FULLY AUTOMATED HIGHLY ACCURATE 3D RECONSTRUCTION FROM MULTIPLE VIEWS

Thomas Ebner, Oliver Schreer, Ingo Feldmann

Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany

ABSTRACT

The reconstruction of real world objects becomes even more important in the view creating highly realistic scenes for Virtual Reality applications. In this paper, we present a fully automated algorithmic pipeline for high-quality 3D reconstruction of real world objects. The proposed method refines an initial 3D model by exploiting the results of additional pairwise stereo depth estimation. An automatic camera selection approach provides different point clouds, which are fused into a common coherent and highly detailed 3D model. The quality of the reconstruction results is discussed in comparison to several state-of-the-art tools, also in the context of automation and performance.

Index Terms— multi-view, stereo image processing, surface reconstruction, 3D modelling

1. INTRODUCTION

3D reconstructions of objects from real world are becoming more and more important in various fields of application [1]. In this paper, we will focus on the reconstruction of architectural sites like buildings and monuments. Such highly realistic 3D reconstructions of static objects can be used for cultural heritage preservation and in virtual museum applications [2].

Numerous solutions specialized on this topic already exist. However, many of these solutions work only on the basis of laser scan data and therefore require specific equipment, which is not available or affordable for everyone. Furthermore, it is inconvenient, if solely point clouds are provided as output format, since they are not directly suitable for integration in commonly used rendering engines. They require further processing steps with separate tools, which makes fully automated workflows difficult.

The core component of the presented approach is the dense depth based 3D surface reconstruction. It starts with a rough point cloud based 3D model, which is further refined with the proposed method. Compared to existing approaches, one novelty is that parts of the 3D model refinement are performed in the related depth maps of the original cameras, with dedicated input view clustering (stereo systems), rather than directly in the model.

The complete processing chain requires a set of input images capturing the object from different views. Usually, 50 to 100 images are sufficient. The processing pipeline initially consists of the estimation of the initial 3D structure using VisualSFM [3]. An automatic camera selection algorithm arranges relevant input images in a number of stereo pairs. A dense depth map refinement is performed to achieve 3D point clouds per stereo pair, followed by a data fusion step to create a final coherent 3D point cloud. Based on Poisson reconstruction and quadric-edge collapse, a manageable 3D mesh and the associated UV atlas are computed, which can then be integrated directly into authoring tools. Details of the individual processing steps are given in the next section.

Since the degree of reconstruction accuracy is one of the key aspects for the proposed algorithmic pipeline, we compare the resulting 3D models to the highly-rated PMVS reference results [7] as well as several state-of-the-art tools from major companies like Agisoft [9] and Autodesk [10][11]. Results of this study are presented in section 3.

2. FULLY-AUTOMATED PIPELINE

The main idea of the proposed algorithmic pipeline is to refine an existing 3D structure by exploiting the results of additional pairwise stereo depth estimation in order to achieve a high-density 3D surface reconstruction.

The pipeline is assembled of several algorithmic components, which are presented in Fig. 1. The blue modules are using commonly available approaches, whereas the green modules refer to our proposed extension for 3D surface reconstruction. The different processing steps are explained in the following sub-sections.

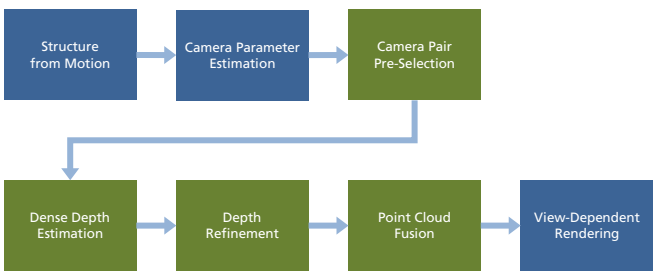

Fig. 1: Algorithmic architecture of the proposed approach.

2.1. Initial estimation of 3D structure

Initially, the 3D structure of the scene is estimated by using well-known structure from motion approaches. In the presented workflow, VisualSFM [3] and SIFT on GPU [4] are used to obtain an initial 3D model and calibration parameters from a given set of unordered input images. The calibration contains the 6DOF pose of the images used in this reconstruction step. The model is described as a sparse 3D point cloud that will be used hereinafter to initialize the dense depth estimation.

