

Towards Automatic 3D Reconstruction of Urban Scenes from Low-Altitude Aerial Images

Adriano B. Huguet Rodrigo L. Carceroni Arnaldo de A. Araújo

*Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627, Pampulha – Belo Horizonte, MG, CEP 31270-010, Brazil
{huguet,carceron,arnaldo}@dcc.ufmg.br*

Abstract

We propose a methodology for reconstructing large-scale architectural scenes from low-altitude aerial images, in an efficient, accurate and fully automatic way. Towards this goal, we have developed an area-based segmentation technique, called Colored Watershed, that is particularly suited to segment objects with homogeneous photometric properties, which are typical of such scenes. This technique is now being combined with a dense-stereo method biased towards depth discontinuities near the edges of the segmented objects. In a final step, parametric models of these segmented objects will be instantiated and directly adjusted to the multiple images available to generate a mixed surface and elevation map for each scene.

1. Introduction

In this paper we consider the problem of recovering the 3D structure of a metropolitan area from a set of images of this area acquired from viewpoints where the effects of perspective foreshortening and occlusion are substantial. The widespread use of Geographical Information Systems (GISs) and the recent advances in the area of Tele-Presence [16, 15] have stressed the practical importance of reconstructing architectural structures from images. Consequently, this general problem has been extensively studied by the Image Processing, Computer Vision and Graphics communities. In reality, widespread scientific interest in this theme dates back to the 19th century, when it was one of the central questions studied in Photogrammetry [11].

Unfortunately, this diversity of efforts has led to a myriad of alternative solutions to the problem, rather than to a universal, well-established methodology. The more GIS-oriented literature (*e.g.* [10]) tends to work with (approximately) orthographic images, captured from high altitude. While this eliminates or at least reduces to a minimum the difficulties caused by perspective foreshortening and occlu-

sion, it requires the use of expensive, high-resolution equipment in the image-capture procedure, or it limits strongly the resolution of the data obtained.

The Computer Vision community, on the other hand, has dedicated a great deal of effort towards the development of general-purpose 3D reconstruction algorithms. In particular, a major breakthrough in this area was the recent development of Space Carving and other scene-space reconstruction techniques [14, 4, 20, 9] that are guaranteed to obtain globally consistent solutions even in the presence of substantial occlusion. Once occlusion has been handled at this global level, the reconstructions can be fine-tuned through the use of dense stereo techniques that tolerate depth discontinuities [19, 21, 6, 17]. A caveat, if one feels tempted to use this kind of approach, is the high price of generality: not only are these techniques quite expensive computationally, but they fail to make use of any specific prior knowledge about aerial images of urban areas in order to obtain more accurate reconstructions.

The benefit of using prior knowledge in the reconstruction process has been demonstrated quite dramatically by the high-accuracy architectural modeling techniques developed within the Graphics community, such as Façade [8]. However, in these approaches the degree of manual intervention in the reconstruction process is high. While this works fine for the modeling of individual buildings, it can render the modeling of entire cities very cumbersome.

The ultimate goal of the work described here is to combine contributions of these various communities into a algorithm designed specifically to reconstruct large-scale architectural scenes from low-altitude aerial images, in an efficient, accurate and fully automatic way. Towards this goal, our strategy is to use pieces of prior knowledge that are valid for a large set of architectural scenes — in contrast to *a priori* geometric models of each particular building, as in Façade — to constrain and guide the 3D reconstruction.

More specifically, we exploit the fact that architectural scenes typically contain a large number of objects with homogeneous photometric properties (roofs, walls, pavement)



Figure 1. Rectified stereo pair created from shots taken 4 seconds apart in a low-altitude flight (1,500 feet relative to the ground) performed over the city of Belo Horizonte, MG, Brazil.

in order to develop an especially tailored technique, named *Colored Watershed*, to segment these objects in individual images. Given this initial segmentation, we can then modify existing dense-stereo algorithms in order to make them strongly biased towards depth discontinuities near the edges of the segmented objects. Finally, given a dense elevation map for the entire scene obtained in the step above, we exploit the fact that architectural scenes typically contain a large number of polyhedral objects (buildings, streets), in order to fit data-driven, parametric models of these objects directly to the multiple images available.

