Summary: This article presents a novel framework for haptic interaction with 3D virtual environments. The system is used by blind or visually impaired people. The important advantage of the proposed method is that it does not handle static virtual scenes, as do most approaches presented in the past. On contrary, it creates the 3D virtual environment dynamically using as input a monoscopic video utilizing efficient signal processing methods in order to extract scene structure from the monoscopic video. The resulting 3D structure data are used to generate the 3D model of the processed scene. The generated model is used to provide haptic access of the processed scene to the blind. Haptic rendering is performed using an efficient method based on superquadrics, which provides fast and smooth force feedback. Experimental evaluation demonstrates the usability of the proposed methods, especially when used to train visually impaired people in simple 3D virtual environments.

1. Introduction

In recent years there has been a growing interest in developing virtual reality (VR) systems for medical applications. Increasing interest is also observed in developing haptic interfaces that allow users to access information presented in 3D virtual reality environments [5]. The greatest potential benefits from virtual environments can be found in applications concerning areas such as education, training the visually impaired, and communication of general ideas and concepts [6]. The technical trade-offs and limitations of the currently developed virtual reality (VR) systems are related to the visual complexity of a virtual environment and its degree of interactivity. A very challenging issue in the scientific area of haptic interaction is the combination of haptic human-computer interaction and a video. The generation of the 3D model of a scene, using as input a monoscopic video, is a special case of the above category.

The estimation of the scene structure from a monoscopic sequence is a pro-
blem 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 [2], has been under extensive research and analysis [15], [11] for the last decades. A variety of different approaches have been proposed in the literature, for the description of the scene structure (lines, curves, surfaces, point features) as well as for the projection model (perspective, orthographic and affine). Different approaches were also exploited in the past to estimate 2D motion, which may then be used for 3D motion estimation and scene structure recovery. Very efficient approaches include those based on optical flow estimation, mathematical transformations and feature-point tracking [12]. Although all these methods yield good results each for different types of sequences, the method in [12] appears to yield generally more robust results, because the motion between consecutive frames is, in the majority of sequences in practice, small enough to allow efficient feature tracking.

The developed framework focuses on the integration of information originating from two main modalities, i.e. video and haptics for a very innovative application, which is 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 the two completely complementary modalities. 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.

In particular, a layered rigid object representation of the scene is used [3], [8]. An EKF based algorithm is applied to estimate the depths of the feature points and dense depth maps are created using 2D Delaunay triangulation. At this point the 3D model of the scene is by applying a superquadric approximation procedure in order to model the observed scene with implicit analytical functions. Finally, force feedback is evaluated using the analytical formulae of the superquadric and the generated force field is provided as input to the haptic interaction system.

The article is organized as follows. In Section 2, the main aspects of the 3D structure reconstruction are described, while Section 3 presents the 3D model generation algorithms. In Section 4 the haptic rendering procedure is described, which utilizes the superquadric modeling of the scene. Finally, experimental results and conclusions are exhibited in Sections 5 and 6 respectively.
2. 3D scene structure extraction

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.

Figure 1: Architecture of the haptic interaction system

Improving on past results [4] Azarbayejani and Pentland proposed in [2] a very robust method for SfM, which in addition to the standard scene structure parameters was also able to estimate focal length. This method, based on Extended Kalman Filtering (EKF) will be briefly described in the sequel.

The camera model used is the central projection model. A translation vector \( t = (t_x, t_y, t_z) \) is used to represent the 3D location of the camera in the current frame, relative to its position in the reference frame. The \( t_x \) and \( t_y \) components correspond to translation along the horizontal and vertical axis of the image plane, while \( t_z \) corresponds to translation along the depth axis and is very sensitive to motion, especially for longer focal lengths. Less sensitive to motion is the product \( t_z \beta \), where \( \beta \) is the inverse focal length \( \beta = 1/f \), which for this reason is estimated instead of \( t_z \). The state vector \( x_i \) of the EKF implementation has \( N+7 \) components [11], where \( N \) is the number of features, which have been tracked.
and $\omega_x$, $\omega_y$, $\omega_z$ are the "Euler angles" [11], [4], which correspond to rotation with respect to each coordinate axis, while $\alpha_1$, ..., $\alpha_N$ are the depths of the N feature points. The measurement vector, used for feedback in a standard EKF implementation, consists of the locations of all tracked features in the new frame.

