WIDE-BASELINE 3D RECONSTRUCTION FROM DIGITAL STILLS

M. Vergauwen[1], F. Verbiest[1], V. Ferrari[2], C. Strecha[1], L. van Gool[1,2]

[1] K.U.Leuven ESAT-PSI
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

[2] ETH Zurich D-ITET-BIWI
Gloriastrasse 35, CH-8092 Zurich, Switzerland

KEY WORDS: wide baseline, 3D reconstruction, invariant neighborhoods

ABSTRACT

The last decade saw tremendous progress in the area of structure-and-motion. Techniques have been developed to compute the camera trajectory and reconstruct scenes in 3D based on nothing but a video or a set of closely spaced images. No extra information or calibration is needed to perform the reconstruction. This paper presents extensions on the traditional shape-from-video pipeline in order to deal with wide-baseline conditions, i.e. views that are much farther apart. The algorithms for both the sparse and dense correspondence search need to be upgraded for the system to deal with these wide-baseline views. These new techniques are discussed in this paper and results are shown.

1 INTRODUCTION

During the last few years user-friendly solutions for 3D modeling have become available. Techniques have been developed (Armstrong 1994, Heyden 1997, Hartley 2000, Pollefeys 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.

This paper presents ongoing work on extensions on the shape-from-video system that has been under development in our institute for the last years.

Typical shape-from-video systems require large overlap between subsequent frames. This requirement is typically fulfilled for video sequences. Existing systems can also deal with still images, provided they are sufficiently close together. Often, however, one would like to reconstruct from a small number of stills, taken from very different viewpoints. Based on local, viewpoint invariant features, wide-baseline matching is made possible, and hence the viewpoints can be further apart.

This paper is organized as follows. First an overview of the typical Shape-from-video pipeline is described. Sections 3 and 4 describe the extensions to the pipeline that allow for wide-baseline matching during both sparse and dense correspondence search. Section 5 shows some experiments and results after which some conclusions are be drawn and some directions on future work are given.

2 SHAPE-FROM-VIDEO PIPELINE

2.1 Description

In the last decade, the computer vision community has witnessed the appearance of self-calibration methods in structure-from-motion. Based on a series of images as its only input, such systems can determine the camera motion and the evolution of the camera settings as well as the 3D shape of the scene up to an unknown scale (a so-called metric reconstruction). Several approaches for such self-calibration have been developed and several systems have been proposed (Armstrong 1994, Heyden 1997, Hartley 2000, Pollefeys 1998). All approaches basically follow the same pipeline although specific implementation choices or strategies can differ. The pipeline that has been implemented in our shape-from-video system is shown in figure 1.

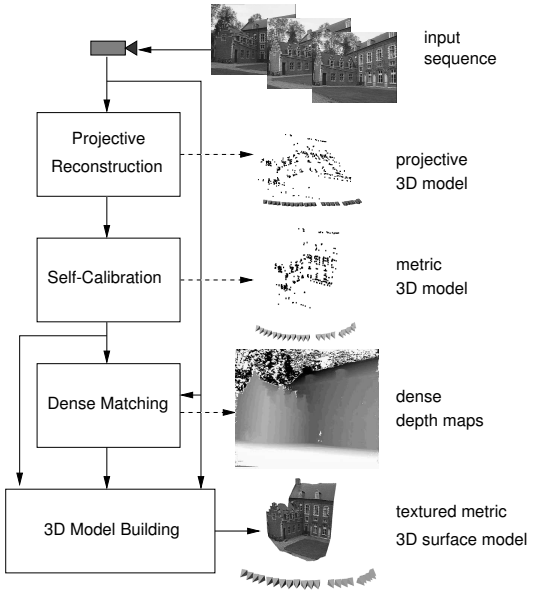

Figure 1: *Typical shape-from-video pipeline*

The pipeline starts with tracking or matching interest points throughout a sequence of views. This is often the most crucial step in the processing since the result of the rest of the pipeline depends on the quantity and quality of the matched features. The consistency of the matched features with a rigid 3D structure imposes constraints that are employed to reconstruct the camera parameters of each

view and a 3D point cloud representing the scene. The matching of these initial interest points is referred to as the *sparse correspondence search*. The projective reconstruction which is the output of the first step in the pipeline is upgraded to a metric one using the self-calibration techniques mentioned before. In our system we employ an algorithm that is primarily based on the absolute quadric approach proposed by Triggs (Triggs 1997) with some adaptations by Pollefeys (Pollefeys 1998).

