On creating depth maps from monoscopic video using structure from motion

On Creating Depth Maps from Monoscopic Video using Structure from Motion

Ping Li[1], Dirk Farin[1], Rene Klein Gunnewiek[2], Peter H. N. de With[1,3]
Eindhoven Univ. of Technology[1] / Philips Research Eindhoven[2] / LogicaCMG[3]
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
{p.li, d.s.farin}@tue.nl / rene.klein.gunnewiek@philips.com / P.H.N.de.With@tue.nl

Abstract

The depth-image-based rendering technique is a promising technology for three-dimensional television (3D-TV) systems. For such a system, one of the key components is to generate a high-quality per-pixel depth map, particularly for already existing 2D video sequences. This paper proposes a framework for creating the depth map from uncalibrated video sequences of static scenes using the Structure From Motion (SFM) technique. This paper describes the architecture and the main components of the proposed framework. The initial experimental results show that SFM can be an effective way for creating the depth map, or it can be used to refine the depth map created by other methods, for example, the Depth From Cues (DFC) technique.

1 Introduction

Depth-image-based rendering (DIBR) 3D-TV has recently received much attention from both academia and industry. In contrast to the conventional stereoscopic video, where two separate video streams (one for the left eye and one for the right eye), need to be encoded and transmitted, only one monoscopic texture video and an associated per-pixel depth sequence need to be encoded and transmitted. This system has two clear advantages. First, it provides a good backward compatibility, since the monoscopic video can be decoded and displayed in a conventional 2D-TV system. Second, the depth information can be encoded with a much higher efficiency than the texture video. Only very little extra bandwidth is needed for transmitting the depth map. Also stereoscopic video systems have a good backward compatibility, but the extra bandwidth needed to transmit the extra view is much higher.

The depth map can be created either by a range camera or by converting the normal 2D video to 3D. During the introduction phase of 3D-TV, conversion of existing 2D videos to 3D is desired [1]. The paper proposes a framework for extracting the depth information from monoscopic videos. Our literature survey has revealed that the existing automatic depth-creation algorithms can be coarsely classified into two categories. One is the SFM approach and the other is the DFC technique that creates the depth from various depth cues such as the gravity, focus/defocus, occlusion, texture, etc.

The SFM approach exploits the physical relation between the motion in the image, motion of the camera, and the motion of the object in the 3D space. One major advantage of this method is that this relation can be well modelled using the pin-hole camera model, epipolar geometry, etc. However, the deficiency is that it cannot handle scenarios containing degenerated motion (e.g., rotation-only camera) or degenerated structure (e.g., coplanar scene) [2]. Moreover, applying SFM to non-static scenes with moving or deformable objects is still a difficult task. In this aspect, DFC has an advantage since it is capable of analyzing all kinds of scenes, including the scenes with moving and deformable objects. Howevever, a significant drawback of DFC is that the heuristic depth cues are hard to model due to the complexity of the scene interpretation. Obtaining an accurate and stable depth map is usually difficult for this type of algorithm. In view of the above observations, our proposed system attempts to integrate the SFM and the DFC methods to improve the depth creation. The system chooses SFM to create the depth map whenever SFM is applicable, as it can give a more stable and accurate depth map. In this case, the heuristic cues are used only as complimentary means for refining or creating the depth map for those parts of the scene where SFM cannot extract good depth information. When SFM is not applicable, our system relies on DFC to extract the depth.

This paper describes the proposed framework. Each of the main components of the proposed framework is briefly addressed. Though the overall framework is presented in this paper, the focus of this paper is on creating the depth using the SFM technique. An SFM algorithm is implemented and an initial depth map is created. The remainder of the paper is organized as follows. Section 2 describes the proposed architecture and its major components. Section 3 describes the SFM algorithm for depth creation from monoscopic videos. Section 4 presents some experimental results. Finally, Section 5 concludes this paper.

2 Architecture

As mentioned in Section 1, due to the inherent advantages and disadvantages of DFC and SFM, our architecture combines both approaches for a better depth creation. Fig. 1 shows the architecture of the proposed system, where we note that the overall architecture comprises of three major components, i.e., the scene analysis, the DFC block, and the SFM block. Though the figure shows the entire depth-creation algorithm, this paper is focusing on the SFM part. The Scene Analysis and the DFC remain as our future work. In this section, we will briefly describe the Scene Analysis and the DFC. SFM will be presented in more detail in Section 3.

