From Images to 3D Models

How computers can automatically build realistic 3D models from images acquired with a hand-held camera

Marc Pollefeys and Luc Van Gool

Nowadays computer graphics allows to render realistic images of the 3D world. However, before such images can be generated, graphics models of the world must be available. Traditionally, these models where obtained using 3D modelling packages. This is a very time consuming process and the achievable level of detail and realism is limited. Therefore, more and more 3D models are obtained by *sampling* the world. Different approaches are available for this. Best known are probably laser range scanners and other devices that project light onto the object and obtain 3D information by observing the reflection. Another approach consists of using images of the scene or objects one wishes to reconstruct. A simple example consists of taking a picture to use it as a texture map for a 3D model. However, not only the appearance, but also the 3D shape can be extracted from images. This will be the main subject of this paper.

The simplest way to understand how this can be done is by considering our own perception of the 3D world. We can perceive depth because we have two eyes that observe our 3D environment from two slightly different viewpoints. The images seen by our eyes are therefore slightly different and this difference is related to the distance of objects.
However, we can also perceive depth in a different way. Closing one eye and moving our head will also allow us to estimate the 3D structure of the environment. In addition, when moving through an environment, we can visually obtain a good estimate of the direction in which we are moving. In fact, both motion and structure estimation are performed simultaneously.

In the area of *computer vision*, researchers have been working for many years to provide similar capabilities for machines. The initial work was targeted towards robotics and automation, e.g. allowing a robot to navigate through an unknown environment. In recent years the focus has shifted to visualization and communication, resulting in much more interaction with the computer graphics community. One of the main focuses has been to provide algorithms that can automatically extract the necessary information from multiple images. In addition, over the last ten years important new insights have been gained in the geometry of multiple images, allowing more flexible approaches to be developed (a good reference for this is the recent book by Hartley and Zisserman (Hartley and Zisserman, 2000)).

In fact, similar work has been going on much earlier in a different context. From the early beginnings of photography, pictures have been used to obtain 3D measurements of the world. In the second half of the 19th century photographs were already used for making maps and measuring buildings. This technique -called *photogrammetry* (Slama, 1980)- is still used today for most map making and surveying. Most attention has gone into achieving high accuracy, by carefully modelling and calibrating the image formation process and dealing with errors in a statistically correct way.

Also in the *computer graphics* community there has been an important effort to construct 3D models, as prerequisites for rendering them. Part of this effort has been directed towards obtaining *image-based models*. The work by Debevec, Taylor and Malik (Debevec et al., 1996) is a good

example of how convincing 3D models can be obtained from a set of images using a limited amount of user interaction. In fact, this work combines methods developed in graphics, vision and photogrammetry.

From Images to 3D Models

Our work also combines insights, methods and algorithms developed in these different fields. The problem we have focussed on is to allow a computer to automatically generate a realistic 3D model when provided with a sequence of images of an object or scene. This problem can be divided in a number of smaller problems.

But before going into more detail, let us consider the simplest case. Given two images of a 3D point, how do we reconstruct that point? Let us assume that we know the relative motion of the camera between the two images and that a model is available that relates pixels in the image to rays passing through the projection centre of the camera. A 3D point can now be reconstructed from its two projections by computing the intersection of the two space rays that correspond to it. This process is called triangulation. Note that we need two things for this: (1) the relative motion and calibration of the camera, and (2) the corresponding image points. Now we will briefly go through the different steps that are needed to extract this information from real images, and see how gradually a complete 3D model can be recovered from a sequence of images.

First, the relative motion between consecutive images needs to be recovered. This process goes hand in hand with finding corresponding image features between these images (i.e. image points that originate from the same 3D feature). The next step consists of recovering the motion and calibration of the camera and the 3D structure of the features. This process is done in two phases. At first the reconstruction contains a projective skew (i.e. parallel lines are not parallel, angles are not correct, distances are too long or too short, etc.). This is due to the absence of a priori calibration. Using a self-calibration algorithm (Pollefeys et al., 1999a) this distortion can be removed, yielding a reconstruction equivalent to the original up to a global scale factor. This *uncalibrated* approach to 3D reconstruction allows much more flexibility in the acquisition process since the focal length and other intrinsic camera parameters do not have to be measured -- calibrated-- beforehand and are allowed to change during the acquisition.

