3D RECONSTRUCTION FOR A CULTURAL HERITAGE VIRTUAL TOUR SYSTEM

Y. Bastanlar [a,*], N. Grammalidis [b], X. Zabulis [c], E. Yilmaz [a], Y. Yardimci [a], G. Triantafyllidis [b]

[a] Informatics Institute, Middle East Technical University, Ankara, Turkey - (yalinb, eyilmaz, yardimy)@ii.metu.edu.tr
[b] Informatics and Telematics Institute, CERTH, Thessaloniki, Greece - (ngramm, gatrian)@iti.gr
[c] Foundation for Research and Technology - Hellas, Inst. of Computer Science, Heraklion, Greece - zabulis@ics.forth.gr

KEY WORDS: Cultural Heritage, Reconstruction, Web-based, Visualization, GIS, Virtual Reality

ABSTRACT:

The aim of this study is to build a Web-based virtual tour system, focused at the presentation of archaeological sites. The proposed approach is comprised of powerful techniques such as multiview 3D reconstruction, omnidirectional viewing based on panoramic images, and their integration with GIS technologies. In the proposed method, the scene is captured from multiple viewpoints utilizing off-the-shelf equipment and its 3D structure is extracted from the acquired images based on stereoscopic techniques. Color information is added to the generated 3D model of the scene and the result is converted to a common 3D scene modeling format. The 3D models and interactive virtual tour tools such as 360° viewing are integrated with GIS technologies in which the excavation site plans can be added as detailed raster overlays.

1. INTRODUCTION

Web-based virtual tour applications constructed by 360º panoramic images are started to being used extensively all over the world. Effectiveness and usability of these tours were discussed by Bastanlar (2007) and Villaneuva et al. (2004). Usage of image-based 3D reconstructions in virtual tours is limited due to their present day lower quality is not appealing yet. Cultural heritage is one of the most important application areas of these technologies. Example studies (Guarnieri et al., 2004; Kadobayashi et al., 2004; Conforti Andreoni and Pinto, 2004) on 3D reconstruction of cultural heritage were performed by merging image data with the output of 3D laser scanner. 3D scanner technology is efficiently developed to scan the environment and add color information to generate the 3D model. However, the necessary equipment is still very expensive and capturing the 3D data and post-processing is very time-consuming.

In this paper, automatic and photorealistic 3D scene reconstruction from images is used to create content for a cultural heritage virtual tour system. With the same aim, Grün et al. (2002) worked to generate 3D reconstruction of a demolished Buddha statue. However the 3D model is not impressive due to the usage of limited number and low quality photographs. Better results were obtained by Pollefeys et al. (1999) who used the recordings by a video camera. Later they applied their technique for the Sagalassos archaeological site (Pollefeys et al., 2004).

The 3D scene could be synthesized, e.g. by a 3D modeller, by performing surface modelling and then adding texture information. Some of the current WWW applications are composed of graphical textures which are displayed via a VRML plug-in. The problem with such synthesized 3D models for cultural applications is that the feeling of reality is lost and the procedure to generate them is tedious and requires highly-experienced personnel.

An automatic procedure for accurate and photorealistic scene modeling that is efficient in terms of the computational resources is not straightforward. We propose a pipeline for reconstruction and presentation of archaeological sites. In short, the steps are:

1) Acquisition of multiple high-resolution images or video-recording and subsequent selection of key frames.
2) Computation of internal camera calibration parameters.
3) Estimation of lens distortion and image rectification.
4) Extrinsic calibration of the acquired images, based on robust feature extraction, tracking and camera motion estimation techniques,
5) Multi-view stereo reconstruction of the scene using the acquired images and intrinsic and extrinsic calibration parameters.
6) Conversion of the reconstruction output to textured VRML format, which includes triangulation of points into a mesh, combination of textures from different images
7) Generation of KML/KMZ file from VRML format.
8) Display of the reconstructed portion of the archeological site with the excavation site plans as detailed raster overlays, on the Google Earth™ system or other GIS tools that support KML/KMZ format.