2.1 Scene analysis

As we discussed in Section 1, depending on the scene contents, the system chooses either SFM or DFC to create the depth map. Thus, analyzing/classifying of the scene contents is the first step in our algorithm and it is crucial for automatic depth creation from monoscopic videos. During the scene analysis, the degenerate motion and structure are detected and the video sequence is partitioned into a number of sub-sequences, where

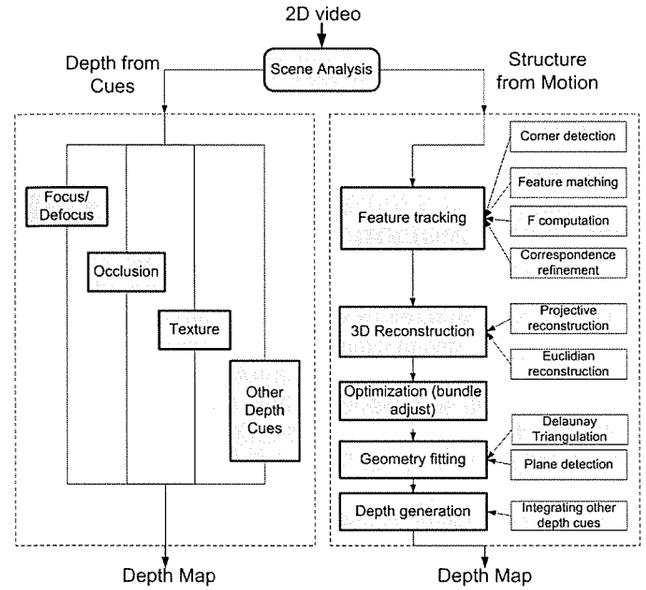


Figure 1: Architecture for depth generation.

either SFM or DFC can be applied. As we will be discussed in Section 3, we use a factorization-based approach for 3D reconstruction, in which the motion and structure for the set of images are computed at the same time. Appropriate partitioning of a long monoscopic video sequence into a number of sub-sequences so that factorization-based SFM can be applied, is very important for the automatic depth creation in our framework. Up to now, we did not yet realize this component. Some research on this topic can be found in [4].

2.2 Depth from cues

The Human Visual System (HVS) obtains depth information from both the disparity information provided by the two eyes and the visual information extracted by the human brain. In our application of creating depth maps from monoscopic videos, SFM is a process that analyzes the disparity information from multiple views in the time axis. To some extent, this process is similar to the disparity processing by the two eyes. Alternatively, DFC tries to extract the depth information by analyzing the heuristic depth cues in individual images. This process to some extent is similar to the visual processing by the brain to extract the depth information using its well-developed knowledge for scene interpretation.

To use the DFC for automatic depth creation from monoscopic videos by a computer, the depth cues such as the occlusion, focus/defocus, etc., must be somehow

described by mathematical models. Due to the little knowledge of the HVS and the complexity of the scene interpretation by the brain, the modelling of the heuristic depth cues is difficult. Obtaining an accurate and stable depth map is usually a problem for this approach. As such, DFC is only used as a fall back in the proposed system. When the scene analysis detects the degenerate motion or structure in the sub-sequence and SFM is not applicable, then DFC is applied.

3 Structure from motion

SFM refers to the problem of recovering the camera motion parameters and 3D scene geometry from a set of images captured by calibrated or un-calibrated cameras. SFM has been a very active research area in computer vision since the early 1980's. State-of-the-art SFM algorithms include the factorization-based approach [5] and the merge-based approach [3]. In general, merging algorithms rely on a good initial estimate of structure, and are also susceptible to drift over a long sequence [4]. Factorization methods calculate the motion and structure using all the tracked feature points with equal weights simultaneously. This approach has been proven to be accurate and robust to noise. This paper adopts the factorization-based approach. In the following, we will briefly describe and comment on the main steps of a the factorization-based SFM algorithm.

3.1 Feature tracking

As pointed out in Section 1, SFM exploits the relation between the camera motion and image motion. In practice, this relation is captured by the feature correspondences (points, lines, curves, surfaces, etc.), which has been extensively studied in past decades. An accurate feature correspondence is crucial to any SFM method.

