



[BY MARC POLLEFEYS AND
LUC VAN GOOL]

FROM IMAGES TO 3D MODELS

How computers can automatically build realistic 3D models
from 2D images acquired with a handheld camera.

Computer graphics allows us to render realistic images of the 3D world. But before these images can be generated, graphics models of the world have to be available. Traditionally, these models are obtained through 3D modeling packages—a time-consuming process that produces only limited detail and realism. 3D models are therefore increasingly obtained by *sampling* the world. A number of approaches are available. 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 consists of using images of a scene or objects one wishes to reconstruct; for example, a picture can be used as a texture map for a 3D model. Along with appearance, an estimate of the 3D shape can also be extracted from images.

The simplest way to understand how a 3D shape can be obtained is by considering our own innate perception of the 3D world. We 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 from our eyes.

However, we also perceive depth another way. Closing one eye and moving our head allows us to estimate the 3D structure of the environment. In

addition, when moving through an environment, we visually obtain a good estimate of the direction in which we are moving. In fact, motion and structure estimation are performed simultaneously.

Computer-vision researchers have worked for years trying to give machines similar capabilities. Their initial work targeted robotics and automation to, say, allow a robot to navigate through an unknown environment. More recently, their focus has shifted to visualization and communication, resulting in much more interaction with the computer graphics community. A main focus now is to provide algorithms that automatically extract the necessary information from multiple images. In addition, important new insights have been gained in the geometry of multiple images, allowing more flexible approaches to 3D modeling from images [6].

Similar work has also been part of another context. From photography's earliest beginnings in the 1830s, pictures have been used to obtain 3D measurements of the world. In the second half of the 19th century, photographs were already being used for map making and measuring buildings. This technique—called “photogrammetry” [11]—is still used today in most map making and surveying. The main photogrammetry effort has gone toward achieving accuracy by carefully modeling and calibrating the image-formation process, dealing with errors in a statistically correct way.

The computer graphics community has also shown a lot of interest in constructing 3D models—as prerequisites for rendering them. Part of this effort has been directed toward obtaining image-based models.

Modeling Medusa in 3D (Marc Pollefeys and
Luc Van Gool).

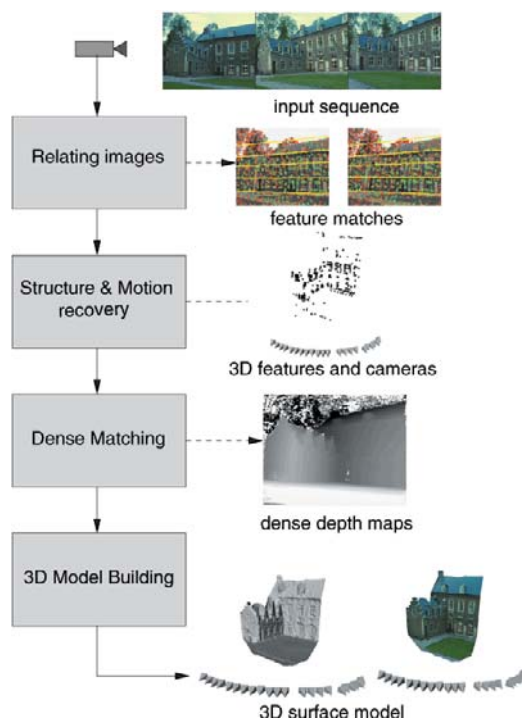
A good example of how convincing 3D models can be obtained from a set of images using a limited amount of user interaction can be found in [1], combining methods from graphics, vision, and photogrammetry.

3D Models, Automatically

Our work combines insights, methods, and algorithms developed in all these fields. We focus on allowing a computer to automatically generate a realistic 3D model when provided a sequence of images of an object or scene. The problem of 3D modeling from images can be subdivided into a number of smaller problems. But before going into detail, consider the simplest case. Given two images of a 3D point, how might we reconstruct that point? Assume we know the relative motion of the camera between the two images and a camera model is available that relates pixels in the image to rays passing through the projection center of the camera. A 3D point can now be reconstructed from its two projections by computing the intersection of the two space rays corresponding to it—a process called “triangulation.” Note that we need two things to reconstruct a 3D point: the relative motion and calibration of the camera; and the corresponding image points. Here, we briefly explore the steps needed to extract this information from real images, showing how a complete 3D model can be recovered from a sequence of 2D images (see Figure 1).