This last step is motivated by the recent work of Yezzi and Soatto [24], which has demonstrated that segmentation and 3D reconstruction are dual problems that should be performed simultaneously, for optimal accuracy. Contrary to their technique, however, the algorithm that we propose here does not assume that a reasonable initial estimate for the shape of each object in the scene is available *a priori*: it works from the raw images, building these shape estimates through the use of special-purpose segmentation and stereo techniques and then it refines them in a final image-driven, optimization step.

Thus, the key contributions of our methodology can be summarized as follows: (1) it allows constraints from multiple viewpoints to be used simultaneously, in order to increase the accuracy of the final segmentation of polyhedral structures; (2) it leads to more accurate reconstructions than general-purpose dense stereo, by exploiting prior knowledge about general properties of architectural scenes; (3) it allows high-precision parametric model fitting to be performed in a fully automatic way; (4) it ultimately solves the segmentation and reconstruction problems jointly, for optimality; (5) it yields scene descriptions more compact than

depth maps alone, yet are as general as them, since residual depth maps are used to represent non-polyhedral structures.

Before we start discussing the specifics of our methodology, we want to stress the fact that this paper describes work that is still in progress. In Section 2, we state the general algorithm of our approach, and in Sections 3 to 5, we describe the major components of this algorithm. While Section 3 describes work that has been completed and is backed by empirical evidence, Section 4 and, especially, Section 5 are more theoretical and still need to undergo rigorous experimental validation.

2. The Reconstruction Algorithm

It should be clear from Section 1 that a major goal of our methodology is to make the image-capture procedure as simple and affordable as possible. While, for instance, Digital Elevation Maps (DEMs) and Digital Orthophoto Quadrangles (DOQs) of most metropolitan areas in the U.S. — with resolutions on the order of ten meters — can be purchased from the U.S. Geological Survey, access to elevation data of poor countries — even at this modest level of resolution — is not so universal. With this scenario in mind, we developed an algorithm that, given as input a sequence of pictures acquired with an off-the-shelf megapixel digital camera attached to a small airplane flying at low altitude over an urban area, is capable of producing a mixed surface and elevation map for this area covered in the flight.

In Figure 1, we show two consecutive frames of a data set acquired as described above. Two important aspects of this data set are the existence of significant and unknown camera motion from frame to frame, and the large overlap between the areas covered by consecutive images. The ab-

sence of *a priori* knowledge about camera motion means that this information must be recovered directly from the data, before dense–stereo techniques can be used to generate DEMs. Fortunately, the large frame–to–frame overlap means that we can perform this task in a fully automatic fashion, through the use of the various epipolar–geometry recovery techniques available in the Computer Vision literature (e.g. [12]), modified to operate with robust metrics. This observation, together with the general ideas discussed in Section 1 leads to the following reconstruction algorithm:

1. For each image:
 - (a) Perform area–based segmentation of salient, uniformly colored / textured image regions;
2. For each pair of images:
 - (a) Recover the pair’s epipolar geometry and use it to rectify the images;
 - (b) Apply a dense stereo algorithm with a bias towards depth discontinuities at the boundaries of the regions segmented in Step 1a;
 - (c) Find stereo correspondences between segmented regions, using the depth map output by Step 2b;
 - (d) Approximate each region matched in Step 2c by a polygon; fit this polygon directly to the input images; if the residual error in the polygon’s image projections is smaller than a pre–defined threshold, use it to replace the region’s depth map;
3. Combine all partial reconstructions obtained in the iterations of Step 2 into a unique 3D model.

While all steps in the algorithm above are non–trivial, solutions to Steps 2a [18], 2c [23] and 3 [7] are well known. In the following sections, we will focus on how to accomplish the less well–established steps 1a, 2b and 2d.

