On-line and Off-line 3D Reconstruction for Crisis Management Applications

Geert De Cubber

Royal Military Academy, Department of Mechanical Engineering (MSTA)

Av. de la Renaissance 30, 1000 Brussels

geert.de.cubber@rma.ac.be

1. Introduction

When confronted with a large crisis, the crisis management teams require a global overview of the crisis scene. In practical situations, however, it is near impossible to obtain such a global overview, due to the abundance of information coming from different sources and the lack of a global model of the crisis scene where all this information can be nicely visualized upon. However, in reality, such 3D models are almost never readily available. As such, they must be constructed on-site in an automated way. Therefore, we applied the presented 3D reconstruction approach to images shot by a semi-autonomous robotic agent [2]. The idea is that this robot is sent out by the crisis management teams to autonomously build up a model of the environment, which can then be used for a better assessment of the situation. This application poses an extra difficulty to the 3D reconstruction approach, as it is required to handle relatively large outdoor environments.

In this paper, we propose an automated 3D reconstruction approach for building a global 3D model. This 3D reconstruction approach is based on dense structure from motion (SFM) recovery from images captured by a camera on-board a semi-autonomous crisis management robot. Dense SFM algorithms aim at estimating a 3D location for all image pixels. The most modern existing dense SFM algorithms minimize the optical flow constraint and enforce smoothness in the depth field in a variational framework. However, due to the noisiness of the optical flow and due to projection ambiguities (leading a.o. to occlusions), these algorithms are still not very robust when confronted with unconstrained 3D camera motion and changing illumination conditions. One could argue that these problems are due to the fact that dense SFM is a relatively new field of research. In this context, we present a methodology which is able to cope with these problems, due to an integration of more reliable sparse motion data, next to the dense motion information. This method is capable of estimating a high quality 3D reconstruction of a scene, which can provide a valuable tool for the crisis management teams. However, this approach has one main disadvantage: with the current state of the art in computing power, it is not possible to process this data in real-time or near-real-time. While awaiting faster computer processing technology, it is thus only possible to use this 3D reconstruction technique as an off-line tool, e.g. for evaluation purposes or for post-disaster needs assessment.

In order to offer the crisis management teams a 3D perception and visualization technology which can directly put to use, we present in this paper also another 3D reconstruction methodology. This approach fuses dense stereo and sparse motion data to estimate high-quality instantaneous depth maps. This methodology achieves near-real-time processing frame rates, such that it can be directly used on-line by the crisis management teams.

*This paper is (co-) funded by the FW6- **IST-045541-**View-Finder Project*

2. Off-line 3D Reconstruction using Dense Structure from Motion

2.1. Methodology

To address the classical dense structure from motion shortcomings, we adopt a dual approach for dense structure estimation, trying to combine the robustness of sparse reconstruction techniques with the completeness of dense reconstruction algorithms. Figure 1 explains the proposed methodology.