First, the relative motion between consecutive images must be computed. This process goes hand in hand with finding corresponding image features between the images, or image points originating 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: The reconstruction first contains a projective skew; parallel lines are not parallel, angles are not correct, and distances are too long or too short. This distortion results from the absence of a priori cal-

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



ibration. A self-calibration algorithm [9] removes this distortion, yielding a reconstruction equivalent of 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, or calibrated, beforehand and are allowed to change during the acquisition.

The resulting reconstruction contains only a sparse set of 3D points, that is, only a limited number of features are considered at first. Although constructing a 3D triangle from these points might be a solu-

tion, this approach would typically yield models with poor visual quality. Therefore, the next step consists of an attempt to match all image pixels of an image with pixels in neighboring images, so these points, too, are reconstructed. This task is greatly facilitated by the knowledge of all the camera parameters obtained in the previous stage. Since a pixel in the image corresponds to a ray in space, and the projection of the same ray in other images can be predicted from the recovered pose and calibration, the search for 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 narrow the search. Algorithms are available for warping the images so the search range coincides with the horizontal scanlines. The algorithm described in [10] performs this task for arbitrary camera motions. An efficient stereo algorithm can be used to compute an optimal match for a whole scanline at once [12]. Thus, we can obtain a depth estimate, or the distance from the camera to the object surface, for almost every pixel in an image. A complete dense 3D surface model is obtained by fusing together the results of all the images. The images used for the reconstruction can also be used for texture mapping, thus achieving a final photorealistic result.

Figure 2. 3D recording of cultural heritage. An image from a video sequence and resulting 3D model (top left); top view of 3D record of excavations (top right); Dionysus statue from museum placed (virtually) in its original location (bottom left); and a frame of a video showing an architect describing a virtual reconstruction of an ancient monument.

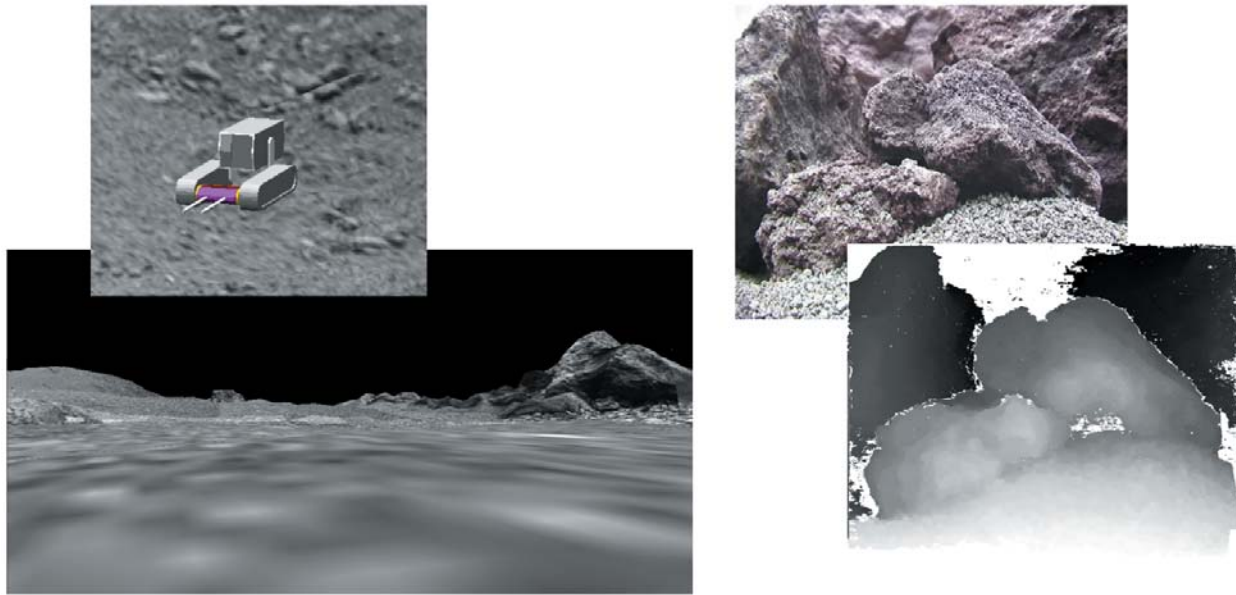


WE can obtain a depth estimate, or the distance from the camera to the object surface, for almost every pixel in an image.