Feature tracking for monoscopic videos is similar to that for multiple views in the sense that both are actually working on multiple images of the same scene from different viewpoints. However, feature tracking for monoscopic video does have its own unique characteristics. One of those is the strict camera-motion constraint. Unlike the multiple-view scenarios where both the external and internal parameters of the camera may change significantly across views, the camera parameters for a monoscopic video usually do not change abruptly. Exploiting this camera-motion constraint is expected to improve the feature tracking significantly, which make an investigation worthwhile. Furthermore, exploiting this feature may also help on our motion and structure recovery process. In the current implementation, the Harris corner detection [8, 9] is used to detect the feature points in the images. The detected feature points are then tracked along images using block matching. Future extension of this work could be to detect the *line* correspondences, which is expected to improve the quality of the 3D reconstruction in the areas that contain little texture but many strong edges.

3.2 Motion and structure recovery

This step is to compute the camera motion and scene structure, based on the detected feature correspondences. It can be divided into two sub-steps: the projective

reconstruction [6] to recover the projective depths for the 3D points and the Euclidian reconstruction [5] to enforce the metric constraints on the recovered camera parameters. In this paper, the factorization-based technique reported in [7] and [5] are used for our projective and Euclidian reconstructions.

3.3 Dense depth map creation using geometry fitting

The feature-based SFM only gives us a sparse depth map. To obtain the dense depth map that is required for our application, the Delaunay triangulation is used, which is shown in Fig. 4(a). The triangulation technique assumes that the complete scene consists of piece-wise planar surfaces described by the triangles. Generally, this assumption works well if the three vertices of the triangle are close to each other and lie in the same object. However, problems may arise in certain cases. As we can note from Fig. 4(b), the depth is not accurate for those triangles covering the edges of two objects and the transition area between foreground and background. Another problem of triangulation is that in some image areas where few feature points can be detected and tracked (the sky, the tree, and the ground in Fig. 2), triangulation is not applicable at all. We refer to these areas as *degenerate areas* in this paper. Extending or inferring the depth into these degenerate areas from their neighborhoods where structure can be reconstructed is desirable. This comes with the geometry fitting in our SFM process. Briefly, the fitting first detects the object geometry (e.g., plane), and then infers depth for the degenerate areas based on the detected geometry. Currently, we are thinking of using the color, texture and edge information together with the reconstructed 3D points to detect the object geometry.

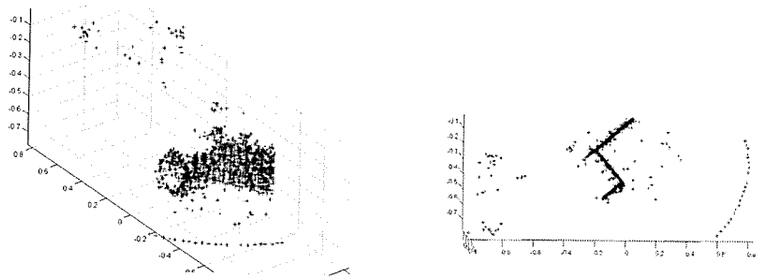
4 Experimental results and discussions

We have implemented an initial SFM algorithm for depth creation. In the algorithm, feature points are detected using Harris corner detection. Then, the feature points are tracked along a number of frames using a block matching technique [3]. After that, the factorization-based projective and Euclidian reconstruction are conducted to recover the camera motion parameters and the scene structure. Finally, dense depth maps are created using the Delaunay triangulation. In this section, we will present results that show each of these steps.

The *castle* sequence (Fig. 2) that is used in [3] for 3D reconstruction is used for our experiment. In the experiment, the feature points are tracked along the first 21 frames of the sequence (Fig. 2 also shows the tracked feature points in the first frame). Fig. 3 depicts the reconstructed scene geometry from two different viewpoints. The figure shows that the reconstructed structure is quite accurate. The three planes that corresponds to the three walls of the house as well as the orthogonality between the walls can be clearly seen from the top view of the reconstructed scene geometry. Furthermore, we also note that the locations of the 21 cameras are accurately recovered (at the bottom of Fig. 3(a) and at the right of Fig. 3(b)). However, for the ground, sky and tree areas where feature tracking is difficult, the structure cannot be reconstructed

Figure 2: Input image 0 and the tracked feature points.

or the reconstructed structure is very sparse.