As is the case for most multiview stereo reconstruction techniques, the accuracy of the final results greatly depends on the quality of both camera calibration and motion estimation (Steps 2 to 4). To efficiently tackle the problem of fully-automatic motion estimation, the proposed approach employs state-of-the-art techniques (Beardsley et al., 1997; Pollefeys et al., 1999; Pollefeys et al., 2004). Custom modifications were made to these techniques to improve accuracy of the calibration results, namely, robust feature point detection and matching using SIFT (Lowe, 2004) and bundle adjustment (Lourakis and Argyros, 2004). Details are presented in Section 2.

* Corresponding author.

The scene is then reconstructed with the technique proposed in Section 3. In this section, an approach for reconstructing wide area-scenes from high-resolution images with the associated computational issues is proposed. In our technique, the conventional space-sweeping approach (e.g. Zabulis et al. 2003) is slightly modified to employ a sweeping spherical, instead of planar, back-projection surface. Result is a more accurate and memory-conserving technique. Moreover, this extension facilitates the acceleration of the methods, based on a coarse-to-fine depth map computation.

The proposed approach offers to the user the ability to reconstruct a scene from a few snapshots acquired with an off-the-shelf camera, preferably of high resolution. This way, a few snapshots suffice for the reconstruction and the image acquisition process becomes much simpler than capturing the scene with a video camera or with a multicamera apparatus (Mordohai et al., 2007).

The final result is a textured mesh in either the Keyhole Markup Language (KML) or Virtual Reality Modeling Language (VRML) formats. The KML output allows integration to the Google Earth™ platform, thus the reconstructed 3D models and their virtual walkthrough applications can easily become a part of a large geographical information system (GIS) in the near future. Section 4 explains the Web-based virtual tour application developed.

2. ROBUST CAMERA MOTION ESTIMATION BASED ON SIFT DETECTION AND MATCHING

Robust estimation of the camera motion is essential, since the accuracy of the produced 3D reconstruction is based on this information. Our work is based on the approach proposed initially by Beardsley et al. (1997), and, subsequently, extended by Pollefeys et al. (1999, 2004) and Tola (2005). The approach establishes correspondences across consecutive images of a sequence to estimate camera motion.

Previous approaches used the Harris corner detector (Harris and Stephens, 1988) to extract point features in images. The matching procedure utilized similarity as well as proximity criteria (Tola, 2005) to avoid spurious matches. In this paper, an alternative procedure was tested, utilizing SIFT feature detection and matching (Lowe, 2004). In both cases (Harris/SIFT), a RANSAC framework is then utilized to remove spurious correspondences, followed by a Levenberg-Marquardt post-processing step to further improve the estimation. Intrinsic camera parameters are estimated a priori through a simple calibration procedure (Bouguet, 2007). Besides reducing the unknowns in the following external calibration and bundle adjustment procedures, intrinsic calibration is used to compensate for radial distortion. As a result, the perspective camera model is better approximated and the system produces more accurate results. The output is an estimation of the essential matrix **E**, which is thereafter decomposed into rotation matrix (**R**) and translation vector (**t**) of the new view. Finally, triangulation is used to estimate the 3D coordinates of the corresponding features.

When a sequence of views is available, the above technique is applied for the first two views and for each new view *i*, the feature detection and matching approaches are applied to establish 2-D correspondences with the previous view *i-1*, which are then matched with the already established 3-D points, using a RANSAC-based technique that yields a robust estimate of the projection matrix P_i of the new view. We have used an efficient Bundle Adjustment procedure (Lourakis and Argyros, 2004) as a final step at each addition of a new view. The procedure is illustrated in Figure 1.

Although several error suppression and outlier removal steps are included, results show that the accuracy of the whole chain greatly relies on the success of the feature detection and matching. Despite the efficiency of the Harris corner detector and the neighborhood-based constraints utilized in correspondence establishment, we observed that SIFT yields better correspondences in terms of number and accuracy. This is especially important for camera positions with wider baselines. For our problem, robustness to large disparities or severe view angle changes is important because the scene is to be reconstructed from a few snapshots instead of a high-frame rate video.