Robustly relating real images. The most challenging issue involves 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, an algorithm might compare intensity values over a small region around a point. However, not all points are suitable for such comparison. When a point cannot be differentiated from its neighbors, it is also not possible to find a unique match with a point in another image. Therefore, points in homogeneous regions or located on straight edges are not suitable for matching at this stage. The typical approach to selecting interesting points uses a feature detector [4] that looks for maximal dissimilarity with neighboring pixels; we typically aim to extract 1,000 feature points per image, well distributed over the entire image.

Potential matches are obtained by comparing feature points with similar coordinates. This constraint is later relaxed in a second stage. Many of these initial matches can be wrong, however, so we need a robust procedure for dealing with them. Applying such a procedure is possible because a set of correct matches has to satisfy some constraints. Since a 3D point has only 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 seven or more point matches. A random sample consensus (RANSAC) procedure [3], typically used for robustly dealing with large numbers of outliers, helps compute the largest set of consistent matches. The algorithm generates a

Figure 3. Planetary rover navigation. Virtual reality simulation (left) and close-range reconstruction of rocky terrain to support rover operations (right).



hypothesis from a set of seven arbitrary matching points and determines which portion of the other matches supports the hypothesis. This step is repeated until the probability that a correct hypothesis was generated exceeds a certain threshold. The RANSAC algorithm allows us to robustly deal with over 50% of incorrect matches. A correct solution can be refined and used to guide the search for additional matches. At the same time, the recovered epipolar geometry makes it possible to generate an initial projective reconstruction [2, 5], a key development in the computer vision literature. Note that this approach assumes the observed scene is rigid.

Avoiding the need for calibration. Traditional work on 3D reconstruction from images in photogrammetry and computer vision requires the cameras be calibrated so the intrinsic camera parameters are known. Over the past decade, research in computer vision has sought to relax this constraint. A key insight spurring our effort to produce 3D models from images is to provide a practical approach for dealing with an uncalibrated zooming camera. Besides the important gain in flexibility, another benefit is that 3D reconstruction can be obtained from archive footage or even from amateur video images.

Avoiding calibration is based on the stratification of geometry. At first only the projective structure of the scene and camera geometry are recovered through a sequential reconstruction approach. This computa-

tion can be independent of calibration. However, the perspective camera model can be too general, and a number of additional constraints are available (such as that pixels have to be rectangular). These constraints allow us to recover the Euclidean structure of the scene (up to a scale factor) [9].

From 3D Models to Images

Once a 3D model is obtained, one of the main applications, besides measuring, involves rendering novel views of the recorded scene. However, alternatives to explicit 3D modeling are available for achieving this particular goal. Over the past six years, purely image-based techniques, such as light-field rendering [8], have been proposed. In the case of image-based techniques, the recorded images are treated 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 selection might be achieved efficiently—using the data we have already computed—is described in [7]. The advantage of this approach is that view-dependent effects are easily reproduced. Other effects, however, require explicit 3D models.

Measurement and visualization. This 3D modeling approach has two types of application—measurement and visualization—both used in most applications. Low cost and flexibility mean many interesting applications are likely possible in a variety of fields, exemplified by the following:

Cultural heritage. Archaeological monuments and sites can be recorded so they can be used to render virtual reconstructions based on archaeological hypotheses. However, archaeology has many more 3D recording needs. For example, recording sufficient measurements is important because evidence is continuously destroyed during excavation. In this context, a cheap and flexible technique for 3D recording is highly advantageous. Excavated blocks can be modeled so construction hypotheses can be verified virtually; it is even possible to record the entire excavation process in 3D (see Figure 2).

Planetary exploration. A prerequisite for successful robot navigation is a good estimate of terrain geometry. The only available option is often limited to obtaining this information through cameras. However, launch and landing can perturb camera settings, so it must be possible to perform a recalibration at the remote landing site. In this context, calibrating from images is clearly important. Figure 3 includes some results obtained in the context of a European Space Agency project called Payload Support for Planetary Exploration. Note that besides robot navigation, the 3D reconstruction of the terrain is useful for providing a simple virtual reality interface for simulations and for visualizing the telemetry data received from remote planets.

