EXTRACTION OF 3D SCENE STRUCTURE FROM A VIDEO FOR THE GENERATION OF 3D VISUAL AND HAPTIC REPRESENTATIONS

K. Moustakas, G. Nikolakis, D. Tzovaras and M. G. Strintzis, Fellow, IEEE

Informatics and Telematics Institute
1st Km Thermi-Panorama Road, P.O. Box 361, 57001
Thermi-Thessaloniki, Greece
Tel.: +302310464160
Fax: +302310464164
e-mail: tzovaras@iti.gr

1. INTRODUCTION

The Augmented and Virtual Reality Laboratory (AVRL) of the Informatics and Telematics Institute of the Centre for Research and Technology Hellas was established on June 2000. The AVRL group deals with applications and services in very promising and innovative research areas. Specifically, our work focuses in virtual reality applications in education and training, haptics in virtual environments for medical and assistive technologies applications and virtual environments in cultural, industrial and e-bussiness applications (http://www.iti.gr/db.php/en/pages/virtualreality.html).

Our work in this paper deals with the incorporation of multimodal interfaces for establishing new applications for special population categories. Specifically, the present work focuses on the integration of information originating from two main modalities, i.e. video and haptics for two very innovative applications, namely:

- a virtual reality environment for controlling and monitoring a remote ultrasound examination (utilizing video data and haptic feedback).
- a haptic representation of a real 3D scene for providing access to the blind and visually impaired to real life scenarios.

The innovative aspect of the proposed approach is the integration and synchronization of two completely complementary modalities, i.e. haptics and video. In order to synchronize them, video information is used by a novel structure from motion algorithm in order to extract a 3D representation of the captured scene which is then provided as input to the haptic interaction system.

The paper is organized as follows. In Section 2, the main aspects of the 3D structure reconstruction and 3D model generation algorithms are described. Section 3 presents two applications of the developed system, which illustrate its applicability in existing problems.

2. REAL TIME 3D SCENE REPRESENTATION

A schematic description of the proposed system is illustrated in Figure 1. Initially, the scene is captured using a monoscopic camera. In the following, the video is processed and information about scene structure and characteristics is extracted. Using this information a 3D model of the scene is generated and used in order to create a haptic representation of the observed scene. A brief description of the signal processing algorithms is presented in the following.

The estimation of the scene structure from monoscopic sequences is a problem admitting an infinite number of solutions since true lengths in the scene are unknown. The resulting mathematically ill-posed problem, commonly called Structure from Motion (SfM) in the literature [1, 2], has been under extensive research and analysis [3] for the last decades.

The proposed framework consists of the following steps: Initially 2D motion estimation is performed, using a feature based method, due to its speed and robustness. The location of the feature points in the 2D projection image plane depends on their 3D coordinates, the relative 3D motion between the scene and the camera position and other parameters such as the camera's focal length. There is no prior knowledge on these parameters and thus, they have to be recovered only from the

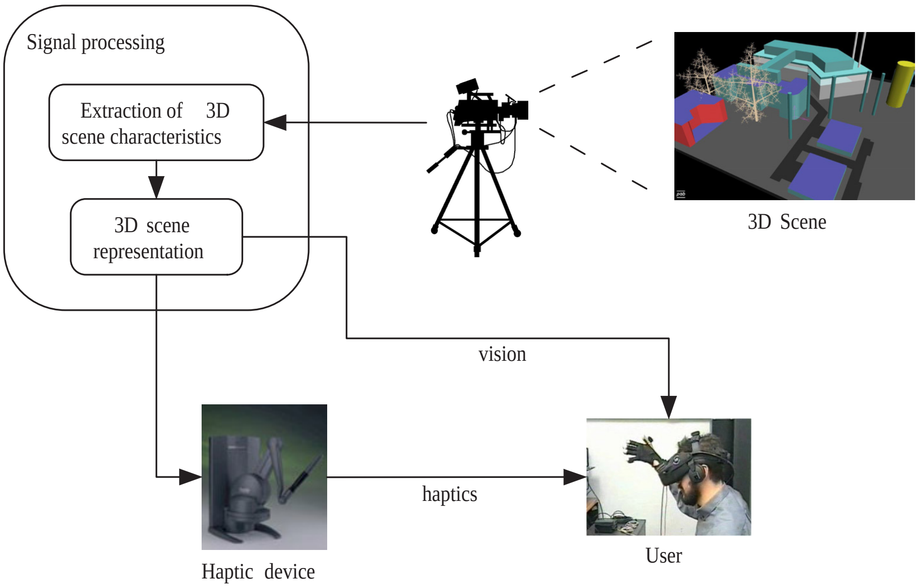


Fig. 1. Architecture of the developed multimodal system

information provided from the 2D motion on the projection plane. For this task Extended Kalman Filtering (EKF) is applied [1, 4] in order to generate an estimate of the scene structure. The latter is combined with an efficient object tracking method and a bayesian framework for occlusion handling, as presented in [4]. Next, the recovered depths of the feature points are used to create dense depth maps via an interpolation method based on 2D Delaunay triangulation. Finally, the obtained scene structure parameters are fed into the 3D model generator system that creates the 3D model of the scene to be used for haptic interaction.

