

# SHAPE FROM VIDEO V.S. STILL IMAGES

Christoph Strecha, Frank Verbiest, Maarten Vergauwen, Luc Van Gool  
KU Leuven ESAT/PSI Belgium

**KEY WORDS:** Wide-baseline stereo, 3D model building, optical flow, minimal path algorithm

## ABSTRACT

In this paper we compare two dense matching approaches. The first one has been developed in the context of shape from video using minimal path search. The second one is a PDE-based approach, and would be expected to give better results for shape from (a small number of) still images and for wide baseline situations. Both methods use as input the fully calibrated camera parameters, that have been obtained after structure from motion recovery from uncalibrated images. Emphasis lies on the usage of only a small amount of (high resolution) images instead of a (low resolution) video sequence. We use ground truth synthetic data as well as real data to compare the two algorithms.

## 1 INTRODUCTION

During the last few years more and more user-friendly solutions for 3D modeling have become available. Techniques have been developed (Armstrong et al., 1994, Heyden and Åstrom, 1997, Hartley and Zisserman, 1998, Pollefeys et al., 1998) to reconstruct scenes in 3D from video or images as the only input.

The strength of these so-called shape-from-video techniques lies in the flexibility of the recording, the wide variety of scenes that can be reconstructed and the ease of texture extraction.

Typical shape-from-video systems require large overlap between subsequent frames. This requirement is typically fulfilled for video sequences. However, video sequences usually have low resolution. Digital cameras today, on the other hand, have resolutions in the order of thousand  $\times$  thousand pixels and more, with far less noise than the video output. Considering this fact, we want to evaluate new possibilities for building accurate 3D models from high resolution cameras. Using such a camera and taking only a few shots of the object or scene is a new interesting application for dense 3D model building, usually referred to as wide-baseline stereo. This opens up a new range of applications as indeed, it may not always be possible to record continuous video of the object of interest, e.g. due to obstacles or time pressure.

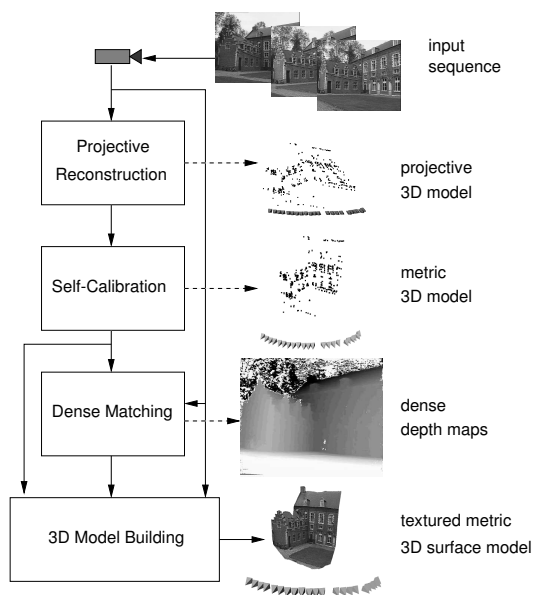


Figure 1: *The structure from video pipeline*

The structure from video pipeline in fig. 1 (Pollefeys et al., 1998), that has already been used successfully for small-baseline situations, needs to be changed in possibly two parts for wide-baseline conditions. First, one requires initial feature matching that should be able to cope with the large change in baseline or scale. This problem has been investigated by (Tuytelaars and Van Gool, 2000, Baumberg, 2000, Matas et al., 2002, Mikolajczyk and Schmid, 2002). They use affine invariant regions for matching images taken from very different view points. An extension of this work to multiple views has recently been published by (Ferrari et al., 2003). With these techniques, it is possible