Virtual, augmented, and mixed reality. For some applications, the recovered camera parameters are important as well. A good example is the insertion of virtual objects into real video sequences; an inserted object should, however, undergo the same relative motion with respect to the camera. The techniques described here are already used to generate special effects (such as the one in Figure 2, bottom left); processing is done off-line. In the future, such computation will be possible in real time, allowing flexible vision-based augmented reality, as well as other aspects of augmented reality by automatically accounting for shadows, occlusions, and lighting.

Conclusion

Obtaining 3D models from images is on its way to being a routine form of model building. It turns out it can even be performed automatically by a computer using a combination of algorithms developed in computer vision, photogrammetry, and computer graphics. Applications are found in a variety of scientific, engineering, industrial, even cultural disciplines, including archaeology, architecture, e-commerce, forensics, geology, planetary exploration, movie special effects, and virtual and augmented reality. Notable advantages of this approach compared to other forms of 3D model building

include greater flexibility and significantly lower cost. ■

REFERENCES

1. Debevec, P., Taylor, C., and Malik, J. Modeling and rendering architecture from photographs: A hybrid geometry and image-based approach. In *Proceedings of SIGGRAPH'96* (New Orleans, LA, Aug. 4–9). ACM Press, New York, 1996, 11–20.
2. Faugeras, O. What can be seen in three dimensions with an uncalibrated stereo rig. In *Proceedings of the 2nd European Conference on Computer Vision (ECCV'92), Lecture Notes in Computer Science, Vol. 588* (Santa Margherita Ligure, Italy, May 18–23). Springer-Verlag, Berlin, 1992, 563–578.
3. Fischler, M. and Bolles, R. Random sampling consensus: A paradigm for model fitting with application to image analysis and automated cartography. *Commun. ACM* 24, 6 (June 1981), 381–395.
4. Harris, C. and Stephens, M. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference* (Manchester, England, Aug. 31–Sept. 2). The University of Sheffield Printing Unit, 1988, 147–151.
5. Hartley, R., Gupta, R., and Chang, T. Stereo from uncalibrated cameras. In *Proceedings of Computer Vision and Pattern Recognition* (Urbana-Champaign, IL, June 15–18). IEEE Computer Society Press, Los Alamitos, CA, 1992, 761–764.
6. Hartley, R. and Zisserman, A. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, U.K., 2000.
7. Koch, R., Heigl, B., and Pollefeys, M. Image-based rendering from uncalibrated lightfields with scalable geometry. In *Multi-image Analysis, Lecture Notes in Computer Science, Vol. 2032*, R. Klette, T. Huang, and G. Gimelfarb, Eds. Springer-Verlag, Berlin, 2001, 51–66.
8. Levoy, M. and Hanrahan, P. 1996. Lightfield rendering. In *Proceedings of SIGGRAPH'96* (New Orleans, LA, Aug. 4–9). ACM Press, New York, 1996, 31–42.
9. Pollefeys, M., Koch, R., and Van Gool, L. Self-calibration and metric reconstruction in spite of varying and unknown internal camera parameters. *Int. J. Comput. Vision* 32, 1 (Aug. 1999), 7–25.
10. Pollefeys, M., Koch, R., and Van Gool, L. A simple and efficient rectification method for general motion. In *Proceedings of the International Conference on Computer Vision* (Corfu, Greece, Sept. 22–25). IEEE Computer Society Press, Los Alamitos, CA, 1999, 496–501.
11. Slama, C. *Manual of Photogrammetry, 4th Ed.* American Society of Photogrammetry, Falls Church, VA, 1980.
12. Van Meerbergen, G., Vergauwen, M., Pollefeys, M., and Van Gool, L. A hierarchical symmetric stereo algorithm using dynamic programming. *Int. J. Comput. Vision* 47, 1–3 (Apr.–June 2002), 275–285.

MARC POLLEFEYS (marc@cs.unc.edu) is an assistant professor in the Department of Computer Science at the University of North Carolina, Chapel Hill.

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 Laboratory of ETH Zurich.

We gratefully acknowledge the financial support of the Fund for Scientific Research, Flanders (Belgium), project G.0223.01 and the European Commission's Information Society Technologies Programme projects: Video Browsing, Exploration, and Structuring (VIBES); 3D Measurements & Virtual Reconstruction of Ancient Lost Worlds of Europe (3D Murale); Interactive and Immersive Video from Multiple View Images (InViews); and Advanced Three-Dimensional Television System Technologies (ATTEST).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.