3. Segmentation

The goal of the initial segmentation step in our algorithm is to isolate simple architectural structures such as roofs, walls and pavement, allowing the posterior application of specific, more accurate reconstruction techniques on the areas covered by these structures. In order to accomplish this task, we use a specially tailored, area–based segmentation algorithm known as *Colored Watershed*, which extends the more traditional Watershed segmentation algorithm [2, 1].

The basic principle of Watershed segmentation is illustrated in Figure 2. Initially, an edge–detection filter is applied to the input image, producing as output an “edgeness” profile, *i.e.*, a monochromatic image where the intensity of each pixel is proportional to the likelihood of existence of an edge at that position in the original image. Watershed segmentation treats this “edgeness” profile as DEM, where

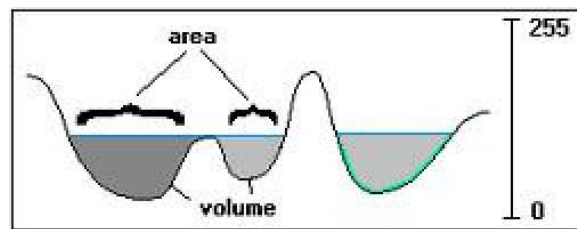


Figure 2. Watershed segmentation.

brightness values represent altitude. A simulation is then performed where the terrain represented by the DEM is gradually flooded with water, which gives rise to “lakes”. The borders of the “lakes” whose volume, surface area and depth satisfy certain pre–defined restrictions are then used as the segmentation boundaries in the original image.

Typical results obtained when this algorithm is applied to aerial images of urban areas, with different volume thresholds, are shown in Figures 3b,c. In both cases it can be observed that Watershed segmentation mistakenly breaks large structures such as the paved streets into several smaller regions. This happens due to a failure of the algorithm in situations where a “watershed” submerges into the rising water and two “lakes” that had been growing independently suddenly get in contact. At this point, the algorithm must decide whether the two “lakes” must be kept separate or whether they must be joined. In the traditional algorithm, this decision is based solely on attributes such as volume, surface area and/or depth of each “lake”, which leads either to improper segmentation of large, homogeneous structures, or to improper grouping of small, distinct structures.

Our contribution to ameliorate this difficulty is to introduce an additional color similarity criterion on the tests performed to decide whether contacting “lakes” should be joined or not. The resulting algorithm, which we denominate *Colored Watershed* segmentation, has a strong bias towards grouping homogeneously colored regions, even if they are already very large. The result obtained when this algorithm is applied to the same aerial image used to test its traditional counterpart is shown in Figure 3d. Notice the difference between the way in which the larger structures are segmented in this figure and the way that they are segmented in Figures 3b,c.

4. Dense Stereo

Dense stereo is a huge optimization problem. It amounts to finding what is the mapping between individual pixels of one image and corresponding pixels of a second image that minimizes some pre–defined global matching error. Research in this area has largely focused on finding restrictions on the set of possible mappings variously strong to make the problem manageable under various degrees of computational–power limitation [19, 21, 6, 17].



Figure 3. Watershed segmentation results. (a) Input image, (b-d) results with a small volume threshold, with a large volume threshold, and with combined volume and color thresholds, respectively.

Methods that are specifically designed to allow discontinuities in the mapping at occlusion boundaries tend to be more accurate than those that enforce smoothness everywhere, but also tend to be much more expensive computationally, because they have to “search” a much larger space of possible mappings to find the optimal solution to the problem. Here, we propose a compromise between these two extremes, in which discontinuities are allowed, but only near segmentation edges previously computed according to the methodology of Section 3.

So far, we have used this idea in a preliminary implementation of a stereo algorithm that takes as input a pair of rectified images and performs one independent optimization for each pair of corresponding epipolar lines. Using the ordering constraint [17], it reduces the problem of finding the mapping for each epipolar line to a shortest path problem that can be solved efficiently via dynamic programming.