The technique in [8] proposes an object-based extension of the standard EKF-based approach. In this work, the image is split into objects that are processed separately. This method yields more satisfactory results than those of non-object based methods, because it is based on a realistic object-based model. The associated EKF algorithm also yields more robust results if the depth and motion variations of the tracked features are small, as is true for objects in most sequences in practice.

Figure 2: Flowchart of the proposed method

In the proposed framework, initially rigid objects are identified in the scene. For this task every segmentation algorithm could be used, which does not tend to create fragmented regions. In the present work a method, which utilizes a k-means approach with connectivity constraints for defragmentation is used. Reliable features are extracted for each object in the first frame, using texture sharpness criteria as described in [14], and they are tracked to the final one using an optimized Kanade-Lucas-Tomasi (KLT) [14] feature tracker. Reliable object
masks are extracted and motion and structure are subsequently estimated using the layered EKF-based algorithm. A schematic description of the proposed algorithm and its sub-procedures is shown in Figure 2.

3. 3D scene representation

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 [13], 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 objects composing the observed scene, which is the case for the presented application for training the blind, 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.

3.1 Superquadric modeling

Superquadrics have been excessively used [16], [7] to model objects from range images and depth maps. Typically, they are a family of analytical implicit surfaces like superellipsoids, superparaboloids, superhyperboloids and supertoroids. However, in the literature [16] the term superquadric is usually used to describe superellipsoids, due to their high applicability. Superquadrics (superellipsoids) are described by the following implicit equation.

\[
F(x, y, z) = \left( \frac{x}{\alpha_1} \right)^{\frac{2}{\epsilon_1}} + \left( \frac{y}{\alpha_2} \right)^{\frac{2}{\epsilon_2}} + \left( \frac{z}{\alpha_3} \right)^{\frac{2}{\epsilon_3}} = 1
\]  

(2)

Function (2) is called inside-outside function, because if for a 3D point \((x, y, z)\) the evaluation of (2) yields \(F(x, y, z) > 1\), the point lies outside the convex surface, if \(F(x, y, z) < 1\) it lies inside and if \(F(x, y, z) = 1\) it lies on the surface. Deformation parameters, which correspond to tapering, bending etc. [16] can be added to the implicit equation so as to produce a more flexible model. Figure 3 illustrates four superquadrics for different values of \(\epsilon_1\) and \(\epsilon_2\).
After the selection of the appropriate superquadric equation to model the 3D data, the problem of modeling the 3D object using a superquadric reduces to the least squares minimization of the nonlinear inside-outside function $F(x, y, z)$ with respect to several shape parameters. In particular,

$$F(x, y, z) = F(x, y, z; \alpha_1, \alpha_2, \alpha_3, \varepsilon_1, \varepsilon_2, \phi, \theta, \chi, t_x, t_y, t_z, K_x, K_y, k, \alpha)$$

(3)

where $x$ is a point in the 3D space, are the superquadric shape parameters, and are the Euler angles and translation vector coefficients respectively, and are tapering deformation parameters and the bending deformation parameters. The above parameters are determined so as to minimize the following mean square error.

$$MSE = \sum_{i=1}^{N} \sqrt{\alpha_1, \alpha_2, \alpha_3} (F(x_i, y_i, z_i) - 1)^2$$

(4)

where $N$ is the number of the obtained 3D data. The term $\sqrt{\alpha_1, \alpha_2, \alpha_3}$ was introduced in [16] so as to accelerate the convergence of the minimization algorithm by transforming the parameter space to a steeper one.

The Levenberg Marquardt method for nonlinear least squares minimization, which is widely used [16], [7], is used in the present article in order to evaluate the shape parameters from the 3D data obtained at the SfM step.

### 3.2 Superquadric modeling of the virtual hand

In the context of the present framework the virtual hand is modelled using superquadrics as described in Section 3.1. The most important benefit of the superquadric modeling of the virtual hand is that collision queries can be re-
solved rapidly. In order to get an accurate superquadric approximation, the virtual hand is segmented into 16 segments as illustrated in Figure 4.