The reconstruction obtained as described in the previous paragraph only contains a sparse set of 3D points (only a limited number of features are considered at first). Although interpolation might be a solution, this typically yields models with poor visual quality. Therefore, the next step consists in an attempt to match all image pixels of an image with pixels in neighbouring images, so that these points too can be reconstructed. This task is greatly facilitated by the knowledge of all the camera parameters that we have obtained in the previous stage. Since a pixel in the image corresponds to a ray in space and the projection of this ray in other images can be predicted from the recovered pose and calibration, the search of a corresponding pixel in other images can be restricted to a single line. Additional constraints such as the assumption of a piecewise continuous 3D surface are also employed to further constrain the search. It is possible to warp the images so that the search range coincides with the horizontal scanlines. An algorithm that can achieve this for arbitrary camera motion is described in (Pollefeys et al., 1999b). This allows us to use an efficient stereo algorithm that computes an optimal match for the whole scanline at once (Van Meerbergen et al., 2002). Thus, we can obtain a depth estimate (i.e. the distance from the camera to the object surface) for almost every pixel of an image. By fusing the results of all the images together a complete dense 3D surface model is obtained. The images used for the reconstruction can also be used for texture mapping so that a final photo-realistic result is achieved. The different steps of the process are

illustrated in Figure 1. In the following paragraphs some of the critical steps are described in some more detail.

