Accurate stereo 3D point cloud generation suitable for multi-view stereo reconstruction

Georgios A. Kordelas [⋆†], Petros Daras [⋆], Patrycia Klavdianos [†], Ebroul Izquierdo [†], Qianni Zhang [†]
[⋆] *Information Technologies Institute, Centre for Research and Technology, Thessaloniki, Hellas*
[†] *Multimedia and Vision Group (MMV). Queen Mary University of London, UK*

***Abstract*—This paper proposes a novel methodology for generating 3D point clouds of good accuracy from stereo pairs. Initially, the methodology defines some conditions for the proper selection of image pairs. Then, the selected stereo images are used to estimate dense correspondences using the Daisy descriptor. An efficient two-phase strategy to remove outliers is then introduced. Finally, the 3D point cloud is refined by combining sub-pixel accuracy correspondences estimation and the moving least squares algorithm. The proposed methodology can be exploited by multi-view stereo algorithms due to its good accuracy and its fast computation.**

***Index Terms*—3D content generation, Multi-view stereo**

I. Introduction

The automatic and accurate 3D modeling of objects and scenes, from multiple photographs or videos, constitutes an important objective in the computer vision and graphics research fields. The realistic 3D models can be exploited in multiple applications, such as computer graphics, TV/film special effects, robot navigation and computer games.

Research in 3D model reconstruction using multi-view stereo algorithms has made signicant progress in the computer vision community. Multi-view stereo (MVS) algorithms take multiple images with pose information as input and produce dense 3D models with increased accuracy. They can be divided into two categories according to the scale of the considered scenes. Small-scale methods include voxel-based approaches [1], [2], which require the definition of a bounding box that encloses the scene, and approaches based on deformable polygonal meshes [3], which require a good estimation of a visual hull to initialize the corresponding optimization procedure. While being accurate, those methods do not scale up to large scale scenes, since their memory and computational requirements increase exponentially with the size of the scene.

On the contrary, depth map fusion methodologies [4] can scale up to very large scenes, sometimes sacrificing some of the accuracy and completeness of volumetric methods [5].

Other large scale methods generate and merge collections of 3D points clouds, which may be then used to generate a mesh surface [6]–[9]. Many of the methods that rely on 3D point clouds, put emphasis on the merging of the point clouds that are generated from different stereo pairs by using visibility constrains to filter erroneous points. Our proposed methodology could foster these approaches by improving the accuracy of the individual point clouds, which are generated from each stereo pair, before point clouds from all stereo pairs are merged.

The rest of this paper is organized as follows. In section 2 the proposed methodology is described. Section 3 provides information on the parameters used, the experimental results and the computational cost. Finally, the conclusions are drawn in Section 4.

II. Stereo dense 3D point cloud generation

In general, the first step of multi-view 3D reconstruction is the computation of camera(s) poses that capture a scene. The Structure-from-Motion (SfM) approach presented in [10] provides an efficient way for computing robustly the camera parameters from a set of user-generated images. In this paper, an efficient methodology for generating an accurate 3D point cloud from a stereo image pair, is presented. The approach can be divided into three stages:

1) During the first stage, the stereo pairs to be used for the generation of each stereo point cloud are appropriately selected, based on specific conditions, in order to ensure the accuracy of reconstruction.
2) The second stage includes the estimation of dense correspondences between the images of the stereo pair, based on fast DAISY [11] descriptor matching. Additionally, a strategy for filtering outlier correspondences is presented.
3) The third stage involves refinement of the generated 3D point cloud. Refinement is accomplished by estimating the correspondences in sub-pixel accuracy and by smoothing the resulting point cloud using the moving least squares algorithm.

The innovation of this method lies mainly in the efficient strategy for removing outliers and in the effective combination of sub-pixel accuracy correspondences estimation with the moving least squares algorithm to improve the accuracy of the generated 3D point cloud. In the following, more details are provided on what each of these stages comprises.