2.2. Camera pair pre-selection

The dense depth map refinement is based on neighbored camera views. Hence, an automatic pairwise camera pre-selection has been developed in order to pre-select optimal stereo camera views. The challenge here is to find neighbored stereo images that are close enough for robust depth estimation by keeping maximum possible difference between stereo views without losing object information.

The algorithm aims to calculate a set of image pairs $\{s_1, s_2, \dots, s_n\}$ with $s_i = [i_j, i_k] \in I \times I$ under the following constraints:

- *Compactness*: Redundant images are excluded.
- *Suitability*: The calculated image pairs are well suited for stereo depth estimation.
- *Coverage*: Minimal loss of content compared to the reconstruction using the maximum set of images pairs $\{[i_j, i_k] | [i_j, i_k] \in I \times I \wedge i_j \neq i_k\}$.

For a given image i_j the input consists of the set of visible points P_j (based on feature visibility) and the calibration data c_j. The *compactness* constraint aims to generate a minimal subset of images $I_{min} \subset I$ so that for all images $i_j \in I$ a certain percentage (70% in our case) of its points P_j could be reconstructed using the images from I_{min}. The *suitability* constraint is achieved as follows: For each image pair $[i_j, i_k] \in I_{min} \times I_{min}$ a reconstruction ratio r is calculated as follows:

$$r(j,k) = \frac{\left\| \{ p \mid p \in P_j \cap P_k \wedge 5° < a_{jk}(p) < 45° \} \right\|}{\left\| P_j \cup P_k \right\|}$$

with $a_{jk}(p)$ is the angle of two viewing rays emanating from p towards the two camera centers. The reconstruction ratio r is affected by the baseline (via the angle criterion) and the relative camera orientation within a stereo pair (via the intersection versus union calculation). A higher ratio means a higher compatibility of the contained images in terms of the stereo reconstruction. In order to force the usage of sufficiently compatible images, all pairs having a reconstruction ratio below a given threshold will be removed.

The *coverage* criterion is addressed by applying a greedy search algorithm: The first step consists in finding the stereo system whose set of reconstructible points (i.e. the denominator of the reconstruction ratio) is maximum. In subsequent steps, the stereo pair is chosen, which maximises the coverage ratio, i.e. the set of reconstructible points (defined as the union of the reconstructible points of all contained stereo pairs). Results of the automatic stereo pair selection are shown in Fig. 2.

Fig. 2: First three stereo systems created by automatic camera selection.

2.3. Initial depth map generation

In general, the dense depth-based surface reconstruction approach requires an initial depth-map for each stereo pair. Hence, the 3D information resulting from structure-from-motion needs to be back-projected in each camera. As the initial 3D point cloud does not contain surface meshes, it is not trivial to derive visibility information for each vertex, i.e. to decide which of the points will be visible in a given camera perspective.

Thanks to the output of the structure-from-motion module, the required information is available. The visible points of the left respectively the right camera of a stereo-system are projected to the belonging camera to obtain an initial sparse depth map and cross-bilaterally filtered to densify this sparse map. The filtered map of the left side is then combined with visible points of the right side and vice versa: All visible points occluding the opposite side's depth map will be combined with their visible points and cross-bilaterally filtered. This way, the detected feature points of two views can be combined without introducing artifacts (as only occluding points will be combined) and thus less filtering is required. At the end, the number of bilateral filter iterations required for closing all holes is treated as a quality measure; i.e. the depth map containing more unfiltered pixels is used as initial depth estimate for the depth refinement. An example of initial and dense depth map is given in Fig. 3.

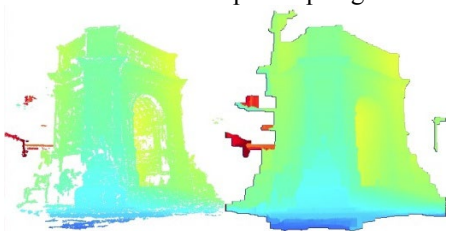

Fig. 3: Initial depth map creation: (left) initial sparse map, (right) initial dense map.