This mesh-based generation of the 3D model is a pure transformation of the 3D data obtained in the structure reconstruction step. More specifically, the feature points that are assigned a depth value [4, 5], are transformed into the vertices of a 3D mesh. The faces of this mesh are constructed utilizing Delaunay triangulation, thus generating a 3D mesh. Due to the limited number of feature points involved into the structure estimation procedure, the mesh based generation of the 3D model is usually not very accurate.

If there is knowledge available about the observed scene, the 3D model generation can be extremely accurate. Parameters of the model like, global and local scaling, translation and rotation are estimated directly from the 3D data obtained at the previous step.

For scenes, that include geometrical surfaces of moderate complexity, a superquadric approximation can be used. This method is widely used in the modeling of 3D data using range images (depth maps) [6]. A non-linear least squares minimization method is utilized in order to estimate the parameters of the superquadrics.

3. EXPERIMENTS AND APPLICATIONS

The proposed method has been used in two main applications:

- Remote ultrasound examination.
- 3D haptic representation for the blind.

3.1. Remote ultrasound examination

This is an application where a doctor performs remotely an echography examination. The main aspects of the system are presented in the following diagram.

As seen in Figure 2 the expert does not have to be near the patient who will only have to be assisted by non-specialised personnel. This paramedical staff has to localise the robot structure on the anatomical region of the patient according to

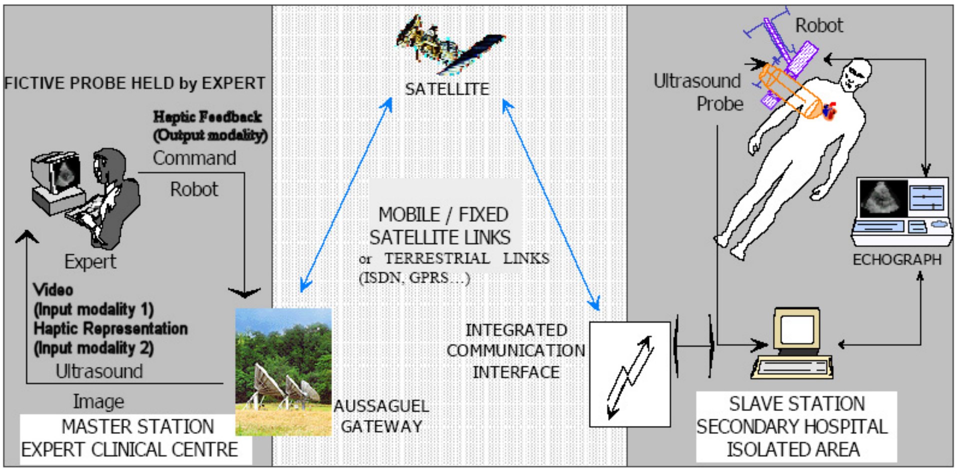


Fig. 2. The remote ultrasound examination system

indications given by the expert. In order to receive the contact force information of the ultrasound probe, the haptic interface at the master station is properly associated to the slave robot.