A. Stereo Pair Selection

Stereo images pair selection is a crucial step to acquire stereo 3D point clouds with good accuracy. The images of an "adequate" stereo pair should have significant overlap to be easily matched, but also to be sufficiently separated, since much closeness may result to point cloud estimation errors. This is quantified, similarly to [8], by measuring the angle θ between the camera principal rays of the stereo images. The condition that θ should satisfy is: $\theta_{min} < \theta < \theta_{max}$.

Afterwards, Quasi-Euclidean epipolar rectification [12] is applied to each stereo images pair that satisfies the previous

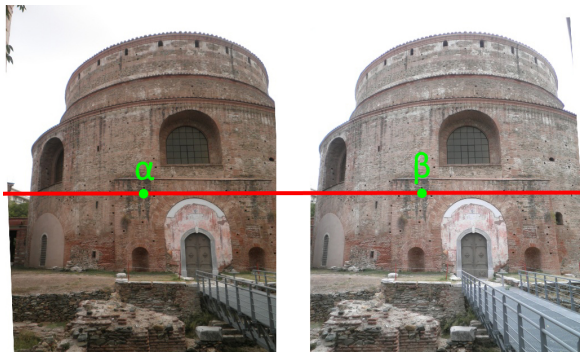


Fig. 1: Correspondence in rectified stereo images.

condition. If the Quasi-Euclidean epipolar rectification error T_{rect} is below a threshold T_{max}, the stereo pair is assumed as suitable for proceeding to the estimation of its point cloud. Consequently, this work, except for the condition based on θ, defines a second condition based on T_{rect} for selecting adequate stereo pairs.

B. Dense correspondences estimation and outliers filtering

During the second step, the DAISY descriptor [11] is exploited to estimate dense correspondences between the images of a stereo pair. Daisy has been selected for this scope, because it has been proved to be very efficient for dense wide baseline matching and at the same time DAISY outperforms the SIFT [13] and SURF [14] descriptors regarding the accuracy of matching.

More specifically, in order to find for a pixel on one image its corresponding pixel to the other image, we search for the pixel's DAISY descriptor the pixel with the nearest DAISY descriptor on the second image. The search is constrained along horizontal epipolar lines, since the images have been rectified. Fig. 1 depicts a pixel correspondence $\alpha - \beta$ on an epipolar line, between a rectified stereo pair. The search for the nearest descriptor is performed using approximate nearest neighbor searching based on randomized kd-trees [15], where trees are searched in parallel. The kd-trees search approach significantly boosts the speed of searching, when compared to exhaustive search.

The correspondence estimation is performed twice. Once having as reference the first image of the stereo pair and once having as reference the second image. Then, the Left-Right consistency check [16] is used for detecting the correspondence outliers.

Except for this common technique, an additional technique to filter outliers in a segment level, and not in a pixel level, is proposed. This technique helps to remove outliers that appear in textureless regions. Initially, mean-shift segmentation is used to partition the image into different segments that contain groups of pixels (the segmentation map of the left image of Fig. 1 is visualized in Fig. 2(a)).

Then for each segment, the percent of pixels that pass the right-left consistency check to the total number of pixels contained in the segment, is computed. If this percent is over 50%, then the correspondences in the segment are considered as inliers. Otherwise, all the correspondences in the segment are considered as outliers.

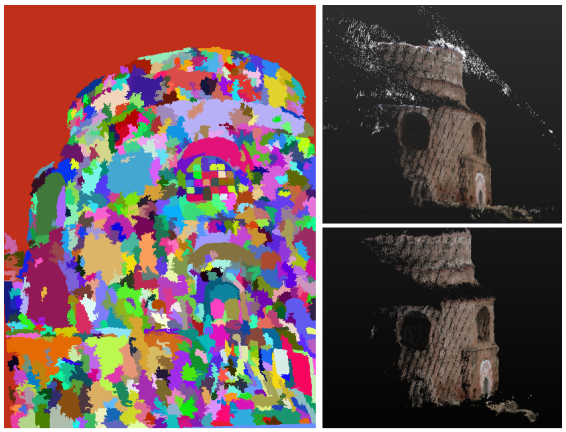


Fig. 2: (a) Mean-shift segmentation map of an image and (b) Generated stereo point cloud without using (upper part) and, when using (bottom part) the proposed outliers filtering strategy.