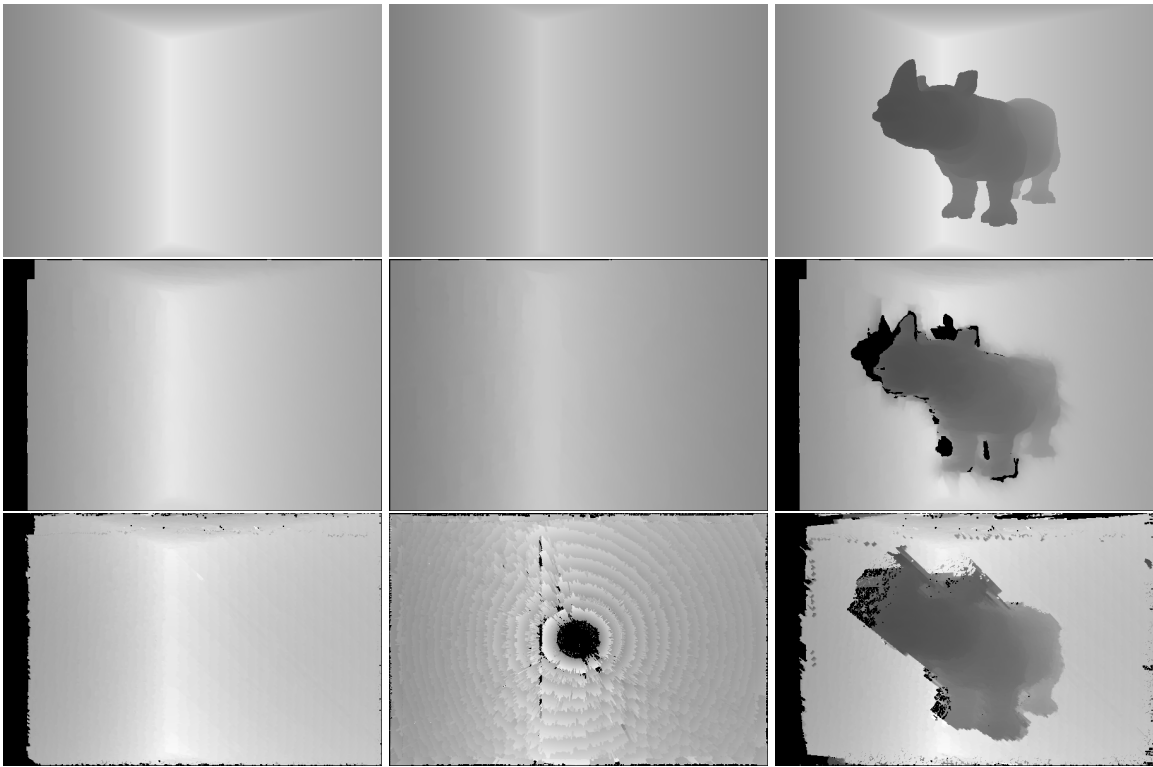


Figure 2: *Depth maps for three (left, middle, right) synthetic scenes: Top ground truth, Middle PDE-based approach, Bottom: Minimal path algorithm*

to self calibrate images in the same way as before under small-baseline conditions (fig. 1) (Pollefeys et al., 1998).

The second change - and this is the focus of this paper - is the dense matching that extends the sparse Euclidean point model to a full 3D model (see fig 1). We compare a minimal path search algorithm using dynamic programming (Van Meerbergen et al., 2002) with a partial differential equation (PDE) based approach (Strecha and Van Gool, 2003, Strecha and Van Gool, 2002) As explained in more detail in section 2, dense matching using the minimal path search algorithm involves the solution of the stereo problem for pairs of rectified images and the fusion of these to a depth map. Especially when dealing with a small amount of images, one would like to combine these two steps into a single estimation scheme. The advantage becomes clear by considering the fact that depending on the geometry some pixels in image 1 could be matched more reliably to image 2 while others have more reliable matches in image 3. An integrated scheme would be able to take advantage of these locally changing reliabilities, and is part of our PDE-based approach discussed in section 3.

Another advantage of the PDE-based approach is that matching is not quantized to a predefined precision, which results in smoother 3D models. Hence, one would expect more accurate results for the PDE-based approach.

On the other hand, since the PDE-based approach is more restricted to the search of corresponding points along lines provided by the geometry of the calibration, this method is more sensitive to inaccuracies in the calibration process. So in case of non-precise calibration one would expect to have less reliable models.

## 2 DYNAMIC PROGRAMMING BASED DENSE MATCHING

With the camera calibration extracted for all viewpoints of the sequence, we can proceed with methods developed for calibrated structure from motion algorithms. The structure from motion algorithm already delivers a sparse surface model based on distinct feature points. To obtain a more detailed model of the observed surface a dense matching technique is used.