![Segmented virtual hand model]

Each finger segment is almost a superquadric on its own. The approximation of the palm is considered more difficult, because it has a more complex shape and overlaps with the TM segment. However, this fact causes no problem to the present approach, since there is no restriction for non-overlapping segments. Assuming that $SQ_i$ represents the superquadric approximation of the $i^{th}$ element of the virtual hand, the superquadric representation of the whole virtual hand (VH) can be described as the union of all superquadrics.

$$ VH = \sum_{i=1}^{16} SQ_i $$

4. Haptic rendering

Haptic perception incorporates both kinesthetic sensing (i.e. of the position and movement of joints and limbs) and tactile sensing (i.e. through the skin). The development of haptic, kinesthetic and tactile devices offers a new dimension of realism to virtual environments and these developments offer further potential applications for advanced multimedia environments. The force feedback device used in the present framework and the methods for evaluating the feedback are described in the subsequent sections.

4.1 Haptic Interaction

The proposed method provides haptic interaction to the user using the CyberGrasp [9] haptic device, which is illustrated in Figure 5b.
CyberGrasp is a force-feedback haptic interface for enabling CyberGlove (Figure 5a) users to experience near-to-realistic force-feedback and to perceive with their hand the volume of computer generated objects. CyberGlove is a low-profile, lightweight glove with flexible sensors, which can measure the position and movement of the fingers and wrist. In order to obtain the position of the whole hand, the MotionStar Wireless\textsuperscript{TM} tracker (Figure 5c) is used [1].

While the user interacts with the application using the CyberGrasp, a virtual hand is displayed in the virtual environment. The user of the applications, which uses the proposed framework is able select and manually apply transformations to the virtual objects and to execute operations according to the application, where it is used. Section 6 describes such possible operations for the two developed applications.

### 4.2 Force feedback evaluation

Consider that point P is detected to lie inside a superquadric as illustrated in Figure 6.
Let also $S_{SQ}^p$ represent the distance of point $P$ from the superquadric, which corresponds to point $P_{SQ}$ on the superquadric surface, i.e. $P_{SQ}$ is the projection of $P$ onto the superquadric. The amplitude of the force fed onto the haptic devices is obtained using a simple spring model as illustrated in Figure 6. In particular:

$$\|F\| = k \cdot S_{SQ}^p$$  \hspace{1cm} (5)

where $k$ is the stiffness of the spring. The rest length of the spring is set to zero so that it tends to bring point $P$ onto the superquadric surface.

The direction of the force feedback is evaluated in most state-of-the-art approaches using the triangulated mesh of the objects. In particular, it is set to be perpendicular to the triangle, for which collision has detected. This approach is not only computationally intensive, but also results in non-realistic non-continuous forces at the surface element boundaries. In the present framework the already obtained superquadric approximation is used in order to rapidly evaluate the force direction. More precisely, the direction of the force feedback is set to be perpendicular to the superquadric surface at point $P_{SQ}$. In particular if

$$r(\eta, \omega) = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \alpha_1 \cos \xi_1 \eta \cdot \cos \xi_2 \omega \\ \alpha_2 \cos \xi_1 \eta \cdot \sin \xi_2 \omega \\ \alpha_3 \sin \xi_1 \eta \end{bmatrix}, \quad \forall \eta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right], \omega \in [-\pi, \pi)$$  \hspace{1cm} (6)

is the parametric definition of the superquadric, the normal vector is defined at point $r(\eta, \omega)$ as the cross product of the tangent vectors along the coordinate curves [10].

$$\begin{bmatrix} \frac{1}{a_1} \cos^{2-\xi_1} \eta \cdot \cos^{2-\xi_2} \omega \\ \frac{1}{a_2} \cos^{2-\xi_1} \eta \cdot \sin^{2-\xi_2} \omega \\ \frac{1}{a_3} \sin^{2-\xi_1} \eta \end{bmatrix}$$

$$n(\eta, \omega) = t_\eta(\eta, \omega) \times t_\omega(\eta, \omega) = s(\eta, \omega)$$

$$s(\eta, \omega) = -a_1 a_2 a_3 \xi_1 \xi_2 \sin^{\xi_1-1} \eta \cdot \cos^{2\xi_1-1} \eta \cdot \sin^{\xi_2-1} \omega \cdot \cos^{\xi_2-1} \omega$$  \hspace{1cm} (7)

where
If several points lie inside the superquadric, the force fed to the haptic device is the average force of all penetrating points. Thus, the force feedback is obtained using equation (9).

\[
F = \frac{k}{N} \sum_{i=1}^{N} S_{SQ}^{i} \frac{n(\eta_i, \omega_i)}{\|n(\eta_i, \omega_i)\|} \tag{9}
\]

The CyberGrasp provides feedback only along the perpendicular direction to the user’s fingers as illustrated in Figure 7.