A technical issue encountered when high resolution images are utilized is that the computation of the SIFT features may require more memory than available. The proposed treatment is to tessellate the image into blocks, compute the features independently in each, and merge the results. To avoid blocking artifacts, the blocks in the above tessellation are adequately overlapping. Duplicate features are often encountered, either due to block overlap or due to collocation of different SIFT that occur at different scales; they are all removed at the merging stage.

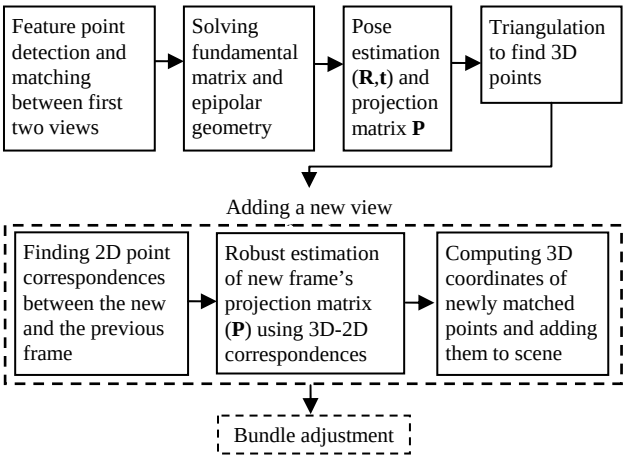

Figure 1. Illustration of the camera motion estimation procedure.

3. 3D RECONSTRUCTION

In this section, an approach for 3D scene reconstruction from high-resolution images is proposed and the associated computational issues are discussed. In the proposed method, the space-sweeping approach is slightly modified to employ a sweeping spherical, instead of planar, backprojection surface (see Zabulis, Kordelas et al. (2006) for an analytical formulation).

The conventional space-sweeping approach is frequently used for multiview stereo reconstruction, due to its computational efficiency and its straightforward acceleration by graphics hardware (Yang et al., 2002; Li et al., 2004). However, it is less

accurate than other approaches that account for the projective distortion due to the orientation of the imaged surface (Zabulis, 2007). For this reason, approaches that employ sweeping in multiple directions (Mordohai et al., 2007) or refine an initial estimation obtained by space-sweeping (Zabulis and Kordelas, 2006) have been proposed.

The proposed technique, based on spherical sweeping, provides higher reconstruction accuracy, especially in the periphery of the images (see Zabulis (2007) for an explanation) and, thus, the available images are more efficiently utilized. In addition, a memory-conserving extension is made to the conventional space-sweeping approaches. This extension also facilitates the acceleration of the methods, based on a coarse-to-fine depth map computation. The importance of memory conservation is twofold. First, the memory of conventional PCs is insufficient to process high-resolution images and using virtual memory renders the process extremely slow. Second, state-of-the-art approaches to stereo reconstruction utilize the graphics hardware to process large amounts of data processing (Mordohai et al., 2007).

The sweeping procedure, which is similar to plane-sweeping, is summarized here briefly. For each depth d_i, the images are backprojected on the, backprojection surface and locally compared. The output of this comparison is a *similarity image* S_i at each depth, whose size is equal to that of the backprojection surface. At each iteration i, the pixels in S_i are compared to their corresponding pixels in S_{i+1} and S_{i-1}. As depth increases, the values for a point in the similarity image correspond to locations along a ray of visibility from the cyclopean eye. The strongest *local* similarity maximum along each such a ray is selected as the optimum depth. The requirement for maxima to be local is used to avoid artifacts that may occur in the textureless areas of the input images.