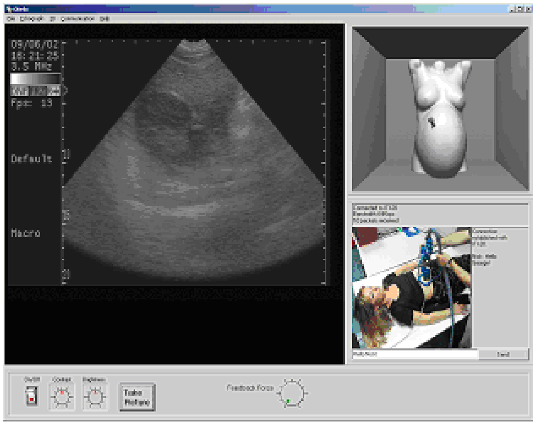


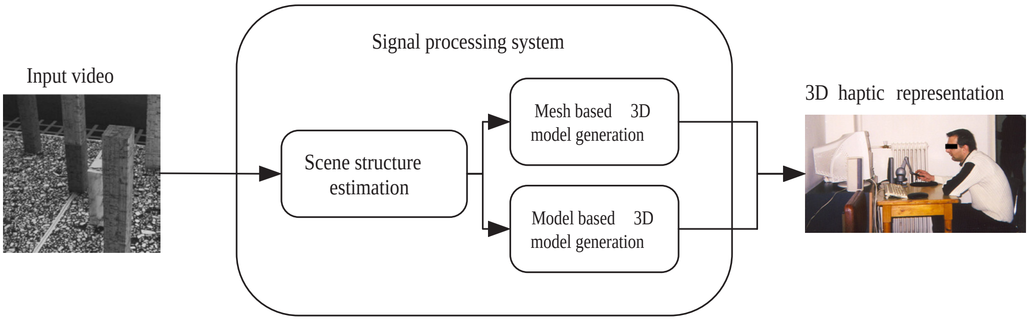
Fig. 3. GUI of the remote ultrasound examination system

A virtual reality environment is used in order to provide the doctor with visual and haptic feedback. It is used for controlling and monitoring the remote ultrasound examination. At the master station site, the expert's role is to control and tele-operate the distant mobile robot by holding a force feedback enabled fictive probe. The PHANToM fictive probe17 can provide sufficient data for the control of the remote site robot. Moreover, the use of this device offers the capability to use force feedback to assist the expert performing the echography examination process. The medical expert must be able to visualize the patient and the slave robot positioning during the medical examination. In order to provide this critical feature in the proposed user interface, the available bandwidth of the communication channel is first detected and evaluated. In the virtual reality environment the expert must be able to use a 3D model that corresponds to the actual part of the patient's body. This can be achieved using a large database of several models or using parametric models for all parts of the human body. The latter is used in our case. A video camera records the patient in the examination area. The method described in the previous section is used in order to extract a 3D video. The extracted depth maps are then used in order to calculate body part representation parameters.

The predescribed system was used in the OTELO IST project (http://www.bourges.univ-orleans.fr/otelo/). The priority order in the system is firstly the ultrasound video, then the master's probe position data (transmitted from master to slave), and finally the force feedback and the robot position feedback data (transmitted from slave to master). In cases that the communication of the force/position feedback data brings significant delay to the system, we prefer not to send this feedback and have all the available bandwidth for the ultrasound video and the probe position data communication. When the channel bandwidth is less than 64Kbps, it is preferable to use one-way communication from the master to slave robot so that the delay times remain under 1 second per minute. In that case the force/position feedback is estimated locally from the reconstructed parametric 3D model.

3.2. 3D haptic representation for the blind

This application focuses on structure reconstruction of a 3D scene. The reconstructed model is accessible using a haptic device, which is used from blind people in order to investigate the processed scene. A block diagram of the application is illustrated in Figure 4.


Input video
Signal processing system
Scene structure estimation
Mesh based 3D model generation
Model based 3D model generation
3D haptic representation

(a)

Fig. 4. Block diagram of the 3D haptic representation generation for the blind people.

In this example, the tower scene is composed of four main parallelepipeda, which are moving mainly across the horizontal direction. The first and last frame of the processed sequence are illustrated in Figures 5a and 5b. After the scene structure is estimated, the resulting depth map for a single frame is shown in Figure 5c. Figure 5d plots the generated 3D model used for haptic interaction.