Thus, if \(n_f\) is the perpendicular direction to a finger, the effective force, \(F_{\text{eff}}\), fed onto the haptic device is:

\[
F_{\text{eff}} = \langle F, n_f \rangle \cdot n_f \tag{10}
\]

![Figure 7: Force feedback for the CyberGrasp](image)

5. Experiments and Applications

As previously mentioned, the developed 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.

In the first case, the “tower scene” is composed of four main parallelepipeds, which are moving mainly across the horizontal direction. The first and last frame of the processed sequence is illustrated in Figure 8a and Figure 8b. After the scene structure is estimated, the resulting depth map for a single frame is shown in Figure 8c. Figure 8d and Figure 8e plot the generated 3D model used for haptic interaction.
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 3. 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 9a. Initially, a ca-
camera tracks existing 3D map models of towns, neighbourhoods 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 9b) in order to teach them how to navigate in those areas.

The proposed haptic interaction system has been evaluated in tests with students of a secondary school and the Blind Association in Thessaloniki, Greece. The evaluation was designed in order to help the qualitative/quantitative estimation of:

- The overall usability of the proposed technologies.
- The acceptance of the tools, the user-friendliness and the points where improvement is needed.
- The acceptance of the demonstration of the novel interaction technologies by the users.

The test procedure consisted of two phases: In the first phase, the users were introduced to the system and they were asked to use it. During this phase, the users were asked questions that focused on usability issues and on their interest in participating to each test. The questionnaire used contained also questions to the test observers, e.g. if the user performed the task correctly, how long did it take him/her to perform the task, etc. The second phase was carried out immediately after the tests, using an after tests questionnaire. Specifically, the users where questioned after finishing all the tests about general issues such as: (a) the benefits and limitations that they foresee on this technology, (b) the usability of the system for recognition, etc.

The system evaluation results have shown that users consider it very innovative and satisfactory in terms of providing realistic and smooth force feedback. The percentage of the satisfied users was reported to be more than 95%.

The Analysis Of Variance (ANOVA) test is used to test differences in means for statistical significance. The ANOVA method was used to compare the performance between male and female users, users familiar to haptic VR systems and those that were not familiar. The time needed to complete each test was used in order to compare the performance of the groups. The result of the ANOVA test are based on the F Distribution. A critical value for the F value is used in order to decide whether the two groups are statistically different or not. The critical value for the parameter $F_{critical}$ of the ANOVA method was calculated to be equal to 4.04 (assuming probability equal to 0.05 and degrees of freedom between groups equal to 1 and within groups equal to 49). When the F result for a pair of groups is grater than the critical value, the difference bet-
ween the mean values for the groups is considered significant and the two groups different from each other. Two groups and 51 measurements were assumed in each case and thus parameters DFS and DFG were computed to be DFS=2-1=1 and DFG=51-2=49.

The gender of the users did not seem to affect the performance results. According to the ANOVA method the F value was 3.33, and the average time was 3.7 min for female users and 4.0 min for male users.

The use of computers affected the performance of the users. The average time for the users that were familiar to haptic VR systems was 3.9 min, while for those that were not familiar the average time was 4.2 min. The F value of the ANOVA test was 4.5 min, which confirms the initial hypothesis that users familiar with haptic VR systems performed better in the environment.

6. Conclusions

In this article a novel framework for the combination of two different modalities, i.e. haptics and video, is proposed. A monoscopic video is processed using efficient structure from motion methods and the resulting 3D structure data are modeled using superquadrics, thus resulting in a 3D virtual environment. Efficient haptic rendering is performed utilizing the analytical formulae the superquadrics, which results in very smooth and realistic force feedback. Finally, an innovative application for training blind users is proposed.

Acknowledgements

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