Memory conservation is achieved by tessellating the backprojection image into, say, $k \times k$ equal spherical segments. This tessellation is parameterized along the two spherical coordinates that, also, correspond to image width and height. The sweeping algorithm is performed independently for each such partition. These partitions overlap slightly, in order to avoid "blocking artifacts" at their boundaries. The amount of overlap is exactly determined by the size of the comparison kernel so that a scene point is not reconstructed twice.

The acceleration of the space-sweeping approach is based on an iterative and coarse-to-fine approach that is combined with the above memory conservation technique. The image data in each iteration are obtained from traditional image pyramids of the input images, starting from the smallest image of the pyramid and advancing a layer in each iteration; at the last iteration the original image is utilized. Also in each iteration, the parameterization of the backprojection surface becomes denser. As described above, the backprojection surface is tessellated and the sweeping algorithm is executed independently for each segment. At each iteration, though, each spherical segment is re-segmented into $k \times k$ more segments. After the 2nd iteration, the range of evaluated depths (d_i) is drastically constrained, based on the reconstruction result previously obtained for the "parent" segment.

The obtained depth map is filtered very conservatively (as in Mulligan et al., 2004), to suppress artifacts at depth discontinuities and remove outliers. By doing so, some valid matches are indeed rejected; however, in the utilized multiview setup the corresponding points are most likely to be reconstructed from another binocular pair. The result is spatially quantized as it is too large ($\propto 10^9$ points for 35 views of 8Mpix each, in this experiment) to fit in memory. To cope with the same limitations the merging process is performed volumetrically, by tessellating the reconstruction volume into cubical segments. Finally, a thin plate interpolating surface is fit (Carr et al., 2001), to yield a mesh outputted into the VRML or KML formats.

In Figure 2, the proposed method is demonstrated for the Dion (Greece) archaeological site. In the experiments presented in this paper, images were 2448 x 3264, 16-bit per layer, color images acquired with a Canon Powershot SLR camera, the number of iterations was 5 and the initial tessellation was 3 x 3. The coarse-to-fine refinement factor was 2, so that in each iteration: (a) the image rows and columns of the stereo and backprojection images were doubled and (b) the number of segments was increased by 4. The above scheme was measured to provide a speedup of ~50 for the scene of this experiment.

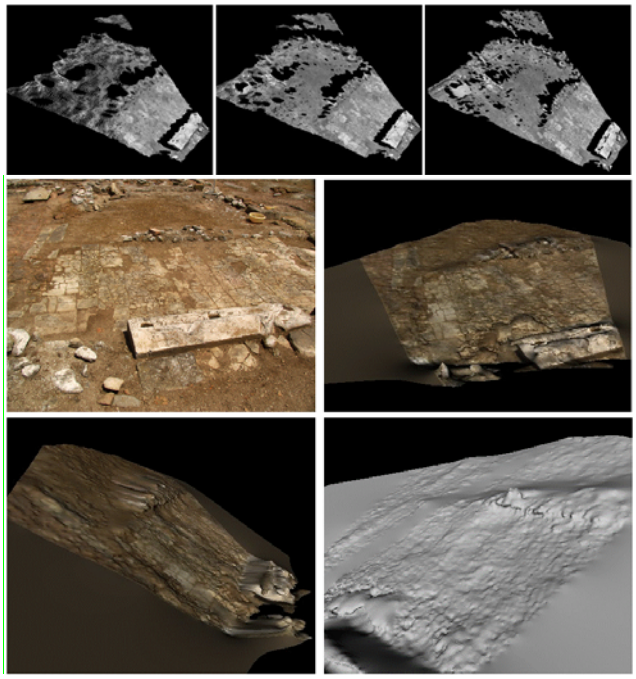

Figure 2. Coarse-to-fine acceleration scheme, for space-sweeping methods. Top row shows the reconstructions for the 3 first iterations of the proposed procedure. In the middle-left an original image from a ~40cm baseline stereo pair (left) is shown. Others are the views of the RBF interpolated reconstruction with and without texture mapping.