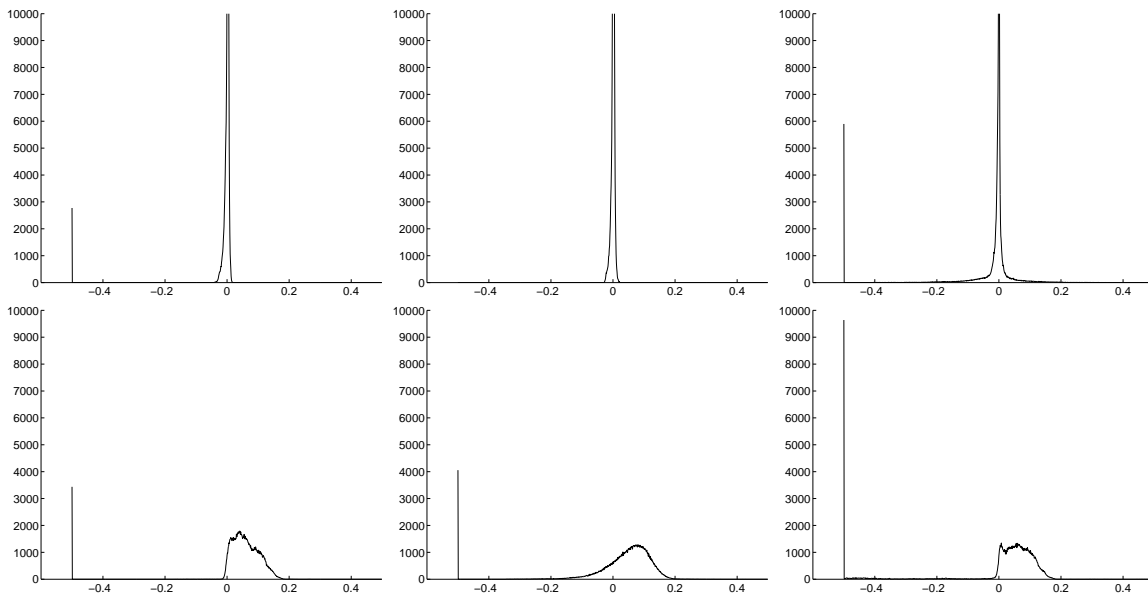


Figure 3: *Synthetic results: Histogram of the relative deviation of the estimated depth maps from the ground truth for each pixel (from fig 2 in the same order); Top: PDE-based approach, Bottom: Minimal path algorithm*

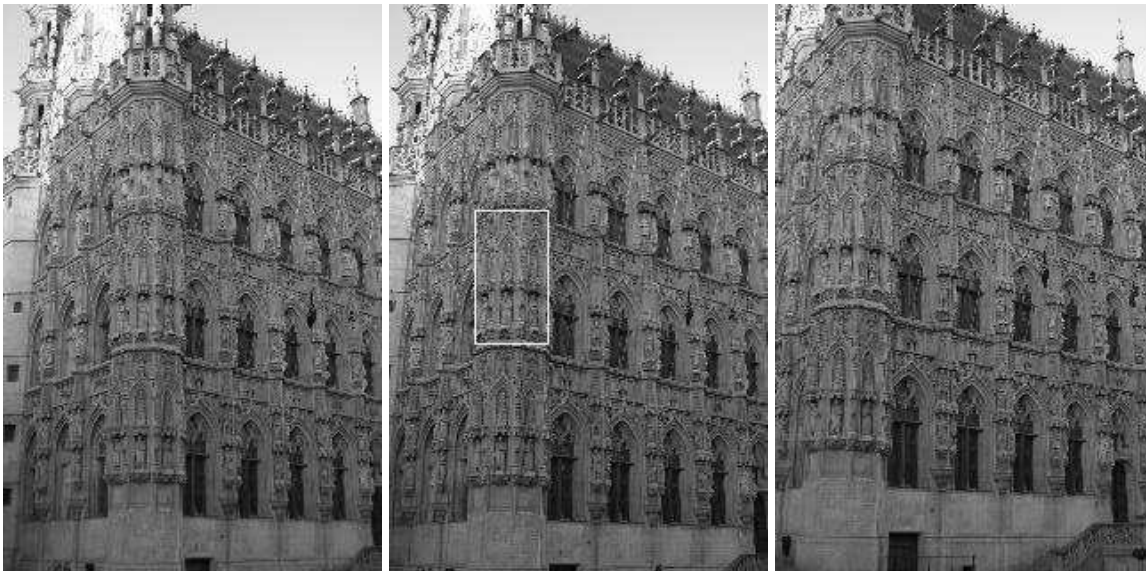


Figure 4: *Three high resolution images ( $3072 \times 2048$  pixels) used in the experiments*