2.4. Depth estimation and fusion

The final step of the depth map estimation is carried out by a highly accurate Patch-Sweeping algorithm [5] refining the dense depth maps from the previous step. Patch-Sweeping assumes that a 3D object surface can be described with quadratic surface elements, which are named as spatial patches. In order to estimate an object surface, the volume of interest is quantized by oriented spatial patches along the viewing rays of a reference camera according to a discrete number of depth layers. These patches represent depth hypotheses, which are evaluated by projecting the patches onto the image planes of all cameras, execute a texture lookup and average the pairwise normalized cross correlation for all images and each patch projection. For every pixel of the reference image, a depth value is assigned by a winner-take-all selection among the matching scores of all depth hypothesis along the corresponding viewing ray. Since the algorithmic concept is highly parallelizable and therefore well suited for GPGPU processing, the enormous computational load can be brought to graphics cards, what enables high quality real-time depth estimation. Finally, the multiple refined depth maps from all stereo pairs are fused into an overall high-resolution 3D model using the visibility-driven patch group generation [6]. By applying this fusion procedure, all 3D points occluding any other depth map are filtered out, resulting in an advanced foreground segmentation. The remaining artifacts have a greater distance to the object to be reconstructed. As a result, they do not occlude any other depth maps.

3. RESULTS

In order to evaluate the quality of the proposed automatic reconstruction pipeline, we compared our results to several state-of-the-art approaches.

Fig. 3 depicts the results on the reference data set *fountain-P11* [13]. Compared to the highly-rated PMVS reference results [7] our method shows a significant improvement of visual quality and geometric detail. Fig. 5 illustrates this on the example of the shaded 3D model. For example, more geometrical details can be seen in the surface of the bricks and the golden fish. Additionally, with our method the border area on the right-hand side of the wall is reconstructed with higher quality and less artefacts.

Moreover, we competed the proposed workflow against several professional tools available on the market, among others 3DF Zephyr [8], Agisoft PhotoScan [9], Autodesk 123D Catch [10] and ReMake [11], and Reality Capture [12]. Since we had no access to their original or intermediate reconstruction data with possibly more details, we reduced our meshes to adequate complexity, what allows for better comparability. The applied post-processing pipeline is also fully automated and involves screened Poisson surface reconstruction [14], followed by a simplification to a dedicated amount of triangles by iterative contraction of edges based on Quadric Error Metrics [15]. Finally, for restoring details that got lost during simplification, the utilization of a texture in contrast to the vertex colors calculated in the patch fusion step is required.

Overall, we found, that ReMake and PhotoScan performed best. Fig. 6 and Fig. 7 highlight some reconstruction details in comparison with these programs for the examples of Arco Valentino in Torino and the Statue of Goethe in Berlin.

Besides the visual quality, these tools have their individual pros and cons regarding automation and performance. Table 1 gives an overview of the most significant findings.

Autodesk ReMake	Our Approach	Agisoft PhotoScan
– supports only JPEG images	+ supports several image formats	+ supports several image formats
+ visual quality of reconstructed object	+ visual quality of reconstructed object	+ visual quality of reconstructed object
– coarse geometry, details mainly through texture	+ highly accurate and dense geometry	+ highly accurate and dense geometry
+ meshes usually watertight	+ meshes usually watertight	– meshes often contain holes
+ fully automated	+ fully automated	– semi-automated
+ easy-to-use workflow	+ easy-to-use workflow	– complicated workflow
+ good background segmentation	+ good background segmentation	– manual masking recommended
– manual parameter adjustments not possible	+ manual parameter adjustments possible	+ manual parameter adjustments possible

Table 1: Comparison of Pros (+) and Cons (–).

4. CONCLUSION

Even though we focused on the development of a fully automated workflow, the proposed algorithmic pipeline allows for targeted interventions in order to fine-tune the generation of the 3D reconstruction results. For example, the characteristics of the input image data sets as well as different demands on 3D structure quality or computational effort can be particularly taken into account. However, the presented approach outperforms current state-of-the-art software in terms of automation and level of quality. As presented in section 3, the geometrical detail is much higher compared with conventional tools. Another advantage of the presented approach is the GPGPU centric implementation, which offers a significant gain in performance.