Figure 3 shows the result of an experiment that compares the reconstructions obtained from the proposed method in Harris and SIFT conditions of the previous section. The images in the first 2 rows show the result of the reconstruction for an early frame (20 views): in the SIFT condition, a larger proportion of the scene is reconstructed. The last row, shows the result of the SIFT condition after 35 frames.

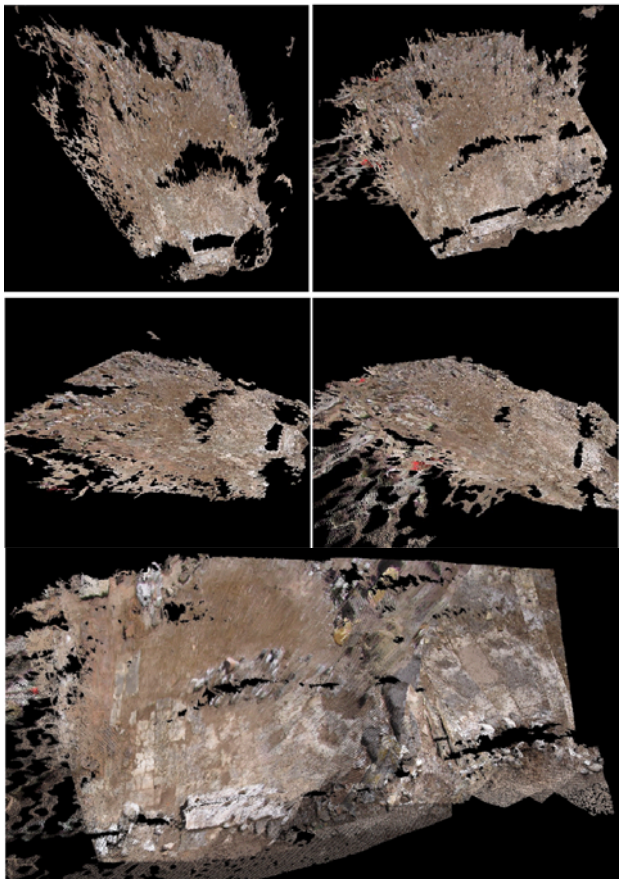

Figure 3. Comparing the reconstructions for the 2 conditions of the experiment in Sec. 3. Top 2 rows show the results for the Harris (left) and the SIFT (right) conditions, for two viewpoints (1st row is top view and 2nd row is side view). The last row of the figure shows a top view of the reconstruction of the same scene, from 35 views and the SIFT condition.

4. VIRTUAL TOUR APPLICATION

The reconstructed VRML models are integrated with GIS technologies within a Web-based virtual tour system, after converting them to the XML-based Collada 3D file format and then referencing to them in Keyhole Markup Language (KML), a format supported by the Google Earth™ GIS platform. Reconstructed part of the archaeological site is placed at its exact location on the terrain. Sample Google Earth™ views for Knossos (Greece) archaeological site are given in Figure 4. If the resolution of Google Earth™ at that location is not satisfactory, excavation site plan can be used as detailed raster overlay, draped over the terrain. Then the reconstructed 3D model will be seen on the site plan.

We added a hyperlink to the application described above, which directs users to a panoramic image based virtual-tour. The main item in this a tour is a viewing window that the user can control. Using Java Applet technology is one proper way of creating such Web-based applications. In addition to the images, audio or textual information related to the site can be presented to the users with extra WWW tools. Using a map of the archaeological site increases the comprehension of the tour and enhances the user's sense of orientation. A step further is making this site plan interactive and integrated with the viewing window. With such tools, more information is communicated to the virtual tour users in an ergonomic fashion (Bastanlar, 2007).

Figure 4. Viewing models in Google Earth™. At the top, overall view of the site together with the reconstructed wall. Bottom-left is the close view of the 3D model of the reconstructed section. The image at bottom-right is a real photograph taken from archaeological site.