Since the calibration between successive image pairs was computed, the epipolar constraint that restricts the correspondence search to a line can be exploited. Image pairs are warped so that epipolar lines coincide with the image scan lines. For this purpose the rectification scheme proposed in (Pollefeys et al., 1999) is used. This approach can deal with arbitrary relative camera motion which is not the case for standard homography-based approaches which fail when the epipole is contained in the image. The approach proposed in (Pollefeys et al., 1999) also guarantees minimal image size. The correspondence search is then reduced to matching along each image scan-line. This results in a dramatic increase of the computational efficiency of the algorithms by enabling several optimizations in the computations. In addition to the epipolar geometry, other constraints like preserving the order of neighboring pixels and bidirectional uniqueness of the match are exploited. However, these are reasonable assumptions for small baseline stereo only. Under general wide-baseline conditions they cannot be made. In (Van Meerbergen et al., 2002) these constraints *are* used to guide the correspondence towards the most probable scan-line match using a dynamic programming scheme. The matcher searches at each pixel in one image for maximum normalized cross correlation in the other image by shifting a small measurement window along the corresponding scan line. The disparity



Figure 5: *Untextured and textured models zoomed to the white rectangle of fig. 4; left: PDE-based using three high resolution images; right: dynamic programming using the same input*

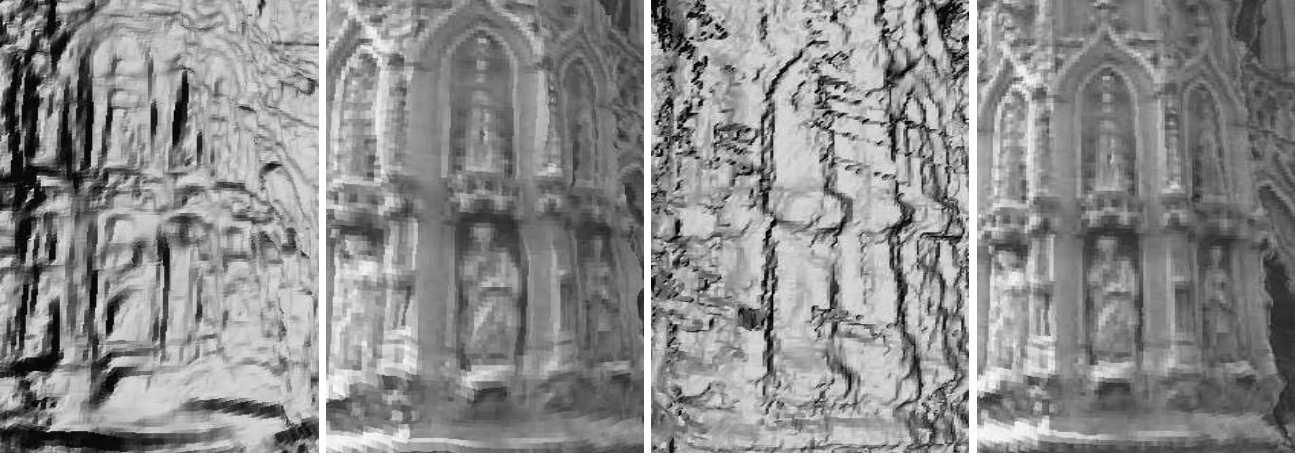


Figure 6: *Untextured and textured models of the algorithms for the low resolution video input captured from the same positions as the images in fig. 4; left: PDE-based approach; right: Dynamic programming*

search range is limited based on the disparities that were observed for the features in the previous calibration stage.

The pairwise disparity estimation allows to compute image to image correspondence between adjacent rectified image pairs up to pixel precision. It yields independent discretized depth estimates for each camera viewpoint. An optimal joint estimate is achieved by fusing all independent estimates into a common 3D model using a Kalman filter. The fusion can be performed in an economical way through controlled correspondence linking and was discussed in more detail in (Koch et al., 1998). This approach can provide a very dense depth map while avoiding most occlusions. The depth resolution is increased through the combination of multiple viewpoints. Larger global baselines are exploited based on the matching of series of subsequent images under small local baseline conditions.

