3D SCENE RECONSTRUCTION BASED ON ROBUST CAMERA MOTION ESTIMATION AND SPACE SWEEPING FOR A CULTURAL HERITAGE VIRTUAL TOUR SYSTEM

Xenophon Zabulis[1], *Nikos Grammalidis*[2], *Yalin Bastanlar*[3], *Erdal Yilmaz*[3], *Yasemin Yardimci Cetin*[3]

Foundation for Research and Technology – Hellas, Institute of Computer Science, Heraklion, Greece[1]
Informatics and Telematics Institute-CERTH, Thessaloniki, Greece[2]
Informatics Institute, Middle East Technical University, Ankara, Turkey[3]

ABSTRACT

An efficient 3D reconstruction technique based on robust camera motion estimation and an improved version of the space-sweeping stereo reconstruction approach is presented. The proposed approach is focused on generation of usable and fully automatic reconstruction of wide-area scenes with the computational resources of a conventional PC. The aim is to use this technique to capture such scenes in 3D utilizing off-the-shelf equipment. Color information is finally added to the derived 3D model of the scene and the result can be converted to common 3D scene modeling formats. The 3D models are integrated with GIS technologies within a web-based virtual tour system.

Index Terms—3D reconstruction, camera motion estimation, cultural heritage

1. INTRODUCTION

In this paper, automatic and photorealistic 3D scene reconstruction from images is used to create content for a cultural heritage virtual tour system. The application supports virtual walkthrough and animated events within photorealistic 3D models of the archaeological sites and also, registers these models onto the Google Earth application. The aimed contribution is a fully automatic procedure for accurate, wide-area, and photorealistic scene modeling that is efficient in terms of the computational resources utilized and may be used for 3DTV content creation.

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. To efficiently tackle the problem of fully-automatic motion estimation, the proposed approach employs state-of-the-art techniques [2][3] and a posteriori accuracy improvement through bundle adjustment [6]. In addition, the features derived from the Scale-Invariant Feature Transform (SIFT) [5] were investigated as to their performance against the, traditional, Harris features [18].

Computational efficiency is achieved by introducing an extension to the space-sweeping stereo reconstruction approach. This approach is frequently used for multiview stereo reconstruction, due to its computational efficiency and its straightforward acceleration by graphics hardware [11-13]. However, it is less accurate than other approaches that account for the projective distortion due to the orientation of the imaged surface [14]. For this reason, approaches that employ sweeping in multiple directions [11] or refine an initial estimation obtained by space-sweeping have been proposed [17].

In the context of wide-area stereo, algorithmic complexity and memory capacity are important, because such scenes require more images to be fully reconstructed. Moreover, the need for memory conservation is reinforced by the recent growth of GPU-based software acceleration. The reason is that state-of-the-art graphics hardware has even less on-board memory than a conventional PC.

Usability was a primary concern when designing our system. The system 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 reconstriction and the image acquisition process becomes much simpler than capturing the scene with a video camera or with a multicamera apparatus [11].

Under these conditions, the fully automatic multi-view reconstruction of a scene is not straightforward and, thus, a complete workpath for this task is proposed. This workpath consists of acquiring the images and then estimating the calibration and camera motion parameters as proposed in Sec. 2. The scene is then reconstructed as proposed in Sec. 3 and 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 GoogleEarth™ platform, thus the reconstructed sites can easily become a part of a large geographical information system (GIS) in the near future.

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

Estimating robustly 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 in [2] and, subsequently, extended in [3][4]. The

978-1-4244-1755-1/08/$25.00 ©2008 IEEE

approach establishes correspondences across consecutive images of a sequence to estimate camera motion.

Previous approaches used the Harris corner detector to extract point features in images. The matching procedure utilized similarity as well as proximity criteria [4], to avoid spurious matches. In this paper, an alternative procedure was tested, utilizing SIFT feature detection and matching [5]. 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 [15]. 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 view rotation and translation (**R**,**T**). 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 [6] as a final step at each addition of a new view. The procedure is illustrated in Fig. 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, SIFT yields better correspondences in terms of number and accuracy. Fig. 2 shows the 3D structure (as viewed from the side of the reconstruction) for a stereo pair obtained using both approaches. For this case, there were ~22% more correspondences with SIFT and they were more accurate yielding a better representation of the planar like scene).