Fig. 4: Original image (left) and reconstructed surface (middle, right) of *fountain-P11* data set.

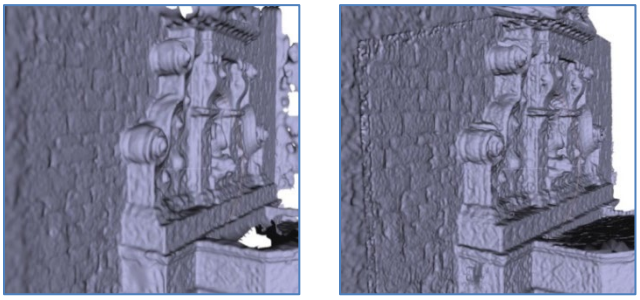

Fig. 5: Shaded 3D model of *fountain-P11* data set reconstructed with PMVS reference method (left) and our approach (right).

Autodesk ReMake | Our Approach | Agisoft PhotoScan

Fig. 6: Comparison of reconstruction details at the example of Arco Valentino in Torino.

Autodesk ReMake | Our Approach | Agisoft PhotoScan

Fig. 7: Comparison of reconstruction details at the example of the Statue of Goethe in Berlin.

5. REFERENCES

[1] Ebner, T., Feldmann, I., Renault, S., Schreer, O., and Eisert, P., "Multi-view reconstruction of dynamic real-world objects and their integration in augmented and virtual reality applications," *Journal of the Society for Information Display*, Vol. 25, No. 3, pp. 151–157, 2017.

[2] Feldmann, I., Schreer, O., Ebner, T., Eisert, P., Hilsmann, Nonne, A. N. and Haeberlein, S., "Digitization of People and Objects for Virtual Museum Applications," *Electronic Media and Visual Arts*, Berlin, Germany, 2016.

[3] Wu, C., "VisualSFM: A visual structure from motion system," 2011.

[4] Wu, C., "SiftGPU: A GPU implementation of scale invariant feature transform (SIFT)," 2007.

[5] Waizenegger, W., Feldmann, I. and Schreer, O., "Real-time Patch Sweeping for High-Quality Depth Estimation in 3D Videoconferencing Applications," *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, pp. 78710E-78710E, 2011.

[6] Ebel, S., Waizenegger, W., Reinhardt, M., Schreer, O. and Feldmann, I., "Visibility-driven patch group generation," *3D Imaging (IC3D), 2014 International Conference on*, IEEE, pp. 1–8, 2014.

[7] Furukawa, Y. and Ponce, J., "Accurate, dense, and robust multiview stereopsis," *IEEE transactions on pattern analysis and machine intelligence*, IEEE, Vol. 32, No. 8, pp. 1362-1376, 2010.

[8] 3Dflow *3DF Zephyr* (v.3.0), http://www.3dflow.net

[9] Agisoft *PhotoScan* (v1.3.0), http://www.agisoft.com

[10] Autodesk *123D Catch*, http://www.123dapp.com/catch

[11] Autodesk *ReMake* (v17.25.0.16), http://remake.autodesk.com

[12] Capturing Reality *RealityCapture*, https://www.capturingreality.com

[13] Strecha, C., von Hansen, W., Van Gool, L., Fua, P. and Thoennessen, U., "On benchmarking camera calibration and multi-view stereo for high resolution imagery," *Computer Vision and Pattern Recognition (CVPR)*, 2008 IEEE Conference on, Anchorage, AK, pp. 1-8, 2008.

[14] Kazhdan, M. and Hoppe, H., "Screened Poisson Surface Reconstruction," *ACM Transactions on Graphics (TOG)*, Vol. 32, No. 3, 2013.

[15] Garland, M. and Heckbert, P. S., "Surface simplification using quadric error metrics," *Proceedings of the 24th annual conference on Computer graphics and interactive techniques (SIGGRAPH '97)*, ACM Press/ Addison-Wesley Publishing Co., New York, NY, USA, pp. 209-216, 1997.