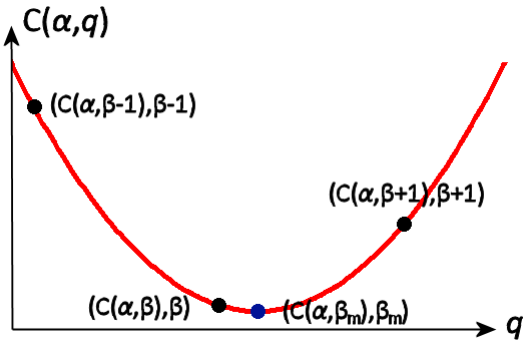


Fig. 3: Sub-pixel accuracy correspondence using quadratic curve fitting.

This strategy assists in filtering numerous outliers. This fact is evident in the visual example of Fig. 2(b). The upper part of Fig. 2(b) shows the point cloud that is generated without using the proposed outliers filtering strategy, while the bottom part of Fig. 2(b) depicts the point cloud after applying the outliers filtering strategy. Obviously, the second point cloud contains less outliers.

C. Point cloud refinement

1) Correspondences estimation in sub-pixel accuracy

So far, the estimated correspondences have pixel accuracy. However, correspondence estimation at sub-pixel accuracy can significantly improve the quality of the generated 3D point cloud, since pixel accuracy matching, results in discrete and not continuous values of depth information.

In order to achieve sub-pixel accuracy the following process is followed. Let us suppose that a pixel α on the left image corresponds to a pixel β on the right image and their matching cost $C(\alpha, \beta)$ has already been estimated. Then, the matching cost $C(\alpha, \beta - 1)$ between the DAISY descriptors of pixels α and $\beta - 1$ and the matching cost $C(\alpha, \beta + 1)$ between α and $\beta + 1$ are estimated. The three points $(C(\alpha, \beta - 1), \beta - 1)$, $(C(\alpha, \beta), \beta)$ and $(C(\alpha, \beta + 1), \beta + 1)$ (these points are visualized in Fig. 3) are used to estimate a quadratic function

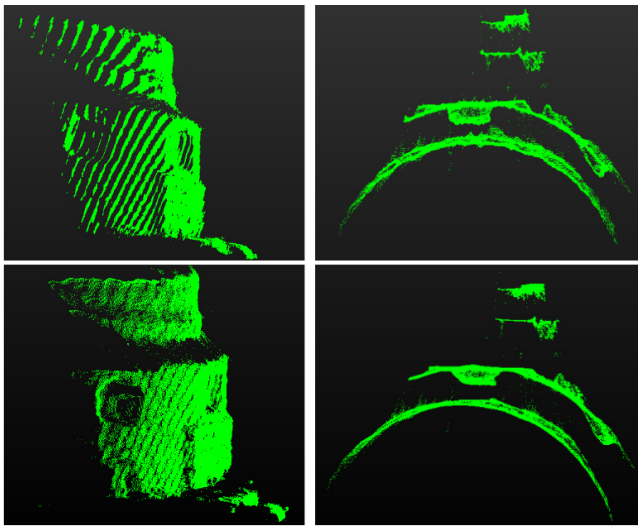


Fig. 4: (a) Point cloud that corresponds to pixel accuracy (upper part) and sub-pixel accuracy (bottom part) correspondences and (b) Point cloud before (upper part) and after (bottom part) applying the Moving Least Squares algorithm.

and estimate the minimum cost $C(\alpha, \beta_{\mathrm{m}})$ of the quadratic function's curve, which corresponds to β_{m}. Consequently, the sub-pixel accuracy correspondence is assumed to be given by the pair (α, β_m), while the pixel accuracy correspondence was given by the pair (α, β).