The result above is interpreted as a requirement of wider baselines for more descriptive features. Also, while tracking Harris corners in videos provides accurate results [3], wide baselines may create problems to correlation-based matching. 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-framerate 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.

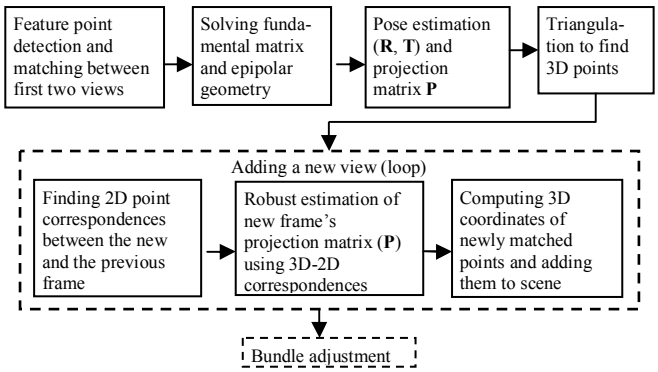

Figure 1: Illustration of the camera motion estimation procedure.

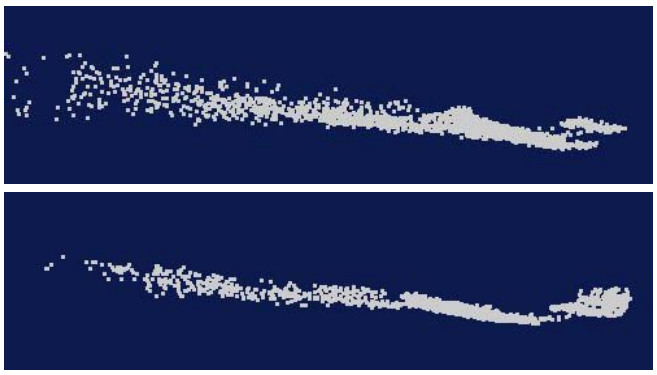

Figure 2: Comparison of the estimated structure when Harris (top) or SIFT (bottom) are used.

3. STEREO RECONSTRUCTION

In this section, an approach for reconstructing wide area-scenes from high-resolution images with the associated computational issues is proposed. In this context, a memory-conserving extension to the conventional space-sweeping approaches is proposed. Moreover, this extension 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 [11]. Since GPUs have even less on-board memory than PCs, RAM memory conservation is, therefore, an important property for GPU-accelerated techniques.

In the proposed approach, the space-sweeping approach is slightly modified to employ a sweeping spherical, instead of planar, backprojection surface (see [19] for an analytical formulation). The technique provides higher reconstruction accuracy, especially in the periphery of the images (see [14] for an explanation) and, thus, the available images are more

efficiently utilized. Otherwise, the sweeping procedure is similar to plane-sweeping and, for this reason, it 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. Memory conservation is crucial, because even though sweeping is performed in a memory efficient manner (similar to [16]), the memory requirements are still large. The reason is that besides the buffer that stores the index of the optimal depth for each pixel in S, two additional such buffers are required to store S_{i+1} and S_{i-1}.

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 [10]), 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 [9], to yield a mesh outputted into the VRML or KML formats.

In Fig. 3, the proposed method is demonstrated. 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.

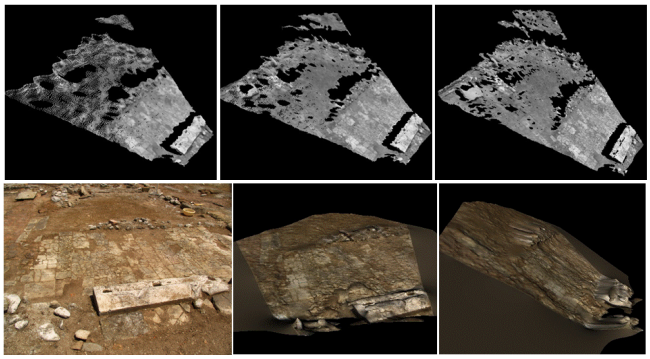

Figure 3: Coarse-to-fine acceleration scheme, for space-sweeping methods. Top row shows the reconstructions for the 3 first iterations of the proposed procedure. Bottom row shows original image from a ~40cm baseline stereo pair (left) and two views of the RBF interpolated reconstruction with texture mapping.