### 3 PDE-BASED DEPTH EXTRACTION

Due to space limitations, we only describe the main idea of the algorithm. For details, we refer to (Strecha and Van Gool, 2003, Strecha and Van Gool, 2002). Given  $N$  images  $i = 1..N$ , that have been calibrated, our PDE-approach is based on the minimization of the following cost functionals for the different cameras, written here in terms of the  $i$ th camera.

$$E_i[d_i] = \int_{\Omega} \sum_{i \neq j}^N c_{ij}(\vec{x}) |I_i(\vec{x}) - I_j(\vec{l}(\vec{x}, d_i))|^2 d\vec{x} + \lambda \int_{\Omega} (\nabla d_i)^T D(\nabla C_i) \nabla d_i d\vec{x} \quad (1)$$

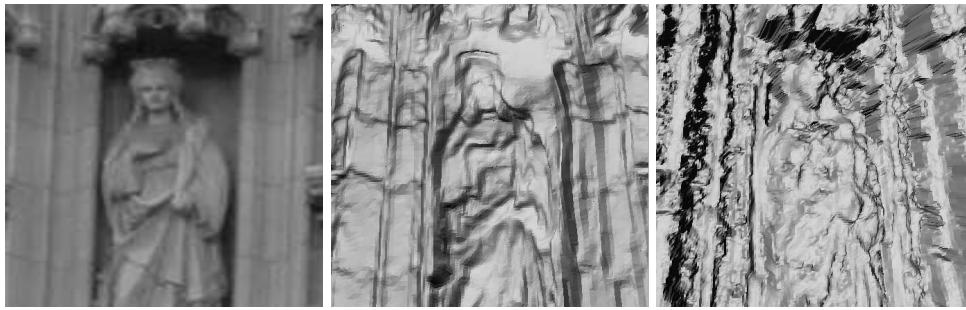


Figure 7: *Untextured and textured models zoomed to the woman's head seen in the models of fig. 5; from left to right: Zoomed image; PDE-based using three high resolution images; dynamic programming using three high resolution images*

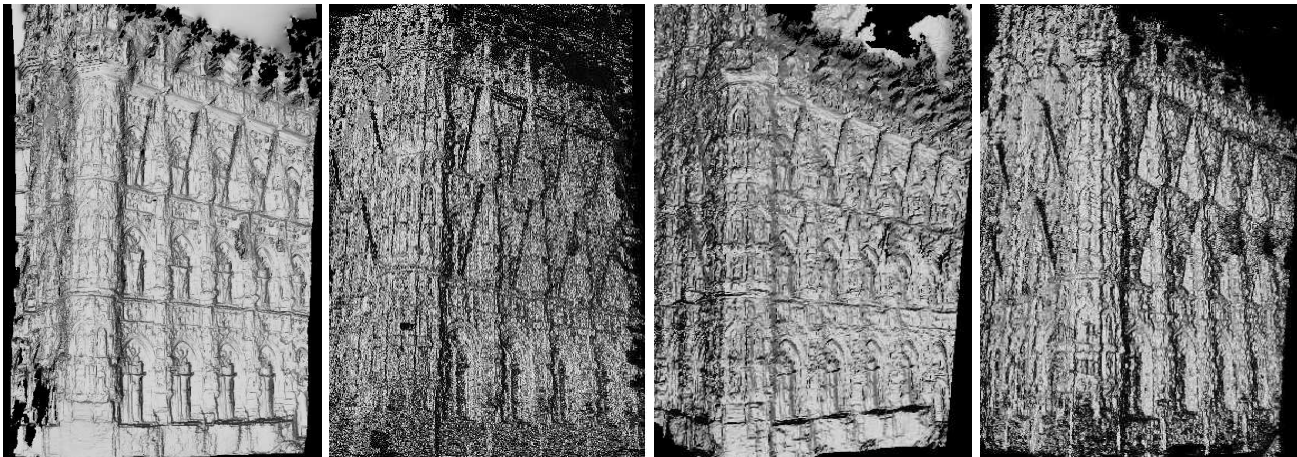


Figure 8: *Un textured full models: from left to right: PDE-based using three high resolution images; dynamic programming using three high resolution images; PDE-based using using video input; dynamic programming using video input*