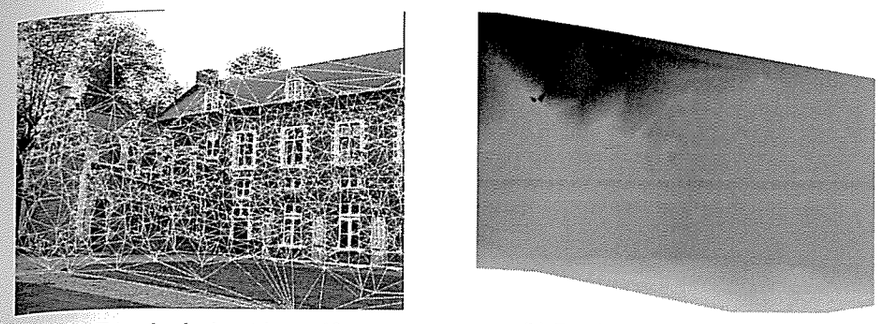


(a) Top-front view

(b) Top view

Figure 3: Reconstructed scene geometry.

Fig. 4(b) shows the dense depth map created using the simple triangulation. From the figure, we observe that the depth map is accurate and fits well to the real scene structure for most of the scene. However, as we explained in Section 3.3, problems occur in the following degenerate areas: 1) the ground, the tree and the sky where features are difficult to detect or to track; 2) the transition areas between the foreground and background where the depths are discontinuous; 3) the connection areas between the two objects, where the triangle covers the edges, e.g., the connection part between the ground and the three walls.



(a) Triangles for input image 12

(b) Depth map for input image 12

Figure 4: The triangles and the dense depth map for input image 12.

5 Conclusion and future work

In this paper, we have presented a framework for creating the per-pixel depth maps from monoscopic videos using the combination of SFM and DFC. The proposed algorithm uses SFM to create the depth map whenever it is applicable and DFC elsewhere. A special feature of our framework is the use of scene analysis to choose between SFM or DFC. An SFM algorithm is implemented for creating the depth map for the *castle* sequence. As a preliminary conclusion from our results, SFM can be used to create a good depth map from monoscopic videos. We expect that the accuracy of the depth map can be significantly improved over that obtained by the DFC-based algorithm because the modelling for DFC is complicated.

We have also observed shortcomings of the current implementation. Based on current results, we list the tasks below that need further investigation: 1) Exploiting camera-motion constraints to improve the feature tracking and the projective reconstruction; 2) Partitioning a long video sequence into sub-sequences for a robust depth creation using SFM; 3) Extending the depth to the degenerate areas and the degenerate sub-sequences where SFM is not applicable.

References

[1] Christoph Fehn, "A 3D-TV System Based On Video Plus Depth Information", *Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1529-1533, Nov. 2003.

[2] Philips H. S. Torr, Andrew W. Fitzgibbon and Andrew Zisserman, "The Problem of Degeneracy in Structure and Motion Recovery from Uncalibrated Image Sequences", MSR-TR-99-03, Microsoft Research, March 1999.

[3] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, R. Koch, "Visual Modeling With a Hand-held Camera", *International Journal of Computer Vision*, 59(3): 207-232, 2004.

[4] Yoon-Yong Jung, Yong-Ho Hwang, and Hyun-Ki Hong, "Frame Grouping Measure for Factorization-based Projective Reconstruction", *17th International Conference on Pattern Recognition*, vol. 4, pp. 112-115, 2004.

[5] Mei Han and Takeo Kanade, "A Perspective Factorization Method For Euclidean Reconstruction With Uncalibrated Cameras", *J. Visual. Comput. Animat.* 2002.

[6] Bill Triggs, "Factorization Methods for Projective Structure and Motion", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1996.

[7] Qian Chen and Gerard Medioni, "Efficient Iterative Solution to M-view Projective Reconstruction Problem", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, June 1999.

[8] Cordelia Schmid, Roger Mohr And Christian Bauckhage, "Evaluation of Interest Point Detectors", *International Journal of Computer Vision*, 37(2): 151-172, 2000.

[9] Wenxin Wang and Robert D. Dony, "Evaluation Of Image Corner Detectors For Hardware Implementation", *2004 Canadian Conference on Electrical and Computer Engineering*, vol. 3, pp. 1285-1288, May 2004.