Fig. 4 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 shows the result of the reconstruction for an early frame (20 views): in the SIFT condition, a larger proportion of the scene is reconstructed (~18%, in this experiment). In addition, as already observed in Fig. 2, in the SIFT condition the resulting reconstruction is less noisy. Therefore, the robustness of camera motion estimation provided by SIFT features has a direct impact on the quality of the reconstruction. The last row, shows the result of the SIFT condition after 35 frames.

4. VIRTUAL TOUR APPLICATION-FUTURE WORK

The reconstructed VRML models are integrated with GIS technologies within a web-based virtual tour system, after first 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 GoogleEarth™ GIS platform. Excavation site plans are used as detailed raster overlays, draped over terrain at the exact locations on the earth. When the reconstructed archaeological site is placed on top of the site plan, users view the reconstruction together with the site plan. Additional information (e.g. site-related audio or text) can be presented to users via hot-spots. We added a hyperlink to the application described above, which directs users to a panoramic image based virtual-tour.

Using a map of the archaeological site increases the comprehension of the tour and enhances the user's sense of orientation. With such tools, more information is communicated to the virtual tour users in an ergonomic fashion [8]. A larger level 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 GoogleEarth™, to render real-time 3D scenes.

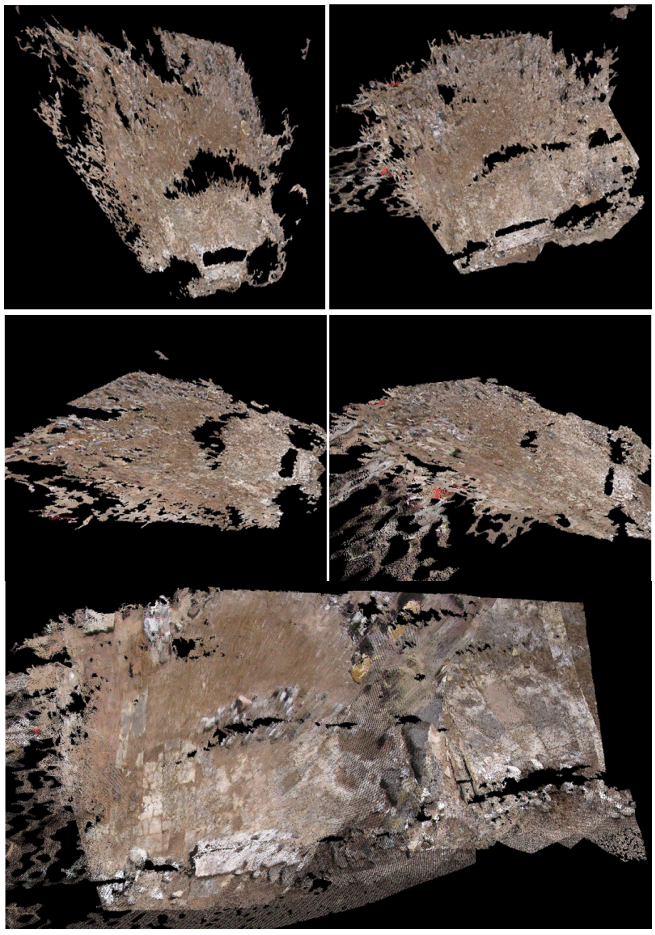

Figure 4: Comparing the reconstructions for the 2 conditions of the experiment in Sec. 3. Top 2 rows show the results for the Harris (left) and the SIFT (right) conditions, for two viewpoints (1st row is top view and 2nd row is side view). The last row of the figure shows a top view of the reconstruction of the same scene, from 35 views and the SIFT condition.

