P, V, S, C)$.

Fig. 4. Histogram of wrong objects associations.



(a) x coordinate

(b) y coordinate

Fig. 5. 3D coordinates of the same object observed by camera1 and camera2, fused and filtered coordinate and ground truth.

Table 0(a) and table 0(b) contain the result matrix for the 4 objects sequences respectively reporting results with the use of position feature and with the complete set of the chosen 4 features.

Presented association results are referred to a sample of 4 objects sequence observed along time: 90 frames contain two occlusion phases where association errors are much more frequent due to the failing of the position and often of the color features. The Y axis has discrete values $\in [0, 1/4, 2/4, 4/4]$ with the meaning of the number of erroneous associations on the total of four. It is easy to be convinced of the higher number of errors presented in figure 4(a) where the sole position is used, comparing to figure 4(a), where the 4-feature set is exploited.

Another interesting result regards the evaluation of the position of the objects estimated in the 3D model and the true position of the objects in the real coordinates (ground truth). As we can see in figure 5(a) and 5(b), the error between the coordinates of the same object observed by camera 1 (in blue) and by camera2 (in red) vs. the ground truth is major than the correspondingly error between fused filtered and filtered position vs. the ground truth. Also in this case, the graphs show that the worst situation for a single camera model happens in the situation of occlusion of the blobs (represented by the azure lines in the pictures): instead, this error is smaller for the fused coordinates.

Lastly, it's possible to compare the computational cost of the 3D vision with the video analysis of each cameras in term of bandwidth and CPU computational load. The 3D model updating requires the reception of a packet composed by integer values: three integers for the identifier label, three for position coordinates, three for speed components and one for the class of the object. So, each object requires 13*32 bit = 416 bit. For ambient intelligence applications we can consider a transmission of 3 packets/second, that implies a bandwidth of around 1248bps.

Instead, if we consider the transmission of the video sequences acquired by a single camera, using colored images with size 720*480pixels, 24 frame/second and using an MPEG2 coding, we need a rate from 4 to 6 Mbps. Obviously, n cameras require n*(4 - 6) Mbps.

Another result is evident in the comparison of the computational load of the CPU. Using a pc configured with a Pentium 4 processor, 2.66 Ghz and 512 MB of RAM, the 3D reconstruction of the scene requires the 35.3% of the CPU load; instead, single camera and dual camera tracking demand 68.6 and 93.4 CPU load percentage.

	bandwidth (Kbps)	**% CPU load**
3D vision	1.5	35.3
Single camera	$4000 - 6000$	68.6
Dual camera	$8000 - 12000$	93.4

Table 2. Comparison of the bandwidth occupation and of the CPU load between 3D, single and camera cameras vision.

The previous results imply that, while multi camera tracking is possible only using pc with high performances, the 3D reconstruction of the scene is allowed also with less capable devices as tablet pc, with the great advantage of portability.

6 Conclusions

In this paper algorithms able to process images from a multi-camera ambient intelligence system and extract features of detected moving objects have been presented. Semantic information extracted from the scene is used by the system for update a dynamic virtual 3D model of the guarded environment. Synthetic automatically-generated 3D scene can be used by an user to be guided into the environment by selecting an arbitrary point of view of the considered area.

7 Acknowledgments

The work was partially supported by the project Virtual Immersive COMmunication (VICOM) founded by the Italian Ministry of University and Scientific Research (FIRB Project).

References

1. T. Rodriguez, P. Sturm, P. Gargallo, N. Guilbert, A. Heyden, J. M. Menendez, and J. I. Ronda, “Photorealistic 3d reconstruction from handheld cameras,” *Machine Vision and Applications*, vol. 16, no. 4, pp. 246–257, sep 2005. [Online]. Available: http://perception.inrialpes.fr/Publications/2005/RSGGHMR05
2. M. Fiocco, G. Boström, J. G. M. Gonçalves, and V. Sequeira, “Multisensor fusion for volumetric reconstruction of large outdoor areas.” in *3DIM*, 2005, pp. 47–54.
3. S. Bahadori and L. Iocchi, “A stereo vision system for 3d reconstruction and semi-automatic surveillance of museum areas,” in *AI*IA 2003: Advances in Artificial Intelligence, 8th Congress of the Italian Association for Artificial Intelligence, Pisa, Italy, September 23-26, 2003, Proceedings of Workshop Intelligenza Artificiale per i Beni Cultural*, Pisa, Italy, Sept. 2003.
4. I. O. Sebe, J. Hu, S. You, and U. Neumann, “3d video surveillance with augmented virtual environments,” in *IWVS ’03: First ACM SIGMM international workshop on Video surveillance*. New York, NY, USA: ACM Press, 2003, pp. 107–112.
5. R. T. Collins, C. Biernacki, G. Celeux, A. J. Lipton, G. Govaert, and T. Kanade, “Introduction to the special section on video surveillance.” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 7, pp. 745–746, 2000.
6. D. Comaniciu, V. Ramesh, and P. Meer, “Real-time tracking of non-rigid objects using mean shift.” in *CVPR*, 2000, pp. 2142–.
7. R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade, “Algorithms for cooperative multisensor surveillance,” in *Proceedings of the IEEE*, vol. 89, no. 10, October 2001, pp. 1456–1477. [Online]. Available: citeseer.ist.psu.edu/collins01algorithms.html
8. T. H. Reiss, “The revised fundamental theorem of moment invariants,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, no. 8, pp. 830–834, 1991.
9. S. Piva, A. Calbi, D. Angiati, and C. S. Regazzoni, “A multi-feature object association framework for overlapped field of view multi-camera video surveillance systems,” in *IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, Como, Italy, 15-16 September 2005, pp. 505–510.
10. R. Y. Tsai, “A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses,” pp. 221–244, 1992.
11. G. Welch and G. Bishop, “An introduction to the kalman filter,” Chapel Hill, NC, USA, Tech. Rep., 1995.
12. Y. Zhao and G. Taubin, “Real-time median filtering for embedded smart cameras.” in *ICVS*. IEEE Computer Society, 2006, p. 55.
13. (2006) The AutoDesk website. [Online]. Available: http://www.autodesk.com/
14. (2006) The Open Scene Graph website. [Online]. Available: http://www.openscenegraph.org/
15. (2006) The Cal3D website. [Online]. Available: http://cal3d.sourceforge.net/
16. (2006) The ReplicantBody website. [Online]. Available: http://sourceforge.net/projects/replicantbody/