Once the calibration of the cameras has been extracted strong multi-view constraints between the images are available. These are employed in the search for many more correspondences. A typical example is the epipolar constraint which limits the correspondence search to a single line. In the approach described in (Van Meerbergen 2002) we search for pixelwise matches. This stage is referred to as *dense correspondence search*. The result of the dense matching step (dense depth maps for each view) allows for dense, textured reconstructions of the recorded scene.

2.2 Wide-baseline matching

Although 3D reconstructions can in principle be made from a limited number of still images, structure-from-motion systems like the ones described above only tend to work effectively if the images have much overlap and are offered to the system as an ordered continuous camera sequence. This is underlined by the name 'shape-from-video'. Our system has been used in many areas and applications. For instance we have tested our system to make 3D records of archaeological, stratigraphic layers during excavations. A large part of the scene consists of sand and there is a general lack of points of interest. When walking around the dig, it proved necessary to take images less than 5° apart, or record a continuous video stream for the system to be able to match the images. In such an application, this is not always possible due to obstacles and it disturbs the normal progress of the excavations as the image acquisition takes too much time. It would be very advantageous if the number of images can be limited to about 10 or so. These images would still cover the whole scene but would be taken from substantially different viewpoints. Recording 'wide-baseline' images could also be done with a digital photo camera rather than a video camera, improving the resolution of the imagery by an order of magnitude.

For the shape-from-video pipeline to be able to deal with wide-baseline imagery, the crucial parts of the system must be successful: both the sparse and dense correspondence search. The existing approaches can not deal with wide-baseline conditions and new strategies have to be developed. The self-calibration procedure itself remains essentially identical. Next, we describe the adapted versions of the two correspondence steps.

3 SPARSE CORRESPONDENCE SEARCH

Consider the wide-baseline image pair of figure 2. The two images have been taken from very different viewing directions. Stereo and shape-from-video systems will most often not even get started in such cases as correspondences are very hard to find.

Figure 2: *Two images of the same scene, but taken from very different viewing directions.*

3.1 Detection of features

As already mentioned, the shape-from-video pipeline splits the correspondence problem into two stages. The first stage determines correspondences for a relatively sparse set of features, usually corners. In the traditional shape-from-video approach, corners are matched from one image to another by searching for similar corners around the same position in the other image. The typical similarity measure used is the normalized cross-correlation of the surrounding intensity patterns. Two problems arise if one wants to deal with the intended wide-baseline conditions. The corresponding point may basically lie anywhere in the other image, and will not be found close to its original position. Secondly, the use of simple cross-correlation will not suffice to deal with the change in corner patterns due to stronger changes in viewpoint and illumination. The next paragraphs describe an alternative strategy, based on affinely invariant regions that is better suited.

When looking for initial features to match, we should focus on local structures. Otherwise, occlusions and changing backgrounds will cause problems, certainly under wide baseline conditions. Here, we look at small regions, constructed around or near interest points. These regions need to be matched, so they ought cover the same part of the recorded scene in the different views. Because the images are taken from very different angles, the shape of

Figure 3: *The regions, extracted by the affine invariant region detector for the images in fig. 2. Only regions that were matched between the two images are shown. The extraction, however, was done independently for each image.*

the regions differs in the different views. The extraction method needs to take this into account. Some extraction and matching algorithms select features in an image, try to find a match by deforming and relocating the region in other images until some matching score surpasses a threshold. In order to avoid this slow and combinatoric search, we want to extract the regions for each image independently. The most important characteristic of the region extraction algorithm is that it is invariant under the image variations one wants to be robust against. This is discussed next.

On the one hand the viewpoint may strongly change. Hence, the extraction has to survive affine deformations of the regions, not just in-plane rotations and translations. In fact, affine transformations also not fully cover the observed changes. This model will only suffice for regions that are sufficiently small and planar. We assume that a reasonable number of such regions will be found, an expectation borne out in practice. On the other hand, strong changes in illumination conditions may occur between the views. The chance of this happening will actually grow with the angle over which the camera rotates. The relative contributions of light sources will change more than the frame-to-frame changes in a video. Our local feature extraction should also be immune against such photometric changes.

If we want to construct regions that are in correspondence

Figure 4: *Three wide-baseline views of the 'bookshelf' scene. The top two images show image 1 and 2 with the corresponding invariant regions. The bottom two images show the same for image 1 and 3.*

irrespective of these geometric and photometric changes and that are extracted independently in every image, every step in their construction ought to be invariant under both these transformations just described. A detailed description of these construction methods is out of the scope of this paper, and the interested reader is referred to papers specialized on the subject (Tuytelaars 1999, Tuytelaars 2000).