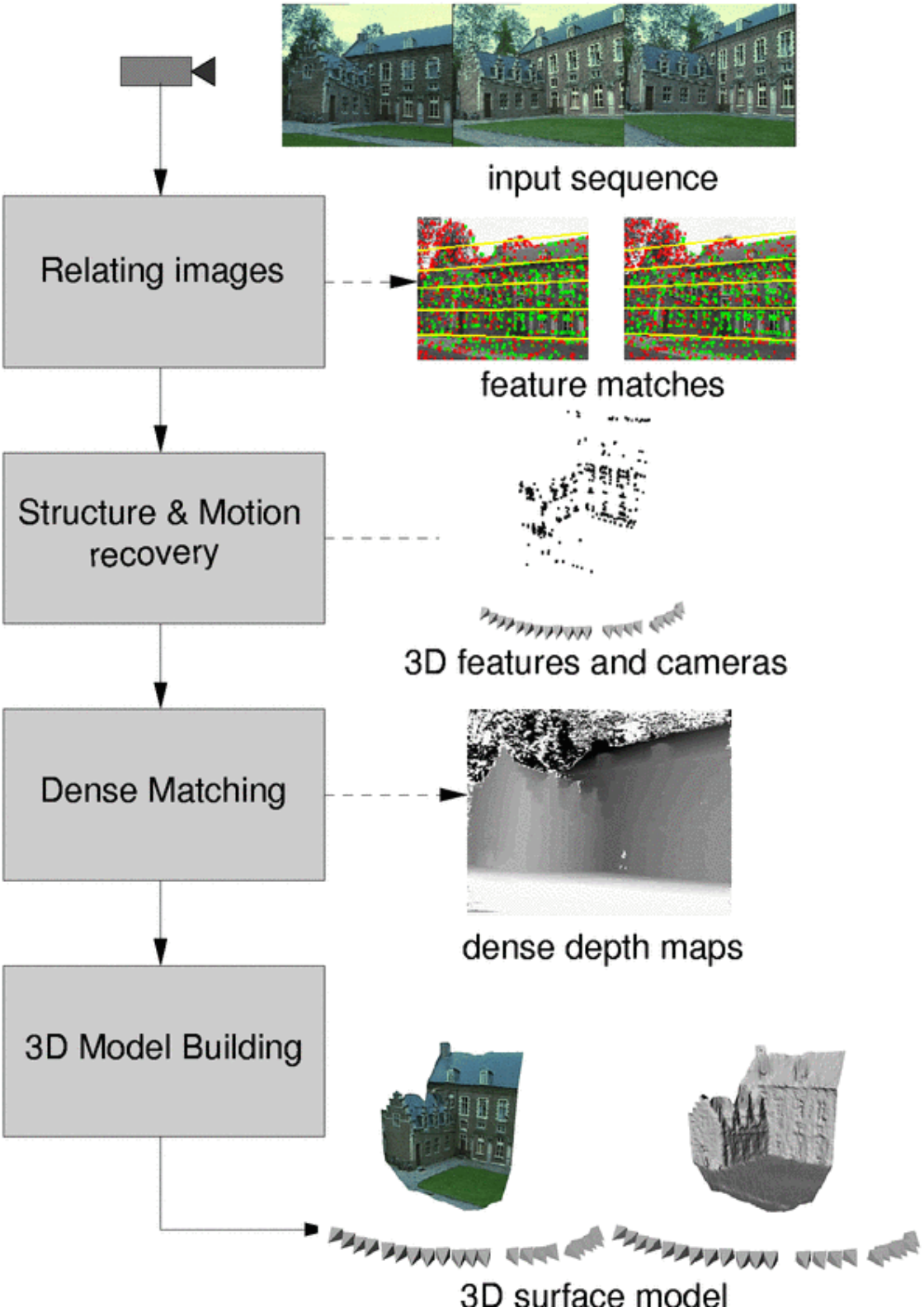

Figure 1 Overview of our image-based 3D recording approach.

Robustly relating real images

The first important difficulty consists of automatically getting initial matches from real images. For the computer an image is just a large collection of pixel intensity values. To find corresponding points in different images one can compare intensity values over a small region around a point. However, not all points are suited for this type of comparison. When a point cannot be differentiated from its own neighbours, one can also not expect to find a unique match with a point in another image. Therefore points in homogeneous regions or located on straight edges are not suited for matching at this stage. The method we use extracts a set of points using a feature detector (Harris and Stephens, 1988) that maximizes dissimilarity with neighbouring pixels. Typically, we aim at extracting 1000 feature points, well distributed over the image, per image.

Potential matches are obtained by comparing feature points with similar coordinates. This constraint is later relaxed in a second stage. Of course, many of these initial matches can be wrong and we need a robust procedure to deal with them. This can be done because a set of correct matches must satisfy some constraints. Since a 3D point only has three degrees of freedom and a pair of image points has four, there should be a constraint for each pair of matching points (up to some degrees of freedom related to the relative camera pose). This internal structure of a set of matches, called the *epipolar geometry*, can be computed from 7 or more point matches. Using a RANSAC procedure (Fischler and Bolles, 1981), the largest set of consistent matches is computed. The algorithm generates a hypothesis from a set of 7 arbitrary matching points and determines which ratio of the other matches supports the hypothesis. This step is repeated until the probability that a correct hypothesis was generated exceeds a certain threshold. This allows us to robustly deal with more than half the initial matches being incorrect. Once a correct solution is found it can be refined and used to guide the search for additional matches. At the same time the recovered epipolar geometry allows us to generate an initial projective reconstruction (Faugeras, 1992, Hartley et al. 1992). Note that this approach assumes that the observed scene is rigid.

Avoiding the need for calibration

The traditional work on 3D reconstruction from images requires the cameras to be calibrated. It was one of the key realisations of our approach to relax this constraint. Besides the important gain in flexibility, one of the benefits is that 3D reconstruction can also be obtained from archive footage or from amateur video images recorded with a varying zoom. Avoiding calibration is based on the stratification of geometry. At first only the projective structure of the scene and camera geometry is recovered. This can be done independently of calibration. The perspective camera model being used at this level is too general and a number of additional constraints are available. These allow us to recover the euclidean structure (up to a scale factor) (Pollefeys et al. 1999a).

From 3D Models to Images

Once a 3D model has been obtained one of the main applications besides measuring consists of rendering novel views of the recorded scene. However, if this is the goal, alternatives to explicit 3D modelling are available. Recently purely image-based techniques such as *lightfield rendering* (Levoy and Hanrahan, 1996) have been proposed. In this case the recorded images are seen as a large collection of 3D lightrays from which the most appropriate ones are selected to fill in the pixels of a new image. How this can efficiently be achieved -using the data we have computed- is described in (Koch et al., 2001). The advantage of this type of approach is that view-dependent effects are easily reproduced. Other effects, however, are easier to cater for given explicit 3D models.