The upper part of Fig. 4(a) shows the point cloud that corresponds to pixel accuracy correspondences, while the bottom part of Fig. 4(a) depicts the point cloud that corresponds to sub-pixel accuracy correspondences. It is evident, by comparing these two parts, that the bottom point cloud is more accurate, since depth information is continuous.

2) Point cloud smoothing

The 2D sub-pixel correspondences estimated in Section II-C1 are converted into 3D point clouds using the projection matrices that were estimated during the SfM process. Afterwards, a final step is applied to improve the reconstruction quality.

More specifically, in order to resample and smooth the generated point cloud the Moving Least Squares (MLS) algorithm, described in [17], is exploited. The upper part of Fig. 4(b) shows the point cloud before applying the MLS algorithm, while the bottom part of Fig. 4(b) after applying the MLS algorithm.

III. Experimental results

A. Set of optimum parameters

The limits for the principal rays' angles are set to $\theta_{min} = 5°$ and $\theta_{max} = 25°$. The rectification error threshold is set to $T_{max} = 0.5 \cdot (D_{max}/640)\ pixels$, where D_{max} is the maximum dimension of the images (width or height), which constitute the image pair, in pixels. In this way, T_{max} is set proportional to the size of the stereo images to be rectified.

The selected parameters for computing the DAISY descriptor are the radius of the descriptor $R = 9$, the number of rings $Q = 3$, the number of histograms on each ring $T = 4$ and the number of bins of the histograms $H = 4$.

The segmentation parameters are the segmentation spatial radius, which is set to $h_r = 3$ and the segmentation feature space radius, which is set to $h_s = 3$. The selection of these strict values ensures that the segmentation map will be of high reliability, meaning that most likely a segment will not overlap a depth discontinuity, and this fact is verified in [18] and [19].

B. Experiments

A stereo pair of images, which has been derived from the Herz-Jesu-P8 [5] (in specific images "0007.png" and "0008.png") is used to visually indicate the improvement introduced by the proposed methodology, regarding the accuracy of the estimated stereo point cloud. The images have been downscaled with a factor of 3, so as to make more obvious the accuracy improvement in visual data of lower resolution. The generated stereo 3D point cloud using this approach is visualized in Fig. 5(a).

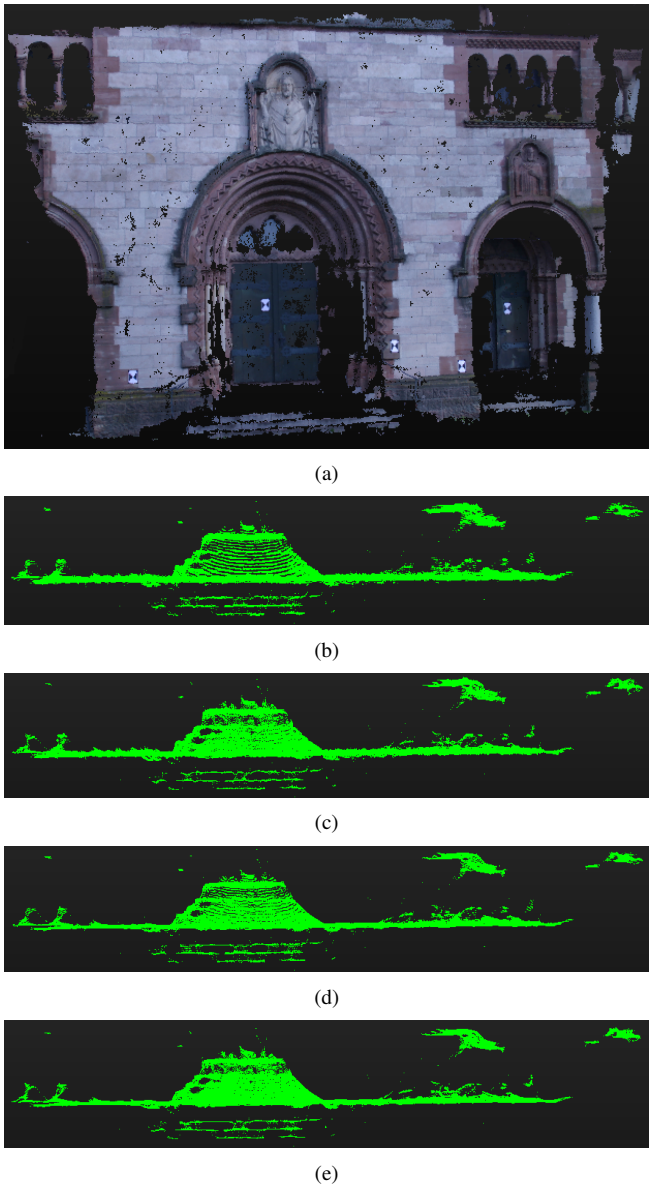
In the following, the stereo point cloud, with or without using the proposed refinement steps, is estimated. The point cloud (observed from the upper viewpoint): (i) without using sub-pixel accuracy nor MLS algorithm is visualized in Fig. 5(b), (ii) using only sub-pixel accuracy is visualized in Fig. 5(c), (iii) using only MLS algorithm is visualized in Fig. 5(d) and (iv) using both sub-pixel accuracy and MLS algorithm is visualized in Fig. 5(e). Evidently, Fig. 5(e) gives the more accurate stereo point cloud.