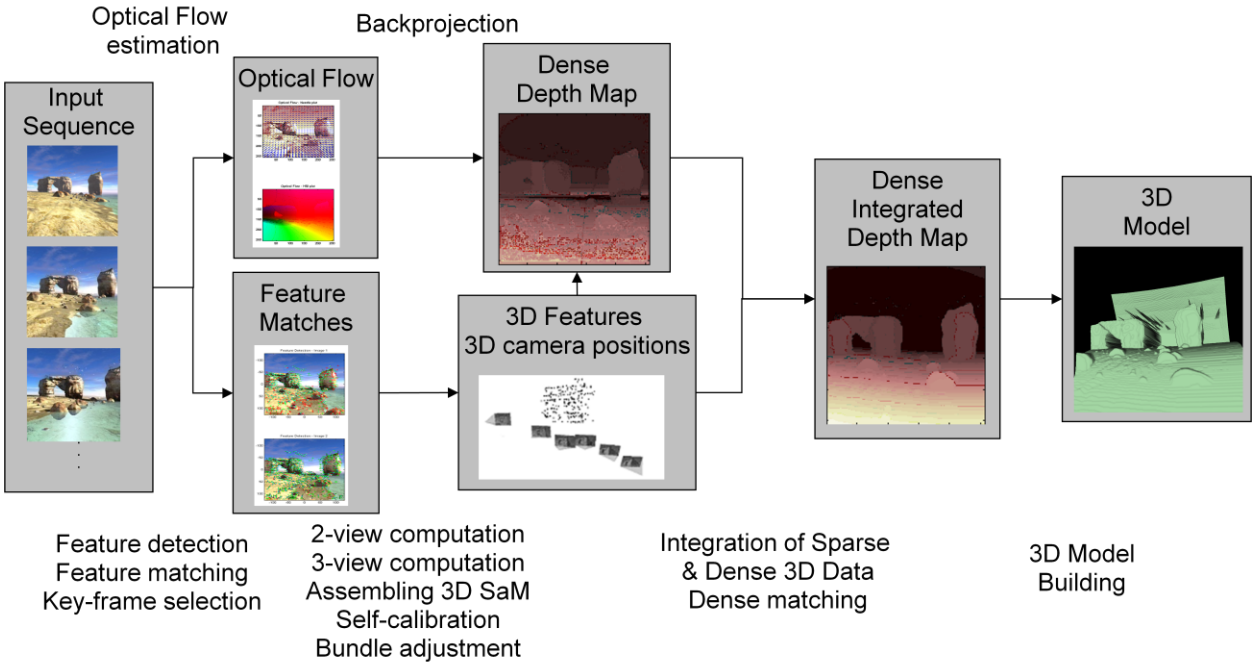


Figure 1: Dense structure from motion processing cue

The first step of the proposed methodology consists of solving the sparse reconstruction problem. Therefore, we used the structure from motion approach presented by Hartley and Zisserman in [3]. These results then serve as initial guesses for the dense reconstruction process, which fuses the sparse data with dense information coming from a densely estimated optical flow field. The optical flow ***u*** is a projection of the 3D motion field and is related to the structure through the rigid motion equation. This relation between optical flow and structure and motion on one hand and the available sparsely reconstructed structure and motion parameters allow for integrated sparse-dense reconstruction. A variational approach is used to tackle this high-dimensional data fusion problem. This methodology formulates the problem of fusing dense image data - in the form of the image brightness constraint from the optical flow – with sparse data - in the form of the epipolar constraint of the sparse reconstruction - as an optimization problem. The result of this optimization approach is a high-quality depth map. Multiple of these depth maps can be integrated to form a global 3D model of the scene, filmed by the camera system.

2.2.Results

The presented 3D reconstruction approach was tested during an integrated crisis management exercise, where an airplane crash was simulated. During this exercise, a semi-autonomous robot [2] was asked by the firefighters to search for human survivors (in this case: the pilot who ejected from the airplane before the crash) near the incident site and while doing this, it was requested to build a 3D model of the environment. Figure 2 shows some images taken by the robot on-board camera during this test, whereas Figure 3 shows the reconstructed 3D model.

Figure 2: Some frames shot by the semi-autonomous robot during the crisis management exercise

The 3D model of Figure 3 shows a good resemblance to the physical nature of the environment and all required features can be identified: the ground plane, the bunker in the back, the canopy... As also the motion of the camera (which is fixed on the robot) is reconstructed using the presented methodology, the robot can be positioned in the virtual environment. As an example of how this 3D model can be efficiently used by crisis management teams, the 3D model of Figure 3 also indicates the position of a human survivor. The presence of the human survivor was detected by a human victim detection algorithm, presented in [1] and this information was fused with the 3D information obtained through the presented depth reconstruction algorithm to locate and visualize the victim in the 3D model.