Applications

The 3D modelling approach explained above has two types of application: (1) the 3D model can be used for measurements; (2) the 3D model can be used for visualization. Most of the real applications make use of both aspects. Because of the cost and flexibility this type of approach has many interesting applications in different fields. Here we will just highlight two applications.

Cultural heritage

A first interesting application is to record archaeological monuments and sites so that they can be completed with virtual reconstructions that are based on archaeological hypothesis. However, archaeology has many more needs. Recording sufficient measurements is very important in the field of archaeology because evidence is continuously destroyed during excavation. In this context a cheap and flexible technique for 3D recording is very advantageous. Excavated blocks can be modelled so that construction hypotheses can be verified virtually. It even becomes possible to record the whole excavation process in 3D. A few examples are shown in Figure 2.

Figure 2 3D recording of cultural heritage: one image from a video sequence and resulting 3D model (top-left), top view of 3D record of excavations (top-right), Dionysus statue virtually placed back from museum to its original on-site location (bottom-left) and one frame of a video where an architect presents a virtual reconstruction of an ancient monument

Planetary exploration

A prerequisite for successful robot navigation is a good estimate of the terrain geometry. Often the only available option consists of obtaining this information with cameras. However, launch and landing can perturb the camera settings, so that it must be possible to perform a re-calibration at the remote landing site. In this context the possibility to calibrate from images is very important. In Figure 3 some results obtained in the context of an ESA (European Space Agency) project are shown. Note that besides robot navigation the 3D reconstruction of the terrain is also very useful to

provide a simple virtual reality interface for simulation and for visualizing the telemetry data received from the planet.

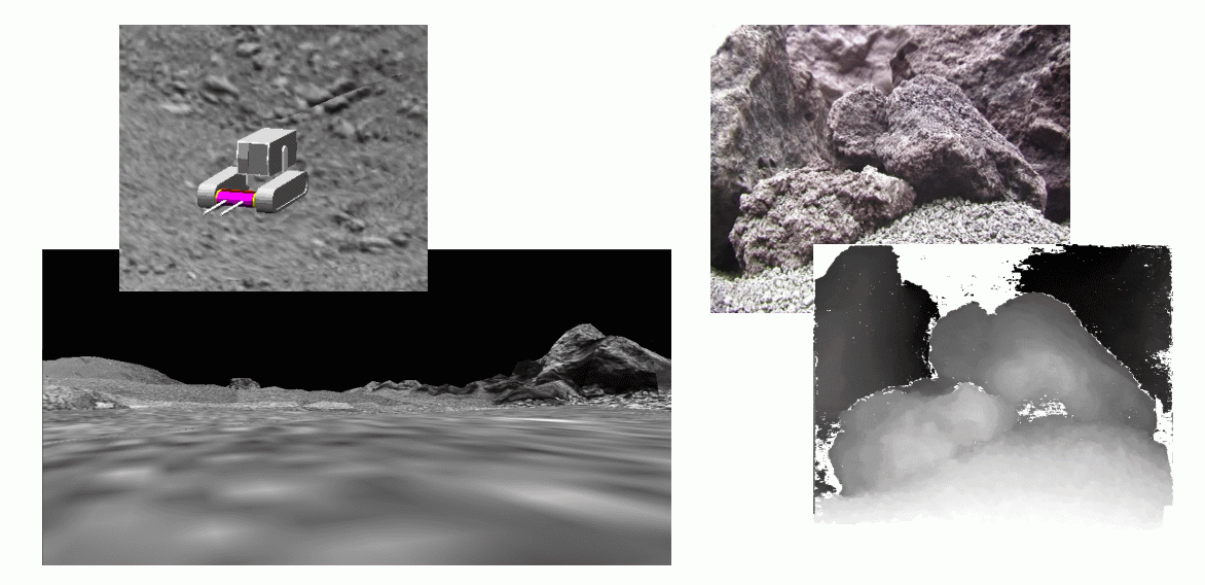

Figure 3 Planetary rover navigation: virtual reality simulation (left), close-range reconstruction of rock docking rover (right).