The minimum of this energy for all cameras (usually not more than 10) is estimated by a system of coupled PDE's. The dense matching of the images does not need a rectification of image pairs. Our PDE-based approach uses a parameterization  $(\vec{l}(\vec{x}, d_i))$  of corresponding pixels by the inverse depth  $d_i$ . The advantage is the possibility to match a pixel in one image to the corresponding pixels in **all** other images using only this single parameter. The inverse depth map  $d_i$  parameterizes in fact a 3D mesh textured with image  $I_i$ . By projecting this mesh to the other cameras  $i \neq j$  one can compute the difference of it with each of the other images  $i \neq j$ . This is realized by the first term in equation 1. The contribution for a match from image  $i \rightarrow j$  is weighted (by  $c_{ij}(\vec{x})$ ) to take care of possible occlusions. The weighting factor is computed dynamically by the difference in matches from image  $i$  to image  $j$  and vice versa. This difference should be zero in most cases indicating a correct match with no occlusions. However, in case the difference is not zero we either have an occlusion or a situation where the match is not yet established. With this weighting mechanism individual pixels can take advantage of their favorite view. The second term in eq. 1 regularizes the depths  $d$ . It forces the solutions to be smooth, while preserving depth discontinuities through anisotropic diffusion. The initialization of the depth maps  $d_i(\vec{x})$  is zero, except for pixels where a depth estimate already exists from the calibration procedure. Using inhomogeneous time diffusion (Strecha and Van Gool, 2003) the diffusion time for these pixels is lowered in order to keep their depth value almost fixed. This results in attracting the other pixels to the minimum of the above energy 1 very efficiently.

## 4 RESULTS

First we applied the algorithms to synthetic data. The cameras (external and internal parameters) are known in this case and have not been extracted by self calibration. Both algorithms used the same



Figure 9: *Four images used in the wide-baseline experiment*

parameter setting for all the three synthetic experiments.

Three images of three small-baseline scenes have been used in order to extract the depth maps seen in fig. 2. The left side depth maps in fig. 2 are made from a  $x$  and  $y$  translating camera scene.

In the middle column one can see the interesting case where the epipole lies inside the image ( $z$ -translating camera). To the PDE-base approach this does not pose any problem. Dynamic programming has difficulties near the epipole. The right images show the behavior with occlusions for a camera path translating in  $x$ - $y$  direction. The deviation of both algorithms from the ground truth is shown in figure 3. We show the histogram of the relative depth deviation from the ground truth

$H((D_{groundtruth} - D_{estimated})/D_{groundtruth})$   
 for the three synthetic experiments. The histogram value at  $-0.5$  collects all the pixels with a depth value outside the deviation range (usually occluded pixels or pixels where no depth value could be assigned). The error of the dynamic programming approach is mainly due to the matching with pixel accuracy that results in a discretization of the possible depth values. The computational speed of the algorithms is in the same order of magnitude (for the synthetic scene:(82, 119, 76sec. dynamic programming, 80, 98, 113sec. PDE-based).

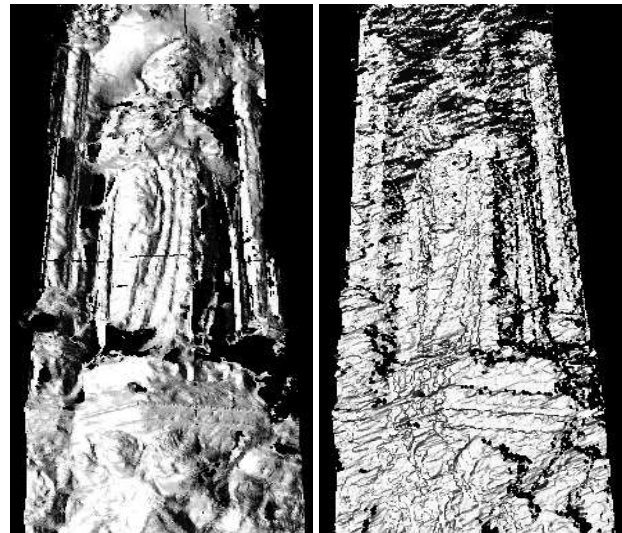


Figure 10: *Details of the wide-baseline experiment: PDE-based (left) and dynamic programming (right)*