Once the scene has been extracted with reasonable accuracy, it can be passed as input to a more sophisticated, but also more computationally complex approach (e.g. [1], [3], [7]). The initial reconstruction of the scene is important in order to restrict the set of possible solutions and, thus, significantly reduce the computational time. The high-accuracy reconstruction results could then facilitate scientific measurements on the 3D model (instead of the actual site) by archaeologists.

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.

REFERENCES

[1] X. Zabulis, A. Patterson, K. Daniilidis, *Digitizing Archaeological Excavations from Multiple Views*, 3DIM, 2003.
[2] P. Beardsley, A. Zisserman, D. Murray, *Sequential Updating of Projective and Affine Structure from Motion*, IJCV, 23(3):235-259, 1997.
[3] M. Pollefeys, R. Koch, M. Vergauwen, L. van Gool, *Hand-held acquisition of 3D models with a video camera*, 3DIM, 1999.
[4] E. Tola, *Multi-view 3D Reconstruction of a Scene Containing Independently Moving Objects*, M.Sc. Thesis, Middle East Technical University, Ankara, Turkey, 2005.
[5] D. Lowe, *Distinctive image features from scale-invariant keypoints*, IJCV, 60(2):91-110, 2004.
[6] M. Lourakis, A. Argyros, *The Design and Implementation of a Generic Sparse Bundle Adjustment Software Package Based on the Levenberg-Marquardt Algorithm*, TR #340, ICS-FORTH, 2004.
[7] M. Pollefeys, L. van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, R. Koch, *Visual modeling with a hand-held camera*, IJCV,59(3): 207-232, 2004.
[8] Y. Bastanlar, *User Behaviour in Web-based Interactive Virtual Tours*, 29th International Conference on Information Technology Interfaces, p. 25-28 June, Dubrovnik, Croatia.
[9] J.C. Carr , R.K. Beatson , J.B. Cherrie , T.J. Mitchell , W.R. Fright , B.C. McCallum , T. R. Evans, *Reconstruction and representation of 3D objects with radial basis functions*, Proc. SIGGRAPH, pp.67-76, 2001.
[10] J. Mulligan, X. Zabulis, N. Kelshikar, and K. Daniilidis, *Stereo-based Environment Scanning for Immersive Telepresence*, IEEE Transactions on Circuits and Systems for Video Technology, 14(3):304-320, 2004.
[11] P. Mordohai, J.Frahm, A. Akbarzadeh, B. Clipp, C.Engels, D. Gallup, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q.Yang , H. Stewenius, H. Towles, G. Welch, R. Yang, M. Pollefeys and D. Nister, *Real-time Video-based Reconstruction of Urban Environments*, 3D-ARCH 2007.
[12] M. Li, M. Magnor, H. P. Seidel, *Hardware-accelerated rendering of photo hulls,* Eurographics, 23(3), 2004.
[13] R.Yang, G.Welch, G.Bishop, *Real-time consensus-based scene reconstruction using commodity graphics hardware*, Pacific Graphics 02.
[14] X. Zabulis, *Utilization of the texture uniqueness cue in stereo*. In Three-Dimensional Television: Capture, Transmission, and Display, Springer Verlag, 2007
[15] J. Bouguet, *Camera Calibration Toolbox for Matlab,* http://www.vision.caltech.edu/bouguetj/calib_doc/
[16] R. Collins, *A Space-Sweep Approach to True Multi-Image Matching*, CVPR, p. 358-363, 1996.
[17] X. Zabulis, G. Kordelas, *Efficient, Precise, and Accurate Utilization of the Uniqueness Constraint in Multi-View Stereo*, 3DPVT, 2006.
[18] C.G. Harris, M. Stephens, *A Combined Corner and Edge Detector*, Proc. of Fourth Alley Vision Conference, Manchester, 1988, p.182-192.
[19] X. Zabulis, G. Kordelas, K. Mueller, A. Smolic, "Increasing the accuracy of the space-sweeping approach to stereo reconstruction, using spherical backprojection surfaces", ICIP, Atlanta GA, 8-11 October 2006.