As mentioned before, these constructions allow the computer to extract the regions in the different views completely independently. After they have been constructed, they can be matched efficiently on the basis of features that are extracted from the color patterns that they enclose. These features again are invariant under both the geometric and the photometric transformations considered. To be a bit more precise, a feature vector of moment invariants is used. Fig. 3 shows some of the regions that have been extracted for fig. 2. We refer to the regions as 'invariant neighborhoods'. Recently, several additional construction methods have been proposed by other researchers (Baumberg 2000, Matas 2001).

3.2 Increasing the multi-view matches

The previously described wide-baseline matching approach is well suited for matching pairs of images. In the shape-from-video or shape-from-stills pipeline, however, one needs correspondences between more than two images in order to compute the camera calibration. In practice it is actually far from certain that the corresponding feature in another view is found by the wide-baseline matching algorithm. This means that the probability of extracting all correspondences for a feature in all views of an image set quickly decreases with the amount of views. Moreover, there is a chance of matching wrong features. In practice, if one is given 3 or more views, the method will mostly find sufficient matches between each pair but the sets of matches will differ substantially and a small number of common features between all views may result.

Figure 5: *The three images of the bookshelf scene, showing the features that could be matched in each of the three views. This intersection of the pairwise matching sets is quite small: only 16 features remain.*

Figure 4 shows 3 views and the matches found between the pairs $\langle 1,2\rangle$ and $\langle 1,3\rangle$. Fig. 5 shows the matches that these pairs have in common. Whereas more than 40 matches were found between the pairs of fig. 4, the number of matches between all three views has dropped sharply, to only 16. When we consider 4 or 5 views, the situation can deteriorate further, and only a few, if any, features may be put in correspondence among all the views (even though there may be sufficient overlap between all the views). Recently we have developed and tested algorithms to counteract this problem. The approach is founded on two main ideas, namely propagation of matches using neighborhood information and based on transitivity.

The first idea, propagation based on neighborhood information, makes use of the information supplied by a correct match to generate other correct matches. Consider a feature A_1 in view v_1 with its matching feature A_2 in view v_2, and a feature B_1 in v_1 which was not matched with its corresponding feature B_2 in v_1. The matching could have failed because of different reasons: maybe B_2 was not extracted during the detection phase or maybe the matching failed. The matching pair $A_1 - A_2$ gives us the affine transformation mapping A_1 to A_2. If B_1 and B_2 are spatially close and lie on the same physical surface, this affine transformation will also map B_1 to B_2 or to a point close by. Therefore, we can use this mapping as a first approximation of B_2. We then search in v_2 for for the real B_2 by maximizing the similarity between B_1 and B_2. We call this process *region propagation*. If B_1 is not close to A_1, or not on the same physical surface, a good similarity is unlikely to arise between the generated region and B_1, so this case can be detected and the propagated region rejected. The propagation approach strongly increases the probability that a feature will be matched between a pair of views, as it suffices that at least one feature in its neighborhood is correctly matched. As a result, also the probability of finding matches among all images of a set increases.

The second idea to increase the quality of multiview feature correspondences is to take the transitivity property of valid matches into account. In our 3 view example, instead of only matching between the view pairs $\langle 1,3\rangle$ and $\langle 1,2\rangle$, we can also match 2 to 3. If, for example, a feature gets matched in $\langle 1,3\rangle$ but not in $\langle 1,2\rangle$, we can look if it is matched in $\langle 2,3\rangle$. If it is, we can decide that either the matching or the matching failure was wrong. Following a majority vote, we might conclude that the match should have been found in $\langle 1,2\rangle$ and obtain a correct feature correspondence along the three views.

In summary, starting from pairwise matches, many more can be generated. Of course, the validity of propagated and implied matches is an issue, and one has to be careful not to introduce erroneous information. Research is being done at the moment to achieve this. The strategies proposed here are akin to recent work by Schaffalitzky and Zisserman (Schaffalitzky 2002). In contrast to their work, there is less emphasis on computational efficiency. In particular, adding transitivity reasoning to the propagation of matches renders our approach slower, but it also adds to the performance. The combined effect of propagation and transitivity reasoning for our example is illustrated in fig. 6. The number of matches along the three views has more than tripled.