Two real sequences have been captured in a small-baseline experiment. One video sequence with resolution  $720 \times 576$  pixels and a sequence of images with a digital camera of  $3072 \times 2048$  pixels, both following approximately the same path. Figure 4 shows three images from the high resolution camera. These three images have been used to extract the models of figs. 5, 7 and 8. We show the textured and untextured models for the part within the white rectangle in fig. 4. Figure 6 shows the reconstruction of the scene by using a low-resolution video captured approximately from the same camera path. An even more detailed view of the models from both algorithms, can be seen in fig.7. Figure 8 shows the complete model. One can see that the dynamic programming approach can

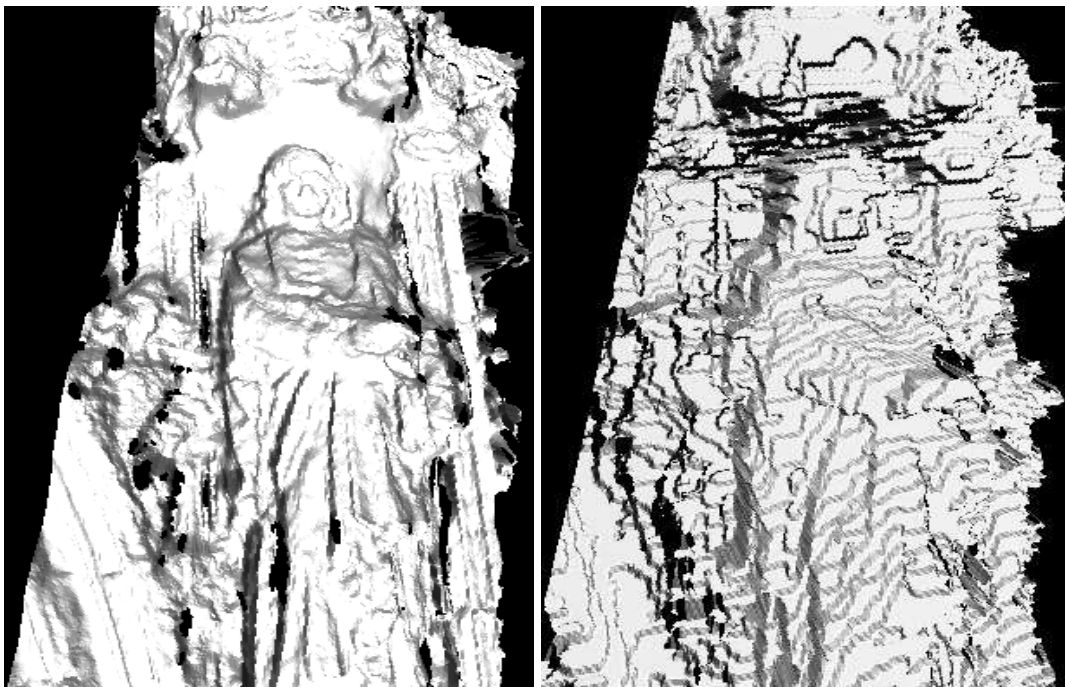


Figure 11: *Details of the wide-baseline experiment: PDE-based (left) and dynamic programming (right)*

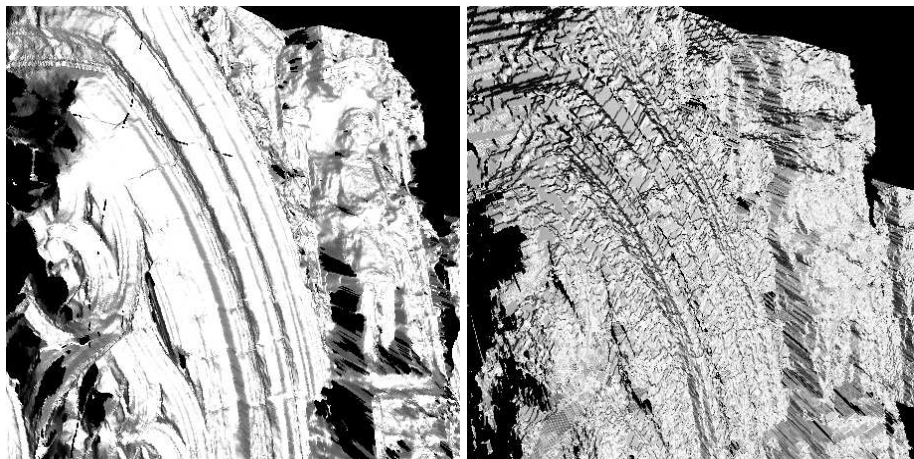


Figure 12: *Details of the wide-baseline experiment: PDE-based (left) and dynamic programming (right)*

produce good results for a low resolution video input (right). Because of linking much more views this model appears to be better than dynamic programming using only three images (middle-left). However, our PDE-based approach has the best results in all cases as one can see also in figures 5 and 7.