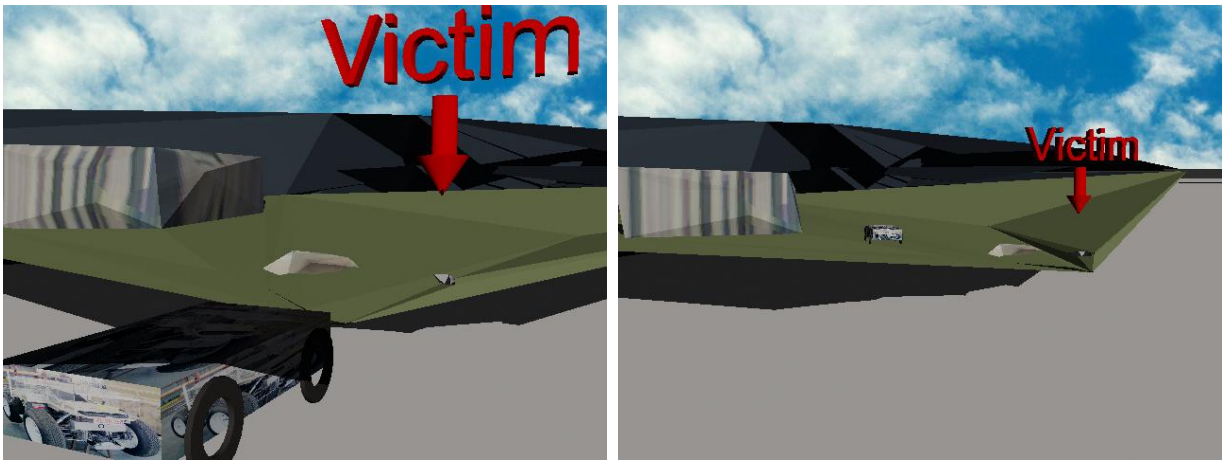


Figure 3: Reconstructed 3D model of the environment, showing the camera/robot position and an indication of the presence of human survivors, by fusing the 3D information with the output of a human victim detector algorithm [1].

3. On-line 3D Reconstruction using Sparse Structure from Motion

3.1. Methodology

Stereo vision is one of the most active fields of research within the computer vision community. Hence, stereo algorithms have gained great maturity over the past decades. The focus of this research work is not to develop new stereo vision algorithms, but rather to investigate if and how the addition of motion data could improve existing stereo algorithms. Therefore, a short overview is given of the working principles of different stereo algorithms. For this, we base ourselves on the excellent taxonomy [5] on dense two-frame stereo vision algorithms, written by Scharstein and Szeliski. This taxonomy is based on the observation that most stereo algorithms perform the following four steps**Error! Reference source not found.**:

1. *Matching cost computation:*

In this step, the stereo algorithm searches a range of disparities and accords a matching cost to each disparity for each pixel. The most popular matching costs are the Absolute intensity Differences (SAD) and the Squared intensity

Differences (SSD). However, next to stereo, structure from motion could also deliver an estimate of the new disparity level for each pixel. Consider 2 left and right stereo images I_{11} and I_{12}, shot at time t_0 by a calibrated stereo rig such that R_{Stereo} and t_{Stereo} are known. In both images, the projections x_{11} and x_{12} of the same 3D point X are visible. At time t_{0+k}, 2 new stereo images, I_{21} and I_{22}, are shot by the same stereo vision system. Structure from Motion is applied on the left images and right images. Here, we use the sparse structure from motion approach presented by Hartley and Zisserman in [3]. This leads to an estimation of the inter-frame camera motion, R_{Motion} and t_{Motion}, and camera projection matrices P_{11}, P_{12}, P_{21}, P_{22} for the 4 cameras. Following the stereo and motion transformations, x_{22} can be written as a function of x_{21}: $x_{22} = P_{22}T_{Motion}^{-1}T_{Stereo}^{-1}T_{Motion}P_{21}^{-1}x_{21}$. The disparity can then be calculated directly from the pixel positions x_{21} and x_{22}: $d_{Motion} = \left(P_{22}T_{Motion}^{-1}T_{Stereo}^{-1}T_{Motion}P_{21}^{-1} - Id\right)x_{21}$. The disparity, estimated through sparse structure from motion as presented above, is then used as an extra term for the cost function: $Cost_{Total} = Cost_{SAD/SSD} + \left(d_{Motion}^2 - d_{Stereo}^2\right)$. It is evident that the computational cost of the matching cost computation rises with the requested maximum disparity, which is a measure for the precision.

2. *Cost support aggregation:*

As in classical stereo processing techniques, a shiftable window is used for cost support aggregation.

3. *Disparity computation and optimization:*

For global methods, this processing step is the most important one. These methods are often formulated as an energy minimization problem with a data term $E_{data}(d)$ representing how well the disparity function d fits with the input image pair and a smoothness term $E_{smoothness}(d)$ enforcing smoothness constraints, to form an energy function $E(d) = E_{data}(d) + \lambda E_{smoothness}(d)$. The definition of the smoothness term is crucial. Indeed, matching of image intensities typically fails on monochrome surfaces, because due to the fact that there are multiple solutions, the numerical stability cannot be assured. What is needed to solve this problem is a regularization term which extrapolates and smooths the structural data over pixels which belong to the same physical object at the same distance. As such, we interfere in the third step of the stereo computation algorithm, by including a smoothness term $E_{smoothness}(d)$. The main problem for smoothing is the preservation of discontinuities. Indeed, regularization should not over-smooth the solution such that depth discontinuities are no longer visible. Nagel and Enkelmann took into account this consideration and proposed in [4] an anisotropic smoothing term which preserves the depth discontinuities. The Nagel and Enkelmann regularization model has already been proven successful in a range of independent experiments and formulates a regularization term of the form: $E_{smoothness} = \nabla\zeta^T \mathbf{D}\left(\nabla I_1\right)\nabla\zeta$, where $\mathbf{D}$ is a regularized projection matrix. Using this approach, discontinuities can be preserved.