In the second example, this approach is used to generate individual stereo point clouds using images captured from the Rotunda Ancient Monument in the city of Thessaloniki. Then, the point clouds are finally concatenated to form the final 3D point cloud. This 3D reconstruction example is depicted in Fig. 6. The right part of Fig. 6, which depicts the overview of the 3D reconstruction, indicates that individual point clouds have satisfactory accuracy, so that they are well registered to form a complete 3D representation of the captured object, even without using any method for combining the individual point clouds.

C. Computational cost

The proposed methodology has been implemented in C++ and it has been tested on a laptop PC with an Intel Core i5-2430M 2.4GHz CPU. A pair of images with size 1024x682 is used to report on the processing time required for each step of the methodology. In the parenthesis is also reported the percentage of total time that is spent in each step. The rectification step (subsection II-A) requires 4.1 sec (5.942%), the dense Daisy computation and correspondence estimation (subsection II-B) requires 20.5 sec (29.71%), the outliers filtering (subsection II-B) requires 4.2 sec (6.087%), the sub-pixel accuracy correspondence estimation (subsection II-C1) requires 5.6 sec (8.116%) and finally the MLS algorithm (subsection II-C2) execution requires 34.6 sec (50.145%). The total processing time is 1.15 min and it is acceptable bearing in mind the obvious improvement in the 3D point cloud accuracy

and that this methodology aims to be used by Multi-view stereo algorithms, which are not real-time applications.



(a) (b) (c) (d) (e)

Fig. 5: (a) Colored stereo point cloud, Point cloud using: (b) neither sub-pixel accuracy nor MLS, (c) only sub-pixel accuracy, (d) only MLS, (e) sub-pixel accuracy and MLS.

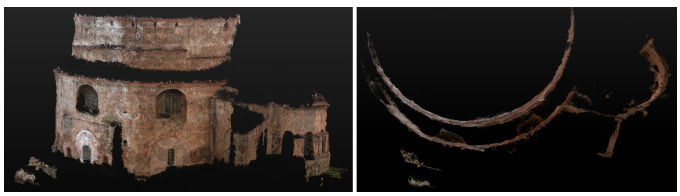


Fig. 6: Rotunda 3D Reconstruction

IV. Conclusions

This approach aims at generating accurate stereo point clouds using a time efficient and accurate methodology. The outliers are removed using an efficient two-phase strategy, while the accuracy of the generated 3D point cloud is improved by combining sub-pixel accuracy estimation and the MLS algorithm, which assist in achieving good reconstruction accuracy even for stereo images of low resolution.