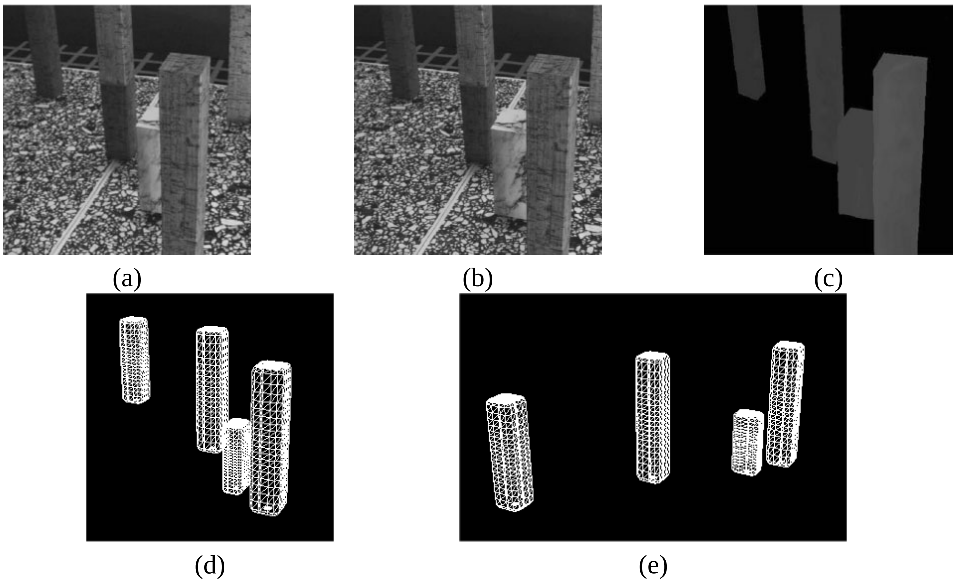
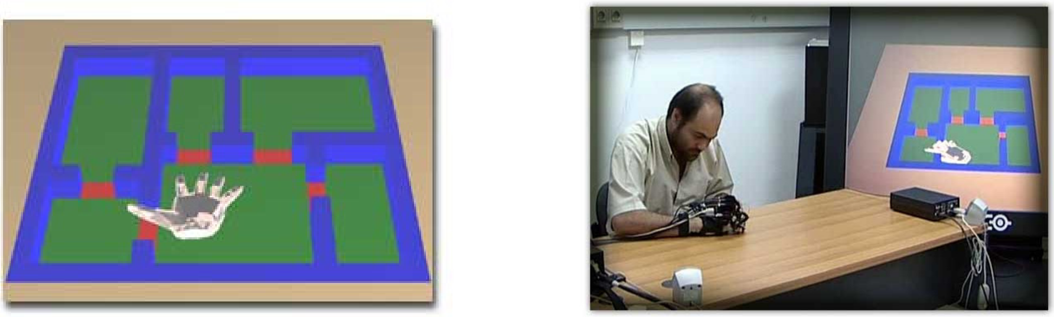


Fig. 5. a) Initial frame, b) Last frame, c) Depth map, d) and e) Resulting model presented from two different viewpoints.

There are two ways to generate such a model. The first is to use the raw information of the depth map and to construct a convex hull utilizing e.g. Delaunay triangulation. On contrary, if there is knowledge about the observed scene available, which is the case for the most application specific tasks, certain models can be assumed, as described in Section 2. In this case we assume that the objects constituting the scene are tower-like. Therefore an accurate haptic representation of the scene can be available.

The main aspects of the above experiment have been used in order to obtain haptic representations of map models, as illustrated in Figure 6a. Initially a camera tracks existing 3D map models of towns, neighborhoods or even apartments, which exist in the schools for the blind. After structure reconstruction is performed, the 3D model for haptic interaction is generated. These 3D models have been used in an experiment for blind people (Figure 6b) in order to teach them how to navigate in those areas.



(a)

(b)

Fig. 6. a) 3D map model, b) A blind person is investigating the 3D map using haptic interaction

In the test, the user should use a maximum of seven minutes to explore the test map. When the user feels to have an ”overview” of the area, she/he has to show to the test leaders the relative position of objects in the scene. Approximately 90% of the users succeeded in identifying the area and 95% of the users have characterised the test as useful or very useful. End users participating in the tests faced no general usability difficulty to the system, especially when they were introduced with an explanation of the technology and after running some exercises to practice the new software.

4. ACKNOWLEDGEMENTS

This work was supported by the Greek Secretariat for Research and Technology Project APEIRO, the EU funded OTELO IST and SIMILAR IST Network of Excellence.

5. REFERENCES

[1] T. Jebara, A. Azarbayejani, and A. Pentland, “3d structure from 2d motion,” *IEEE Signal Processing Magazine*, vol. 16, no. 3, May 1999.

[2] A. Chiuso, P. Favaro, H. Jin, and S. Soatto, “Structure from motion causally integrated over time,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 523–535, April 2002.

[3] S. Soatto and P. Perona, “Reducing structure from motion: A general framework for dynamic vision,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 9, pp. 943–960, September 1998.

[4] K. Moustakas, D. Tzovaras, and M.G. Strintzis, “A non causal bayesian framework for the generation of stereoscopic image sequences,” in *2nd International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT’04)*, Thessaloniki, September 2004, to appear.

[5] S. Diplaris, N. Grammalidis, D. Tzovaras, and Michael G. Strintzis, “Generation of stereoscopic image sequences using structure and rigid motion estimation by extended kalman filters,” in *IEEE International Conference on Multimedia and Expo*, Lausanne, Switzerland, 2002.

[6] F. Solina and R. Bajcsy, “Recovery of parametric models from range images: The case for superquadrics with global deformations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 2, pp. 131–147, 1990.