The last experiment is a wide-baseline high resolution experiment, where the images are too far apart to calibrate them with small-baseline techniques. Instead we used the affinely invariant regions of (Tuytelaars and Van Gool, 2000, Ferrari et al., 2003) to find the initial matches. Figure 9 shows the original images. Zoomed models of the parts indicated by a white rectangle for both algorithms of the top-left image in this figure can be seen in figures 10, 11 and 12. Since the matching is not restricted to pixel accuracy in the PDE-based approach the models are more smooth and detailed. The effect of the restriction to pixel accuracy for the dynamic programming approach is clearly visible in figure 11(right).

## 5 CONCLUSIONS

Using our PDE-based dense depth estimation scheme one would indeed be able to replace a structure from video pipeline by structure from high resolution still images. This approach gives also accurate results when using only a few images. In that case the PDE-based approach outperforms even the dynamic programming using many more images. By defining the correspondence search directly in 3D and using a fine grained *subpixel* accuracy matching the PDE-based approach outperformed the dynamic programming approach in the level of detail and in the applicability to more general scenes. Further work should evaluate the precision of the calibration. This work can be seen as a step forward from structure from motion being a tool to compute visually attractive 3D models to becoming a real 3D measurement system.

**Acknowledgment:** The authors gratefully acknowledge support by K.U.Leuven GOA project ‘VHS+’ and EU IST project ‘CogViSys’.

## REFERENCES

- Armstrong, M., Zisserman, Z. and Beardsley, P., 1994. Euclidean structure from uncalibrated images. 5th BMVC.
- Baumberg, A., 2000. Reliable feature matching across widely separated views. CVPR pp. 774–781.
- Ferrari, V., Tuytelaars, T. and Van Gool, L., 2003. Wide-baseline multiple-view correspondences. CVPR pp. 718–725.
- Hartley, R. and Zisserman, A., 1998. Multiple View Geometry in Computer Vision. Cambridge University Press.
- Heyden, A. and Åstrom, K., 1997. Euclidean reconstruction from image sequences with varying and unknown focal length and principal point. CVPR.
- Koch, R., Pollefeys, M. and Van Gool, L., 1998. Multi viewpoint stereo from uncalibrated video sequences. Proc. ECCV’98 pp. 55–71.
- Matas, J., Chum, O., Urban, M. and Pajdla, T., 2002. Robust wide baseline stereo for maximally stable external regions. BMVC pp. 414–431.
- Mikolajczyk, K. and Schmid, C., 2002. An affine invariant interest point detector. ECCV 1, pp. 128–142.
- Pollefeys, M., Koch, R. and Van Gool, L., 1999. A simple and efficient rectification method for general motion. ICCV pp. 496–501.
- Pollefeys, M., Koch, R., Vergauwen, M. and Van Gool, L., 1998. Metric 3d surface reconstruction from uncalibrated image sequences. Proc. SMILE Workshop (post-ECCV’98), LNCS 1506 pp. 138–153.
- Strecha, C. and Van Gool, L., 2002. Pde-based multi-view depth estimation. 1st Int. Symp. of 3D Data Processing Visualization and Transmission pp. 416–425.
- Strecha, C. and Van Gool, L., 2003. Dense matching of multiple wide-baseline views. accepted for ICCV 2003.
- Tuytelaars, T. and Van Gool, L., 2000. Wide baseline stereo matching based on local, affinely invariant regions. Proc. British Machine Vision Conference - BMVC pp. 412–422.
- Van Meerbergen, G., Vergauwen, M., Pollefeys, M. and Van Gool, L., 2002. A hierarchical stereo algorithm using dynamic programming. International journal of computer vision - special issue on stereo and multi-baseline vision 47(1-3), pp. 275–285.