Virtual Reality, Augmented Reality and Mixed Reality

For some applications the recovered camera parameters are also very important. A good example is the insertion of virtual objects into a real video sequence. In this case it is important that the inserted object undergoes the same relative motion with respect to the camera. Today, the techniques described in this paper are used to generate special effects (see bottom left of Figure 2). The processing is done off-line. In the future this type of computation will also become possible in real-time allowing flexible vision-based augmented reality and other aspects such as shadows, occlusion and lighting will automatically be taken into account.

Conclusion

Obtaining 3D models from images is possible. It turns out that this task can even be performed automatically by a computer. Applications for this type of technology can be found in many different areas, just to name a few: archaeology, architecture, forensics, geology, planetary exploration, e-commerce, movie special effects, virtual and augmented reality, etc. Important advantages of this type of approach compared to others are the flexibility and a lower cost.

Authors

Marc Pollefeys (Marc.Pollefeys@esat.kulueven.ac.be) is a post-doctoral fellow of the Fund for Scientific Research - Flanders (Belgium) attached to the Centre for Processing of Speech and Images of the Katholieke Universiteit Leuven.

Luc Van Gool (Luc.VanGool@esat.kuleuven.ac.be) is a professor of computer vision in the Center for Processing of Speech and Images of the Katholieke Universiteit Leuven and in the Communication Technology Lab of ETH Zürich.

Acknowledgement

Maarten Vergauwen, Kurt Cornelis, Frank Verbiest, Jan Tops and Reinhard Koch also contributed to the work presented in this paper. The financial support of the FWO project G.0223.01 and the IST projects VIBES and InViews are gratefully acknowlegded.

References

P. Debevec, C. Taylor and J. Malik, 1996. ``Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-Based Approach", *Proc. SIGGRAPH'96*, pp. 11-20.

O. Faugeras, 1992. ``What can be seen in three dimensions with an uncalibrated stereo rig", *Computer Vision - ECCV'92*, Lecture Notes in Computer Science, Vol. 588, Springer-Verlag, pp. 563-578.

M. Fischler and R. Bolles, 1981. ``RANdom SAmpling Consensus: a paradigm for model fitting with application to image analysis and automated cartography", *Communications of the ACM*, 24:381-95.

C. Harris and M. Stephens, 1988. ``A combined corner and edge detector", *Fourth Alvey Vision Conference*, pp.147-151.

R. Hartley, R. Gupta, and T. Chang, 1992. ``Stereo from uncalibrated cameras", *Proc. Conference Computer Vision and Pattern Recognition*, pp. 761-764.

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

R. Koch, B. Heigl, and M. Pollefeys, 2001. “Image-Based Rendering from Uncalibrated Lightfields with Scalable Geometry”, In R. Klette, T. Huang, G. Gimel'farb (Eds.), *Multi-Image Analysis*, Lecture Notes in Computer Science, Vol. 2032, pp.51-66, Springer-Verlag.

M. Levoy and P. Hanrahan, 1996. ``Lightfield Rendering", *Proc. SIGGRAPH '96*, pp 31-42, ACM Press, New York.

M. Pollefeys, R. Koch and L. Van Gool, 1999. ``Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters", *International Journal of Computer Vision*, 32(1), 7-25.

M. Pollefeys, R. Koch and L. Van Gool, 1999. "A simple and efficient rectification method for general motion", *Proc.ICCV'99 (international Conference on Computer Vision)*, pp.496-501, Corfu (Greece).

C. Slama, 1980. *Manual of Photogrammetry*, American Society of Photogrammetry, Falls Church, VA, USA, 4th edition.

G. Van Meerbergen, M. Vergauwen, M. Pollefeys, L. Van Gool, 2002. ``A Hierarchical Symmetric Stereo Algorithm Using Dynamic Programming", *International Journal of Computer Vision*, Vol. 47, No. 1-3.