The proposed approach could be exploited by multi-view algorithms, which attach great importance to the combination of sets of stereo point clouds and not to the computation of the individual stereo point clouds. For instance, the method in [8] uses a complex methodology that verifies the accuracy of each 3D point on more multiple depth maps and does not give weight to the individual stereo point cloud computation.

Future work will examine the exploitation of this approach, within a general framework that will also contain a methodology for the efficient combination of individual stereo point clouds.

Acknowledgment

The authors are grateful for support from the EU-funded IP project REVERIE under contract 287723.

References

[1] J. Pons, R. Keriven and O. Faugeras, *Multi-View Stereo Reconstruction and Scene Flow Estimation with a Global Image-Based Matching Score*, IJCV, vol. 72, pp. 179-193, 2007.

[2] S. Tran and L. Davis, *3D Surface Reconstruction Using Graph Cuts with Surface Constraints*, In Proc. ECCV, pp. 219-231, 2006.

[3] Y. Furukawa and J. Ponce, *Carved Visual Hulls for Image-Based Modeling*, IJCV, vol. 81, pp. 53-67, 2009.

[4] S. Fuhrmann and M. Goesele, *Fusion of depth maps with multiple scales*, ACM TOG, 2011.

[5] C. Strecha, W. von Hansen , L. Van Gool, P. Fua, U. Thoennessen, *On Benchmarking camera calibration and multi-view stereo for high resolution imagery*, In Proc. CVPR , 2008.

[6] Y. Furukawa and J. Ponce, *Accurate, Dense and Robust Multi-View Stereopsis*, PAMI, vol. 32, pp. 1362-1376, 2010.

[7] V. Hiep, R. Keriven, P. Labatut and J. Pons, *Towards highresolution large-scale multi-view stereo*, In Proc. CVPR, pp. 14301437, 2009.

[8] E. Tola, C. Strecha and P. Fua, *Efficient large-scale multi-view stereo for ultra high-resolution image sets*, Machine Vision and Applications, vol. 32, pp. 903-920, 2012.

[9] N. Salman, M. Yvinec, *Surface Reconstruction from Multi-View Stereo of Large-Scale Outdoor Scenes*, International Journal of Virtual Reality, 2010.

[10] N. Snavely, S. Seitz, and R. Szeliski, *Modeling the world from internet photo collections*, IJCV, vol. 80, pp. 189210, 2008.

[11] E. Tola, V. Lepetit and P. Fua, *Daisy: an efcient dense descriptor applied to wide baseline stereo*, PAMI, vol. 32, pp. 815-830, 2010.

[12] A. Fusiello and L.Irsara, *Quasi-Euclidean epipolar rectification of uncalibrated images*, Machine Vision and Applications, vol. 22, pp. 663-670, 2010.

[13] D.G. Lowe, *Distinctive Image Features from Scale Invariant Keypoints*, IJCV, vol. 20, no. 2, pp. 91-110, 2004.

[14] H. Bay, A. Ess, T. Tuytelaars, L. V. Gool, *SURF: Speeded Up Robust Features*, CVIU, vol. 110, no. 3, pp. 346-359, 2008.

[15] M. Muja and D. Lowe, *Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration*, VISAPP, 2009.

[16] S. Mattoccia, F. Tombari, and L. D. Stefano, *Stereo vision enabling precise border localization within a scanline optimization framework*, In Proc. ACCV, pp. 517-527, 2007.

[17] M. Levin, *Mesh-independent surface interpolation*, GMSV, Springer-Verlag, pp. 37-49, 2003.

[18] L. Di Stefano, F. Tombari, S. Mattoccia, *Segmentation-Based Adaptive Support for Accurate Stereo Correspondence*, In Proc. PSIVT, 2007.

[19] T. Liu, P. Zhang, and L. Luo, *Dense stereo correspondence with contrast context histogram, segmentation-based two-pass aggregation and occlusion handling*, In Proc. PSIVT, 2009.