4 DENSE CORRESPONDENCE SEARCH

As shown in figure 1, the sparse point-cloud reconstruction and self-calibration stages are only the first part of the pipeline. These stages were in need of the improved spare correspondence search to overcome wide-baseline views as explained in the previous section. The following step in the pipeline deals with the problem of finding dense matching information, i.e. finding matches between image pairs for almost every pixel in the images. Our traditional shape-from-video approach uses a dynamic programming algorithm to search for dense correspondences along epipolar lines (Van Meerbergen 2002). Given the information from the sparse reconstruction, it can deal with images that are farther apart than in a typical video sequence. It has difficulties, however, to handle more extreme cases.

Under wide baseline conditions, disparities tend to get larger, a smaller part of the scene is visible to both cameras, and intensities of corresponding pixels vary more. In order to better cope with such challenges, we propose a scheme that is based on the coupled evolution of Partial Differential Equations. This approach is described in more detail in a paper by Strecha *et al.* (Strecha 2002). The point of departure of this method is a PDE-based solution to optical flow, proposed earlier by Proesmans *et al.* (Proesmans 1994). In a recent benchmark comparison between different optical flow techniques, this method performed particularly well (McCane 2001).

An important difference with classical optical flow is that the search for correspondences is 'bi-local', in that spatio-temporal derivatives are taken at two different points in the two images. Disparities or motions are subdivided into a

Figure 6: *The features that could be matched in each of the 3 views of fig. 5 after propagation and transitivity reasoning. The number of matches has increased to 58.*

current estimate and a residue, which is reduced as the iterative process works its way towards the solution. This decomposition makes it possible to focus on the smaller residue, which is in better agreement with the linearisation that is behind optical flow. The non-linear diffusion scheme in the Proesmans *et al* approach imposes smoothness of nearby disparities at most places – an action which can be regarded as the dense counterpart of propagation – but simultaneously allows for the introduction of discontinuities in the disparity map.

The method of Strecha *et al.* (Strecha 2002) generalizes this approach to multiple views. The extraction of the different disparities is coupled through the fact that all corresponding image positions ought to be compatible with the same 3D positions. The effect of this coupling can be considered the dense counterpart of the sparse transitivity reasoning. Moreover, the traditional optical flow constraint that corresponding pixels are assumed to have the same intensities, is relaxed. The system expects the same intensities *up to scaling*, where the scaling factor should vary smoothly between neighboring pixels at most places.

5 EXPERIMENTS

Throughout this paper images of has been shown of a bookshelf with some books and a bottle. The original images are too far apart for our traditional shape-from-video pipeline to process. As shown in figures 3 and 6 enough invariant regions can be extracted and matched for the sparse reconstruction process to succeed. The PDE based dense correspondence scheme of section 4 delivers dense 3D models of the scene. Figure 7 shows this resulting reconstruction. Both textured and untextured views of the resulting 3D model are shown.

Fig. 8 shows three images of an excavation layer, acquired at the Sagalassos site in Turkey. This is one of the largest scale excavations currently ongoing in the Mediterranean, under the leadership of prof. Marc Waelkens. These images have less structure than the ones of the bookshelf and are too far apart for our shape-from-video process to get its corner matching started successfully. Again, invariant neighborhoods haven been matched and the PDE-based dense correspondence search succeeded in finding matches for most other pixels. A side view of the resulting 3D model is shown in fig. 9, with and without the texture.

6 CONCLUSIONS AND FURTHER WORK

Three-dimensional reconstruction from still images often introduces ‘wide baseline’ problems, especially when one wants to limit the amount of images to be taken. The traditional shape-from-video approach is in need of improvements to deal with these wide-baseline problems. This paper presented solutions for two crucial stages in the pipeline, namely the sparse and dense correspondence search. Ongoing work is mainly focused on issues of efficiency.

Figure 7: *Textured and untextured views of the dense reconstruction of the bookshelf scene shown in figure 3. The images are too far apart for our traditional shape-from-video pipeline to match.*

Figure 8: *Three input images of an excavation layer at an archaeological site. The images are too far apart for our shape-from-video process to match features between the views.*