In Figure 5, a screenshot of the virtual tour page is shown, which is implemented for the ancient settlement Selime Castle in Cappadocia, Turkey. At bottom-left the viewing window and at the right the site plan are located. The section of the site that is currently presented in the viewing window, field of view (FOV) and direction of view are indicated in the floor plan. It is updated accordingly as the user changes these controls.

A larger degree of immersiveness can be experienced by viewing the reconstructed 3D models on autostereoscopic displays, which can be achieved by using a special plug-in, TriDef™ Visualizer for Google Earth™, to render real-time 3D scenes. We used this property to 3D render the scene for a stereoscopic notebook PC.

The pilot application implemented so far can be reached at http://www.ii.metu.edu.tr/~3daegean/recent.htm

5. CONCLUSIONS

In this study, a Web-based virtual tour system is built for the presentation of cultural heritage. In the proposed approach, the scene is captured from multiple viewpoints utilizing off-the-shelf equipment. We developed and presented the techniques to extract the 3D structure from the acquired images based on stereoscopic techniques. For presentation and 3D modeling of outdoor cultural heritage, the proposed approach as a whole constitutes an economic and practical alternative to the 3D

scanner technology. Generated 3D model of the scene, detailed site plans and interactive virtual tour tools such as 360° viewing were integrated with GIS technologies.

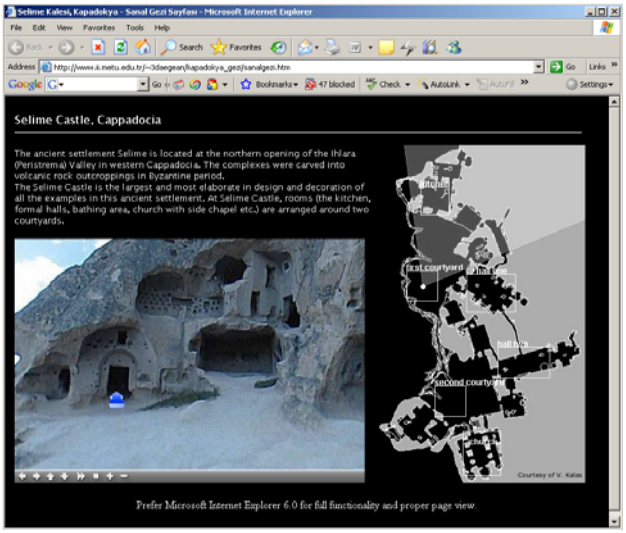

Figure 5. A screenshot from the panoramic image based virtual tour.

REFERENCES

Bastanlar, Y., 2007. User Behaviour in Web-based Interactive Virtual Tours, In: *Proc. of 29th International Conference on Information Technology Interfaces*, Dubrovnik, Croatia.

Beardsley, P., Zisserman, A., Murray, D., 1997. Sequential Updating of Projective and Affine Structure from Motion, *IJCV*, 23(3), pp.235-259.

Bouguet, J., 2007. *Camera Calibration Toolbox for Matlab*, http://www.vision.caltech.edu/bouguetj/calib_doc/ [28/02/2008]
Carr, J.C., Beatson, R.K., Cherrie, J.B., Mitchell, T.J., Fright, W.R., McCallum, B.C., Evans, T.R., 2001. Reconstruction and representation of 3D objects with radial basis functions, *Proc. SIGGRAPH*, pp.67-76.

Conforti Andreoni, D., Pinto, L., 2004. The Creation of The Digital Models for The Protection of Cultural Heritage: The Baptistery of Cremona, *ISPRS Comm. V Sym.*

Guarnieri, A., Vettore, A., El-Hakim, S., Gonzo, L., 2004. Digital Photogrammetry and Laser Scanning In Cultural Heritage Survey, *ISPRS Comm. V Sym.*

Grün, A., Remondino, F., Zhang, L., 2002. Reconstruction of the Great Buddha of Bamiyan, Afghanistan, *International Archives of Photogrammetry and Remote Sensing*, 34(5), pp. 363-368.