4. *Disparity refinement or post processing:*

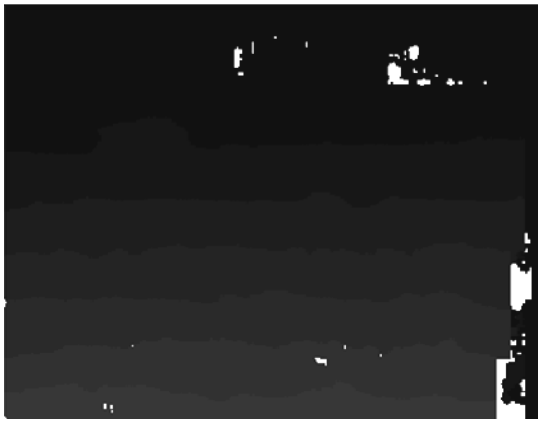
In the course of this work, we do not consider"post-processing" steps, as they are too application dependent.

3.2.Results

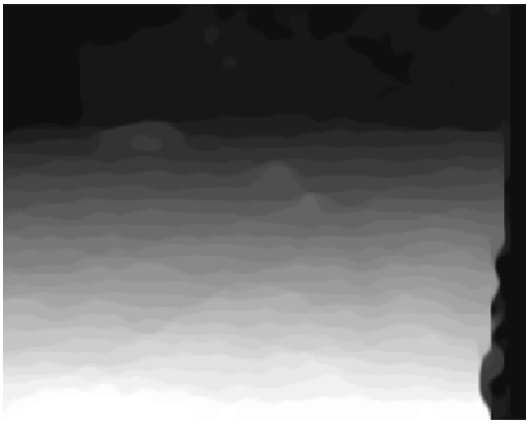
Figure 4 shows the results of the presented dense motion-augmented stereo estimation algorithm. By comparing the disparity estimation result of the presented methodology, as presented by Figure 4d to the result of the classical stereo approach without the presented augmentations of Figure 4c, it is evident that the proposed methodology outputs superior results compared to the classical method. In contrast to the pure stereo result, the disparity map as estimated by our method presents no disturbing holes, the depth gradient of the ground plane is well-visible, the 2 obstacles on the ground can also be easily discerned on the disparity map, and even the building, which is very far away, can be distinguished on the disparity map of Figure 4d. The processing time required to estimate a dense depth map using the presented methodology is about 1 second, which is still reasonable for near-real time applications and it is to be expected that in the near future, with the constant increase in processing power, the calculation time will go down substantially, allowing full real-time operation.

a) Left Image

b) Right Image



c) Disparity Map from Pure Stereo



d) Disparity Map from Dense Stereo + Motion

Figure 4: Dense Motion-augmented Stereo Estimation

4. Conclusions

In this paper, we have presented 2 approaches for the extraction of 3D information from a scene from visual data for crisis management applications. A first, off-line, dense structure from motion based approach mixes sparse and dense motion data in a variational framework, providing a high quality 3D reconstruction of the scene. The visualization of the virtual 3D scene with added localized information, as presented by Figure 3, provides a powerful tool for the human crisis management teams to augment their situational awareness without increasing the cognitive load too much, as the whole process of data acquisition by the robot and processing by the presented algorithm is automated. A second, on-line, methodology mixes sparse motion and dense stereo data in an integrated framework, providing high-quality depth maps at near-real-time framerates

References

[1]. G. De Cubber, G. Marton. Human Victim Detection. Third International Workshop on Robotics for risky interventions and Environmental Surveillance-Maintenance, RISE'2009- Brussels, Belgium. January 2009.

[2]. D. Doroftei, G. De Cubber, E. Colon and Y. Baudoin. Behavior Based Control For An Outdoor Crisis Management Robot., RISE'2009- Brussels, Belgium. January 2009.

[3]. Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, Second edition, 2004.

[4]. H.H. Nagel, W. Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Tran. on Pat. An. and Machine Intelligence*, 8(5):565--593, 1986.

[5]. Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vision, 47(1-3):7–42, 2002.