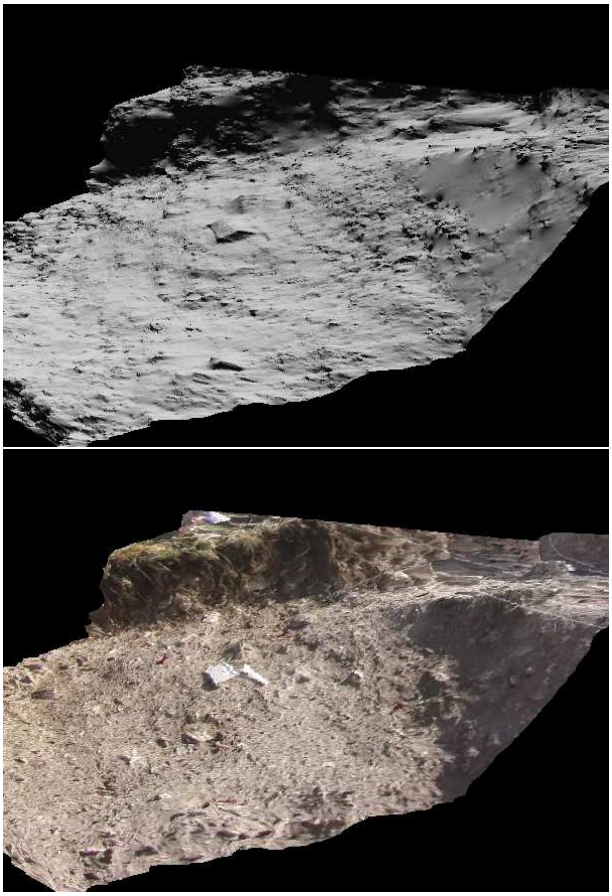

Figure 9: *The reconstruction extracted from the relatively wide baseline images of fig. 8, with and without texture.*

ACKNOWLEDGMENTS

We acknowledge support from the European IST 'ATTEST' project

REFERENCES

M. Armstrong, A. Zisserman, and P. Beardsley, *Euclidean structure from uncalibrated images*, 5th BMVC, 1994

A. Baumberg, *Reliable Feature Matching Across Widely Separated Views*, proc. CVPR, pp. 774-781, 2000

J. Feldmar and N. Ayache, *Rigid, affine and locally affine registration of free-form surfaces*, TR INRIA Epidaure, No. 2220, 1994

R. Hartley and A. Zisserman, *Multiple View Geometry*, Cambridge University Press, 2000

A. Heyden and K. Astrom, *Euclidean reconstruction from image sequences with varying and unknown focal length and principal point*, Proc. CVPR, 1997

J. Matas, O. Chum, M. Urban, and T. Pajdla, *Distinguished regions for wide-baseline stereo*, Research Report CTU–CMP–2001–33, Center for Machine Perception, Czech Techn. Un., Prague, November 2001.

B.McCane, K.Novins, D.Crannitch and B.Galvin, *On Benchmarking Optical Flow*, Computer Vision and Image Understanding, 84(1):126-143, 2001.

M. Pollefeys, R. Koch, and L. Van Gool, *Self calibration and metric reconstruction in spite of varying and unknown internal camera parameters*, Proc. ICCV, pp.90-96, 1998

M. Proesmans, L. Van Gool, E. Pauwels, and A. Oosterlinck, *Determination of optical flow and its discontinuities using non-linear diffusion*, Proc. ECCV, Stockholm, pp. 295-304, May 1994

M. Proesmans, L. Van Gool, and A. Oosterlinck, *One-shot active 3D shape acquisition*, Proceedings of the International Conference on Pattern Recongnition, pp.336-340, 25-29 Augustus 1996, Vienna, Austria

F. Schaffalitzky and A. Zisserman, *Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?"*, Proc. ECCV, pp. 414-431, Copenhagen, 2002

C. Srecha and L. Van Gool, *PDE-based multi-view depth estimation*, Proc. 1st Int. Symp. on 3D Data Processing, Visualization, and Transmission (3DPVT), pp. 416-425, Padova, june 19-21, 2002

W. Triggs, *Auto-calibration and the absolute quadric*, Proc. CVPR, pp. 609-614, 1997

T. Tuytelaars, L. Van Gool, L. D'haene, R. Koch, *Matching Affinely Invariant Regions for Visual Servoing*, International Conference on Robotics and Automation, Detroit, pp. 1601-1606, May 10-15, 1999

T. Tuytelaars and L. Van Gool, *Wide baseline stereo matching based on local, affinely invariant regions*, Proc. British Machine Vision Conference, Vol. 2, pp. 412-425, Bristol, 11-14 sept, 2000

L. Van Gool, T. Moons, E. Pauwels, and A. Oosterlinck, Semi-differential invariants, in *Applications of invariance in vision*, eds. J. Mundy and A. Zisserman, pp. 157-192, MIT Press, Boston, 1992

G. Van Meerbergen, M. Vergauwen, M. Pollefeys, and L. Van Gool, *A hierarchical symmetric stereo algorithm using dynamic programming*, International Journal of Computer Vision, 47(1-3):275-285, 2002