Harris, C.G., Stephens, M., 1988. A Combined Corner and Edge Detector, *Proc. of Fourth Alley Vision Conference*, Manchester, U.K., p.182-192.

Kadobayashi R., Kochi, N., Otani, H. Furukawa, R., 2004. Comparison and Evaluation of Laser Scanning and Photogrammetry and Their Combined Use for Digital Recording of Cultural Heritage, *ISPRS Comm. V Sym.*

Li, M., Magnor, M., Seidel, H.P., 2004. Hardware-accelerated rendering of photo hulls, *Eurographics*, 23(3).

Lourakis, M., Argyros, A., 2004. The Design and Implementation of a Generic Sparse Bundle Adjustment Software Package Based on the Levenberg-Marquardt Algorithm, TR #340, ICS-FORTH.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints, *IJCV*, 60(2), pp.91-110.

Mordohai, P., Frahm, J., Akbarzadeh, A., Clipp, B., Engels, C., Gallup, D., Merrell, P., Salmi, C., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewenius, H., Towles, H., Welch, G., Yang, R., Pollefeys, M., Nister, D., 2007. Real-time Video-based Reconstruction of Urban Environments, *Proc. of 3D-ARCH 2007 Workshop, 3D Virtual Reconstruction and Visualization of Complex Architectures*, ETH Zurich, Switzerland.

Mulligan, J., Zabulis, X., Kelshikar, N., Daniilidis, K., 2004. Stereo-based Environment Scanning for Immersive Telepresence, *IEEE Transactions on Circuits and Systems for Video Technology*, 14(3):304-320.

Pollefeys, M., Koch, R., Vergauwen, M., van Gool, L., 1999. Hand-held acquisition of 3D models with a video camera, In: *Proc. of the International Conference on 3-D Digital Imaging and Modeling* (3DIM 1999).

Pollefeys, M., van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., Koch, R., 2004. Visual modeling with a hand-held camera, *IJCV*, 59(3), pp.207-232.

Tola, E., 2005. *Multi-view 3D Reconstruction of a Scene Containing Independently Moving Objects*, M.Sc. Thesis, Middle East Technical University, Ankara, Turkey.

Villaneuva, R., Moore, A., Wong, B.L.W., 2004. Usability Evaluation of Non-immersive, Desktop, Photo-realistic Virtual Environments. In: *The 16th Annual Colloquium of the Spatial Information Research Centre* (SIRC 2004).

http://www.business.otago.ac.nz/SIRC05/conferences /2004/28_Villanueva.pdf [accessed 28/02/2008]

Yang, R., Welch, G., Bishop, G., 2002. Real-time consensus-based scene reconstruction using commodity graphics hardware, *Proc. of Pacific Graphics*, Beijing, China.

Zabulis, X., Patterson, A., Daniilidis, K., 2003. Digitizing Archaeological Excavations from Multiple Views. In: *Proc. of the International Conference on 3-D Digital Imaging and Modeling* (3DIM 2003).

Zabulis, X., Kordelas, G., 2006. Efficient, Precise, and Accurate Utilization of the Uniqueness Constraint in Multi-View Stereo, *Proc. of 3DPVT*.
Zabulis, X., Kordelas, G., Mueller, K., Smolic, A., 2006. Increasing the accuracy of the space-sweeping approach to stereo reconstruction, using spherical backprojection surfaces, *ICIP 2006*, Atlanta, USA.

Zabulis, X., 2007. Utilization of the texture uniqueness cue in stereo. In *Three-Dimensional Television: Capture, Transmission, and Display*, Springer Verlag.

ACKNOWLEDGEMENT

The authors are grateful for support through the 3DTV European NoE, FP6 IST Programme and TUBITAK-GSRT Joint Research Project (105E187). They also thank Engin Tola, for providing the implementation of his structure-from-motion technique and Veronica Kalas, for her guidance and providing material for Selime Castle in Cappadocia. The images of Dion are courtesy of the Archeaological Service of Greece at Dion, Katerini, Greece.