In Figure 4 we display the results obtained when three variants of this dynamic-programming algorithm are ap-

plied to the stereo pair of Figure 1. The depth map in Figure 4a is the result obtained when the a priori segmentation is disregarded and it is assumed that depth is continuous everywhere. Although this variant correctly “highlights” the major architectural structures such as the tallest buildings, it tends to blur the actual occlusion boundaries, which is a typical shortcoming of most dense-stereo techniques.

This should be contrasted with Figures 4b,c, where jumps in disparity are allowed within each epipolar line, but only at the previously computed Watershed segmentation boundaries. More specifically, part (b) is the result obtained when the segmentation information is used to correct the depth map of part (a) a posteriori, so that depth becomes constant within each segmented object and possibly discontinuous at the object boundaries. Part (c), on the other hand, is based on the use of segmentation information on-the-fly, as one of the criteria that define the shortest path sought by the dynamic programming algorithm.

While both approaches reduce the blurring of actual

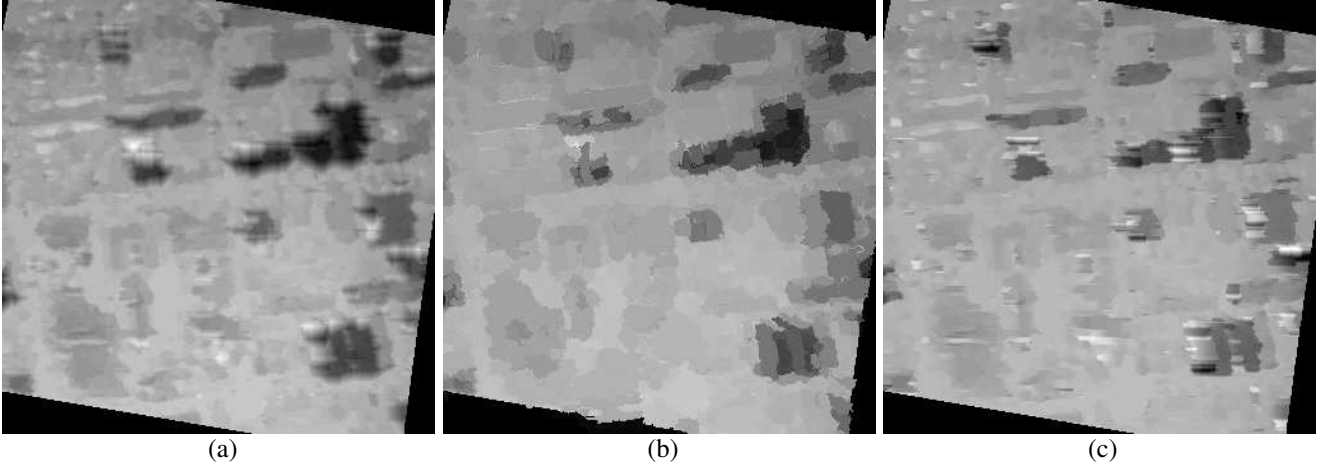


Figure 4. Depth maps recovered by applying dense stereo techniques to the images in Figure 1.

depth edges significantly, the variant that uses the segmentation boundaries on-the-fly in general makes better judgments about relative heights of the buildings. It does produce absurd results within the regions that appear in the left (reference) input image but are occluded in the right input image, but this is a shortcoming of our preliminary implementation, not a fundamental limitation of the technique.

This problem arises because our current implementation does not try to identify occlusions. It always tries to find a match for every pixel within each epipolar line of the reference image. Within regions that are occluded in the other image, this choice of best matches is uncorrelated with the scene’s actual 3D structure. And, to make matters worse, in our implementation no consistency is enforced between consecutive epipolar lines, which leads to absurd high-frequency variations in the vertical direction, as observed in Figure 4c. By embedding our idea of using segmentation boundaries on-the-fly within a stereo algorithm that models occlusions explicitly and that favors reconstructions that preserve continuity between epipolar lines (e.g. [17]), we should be able to overcome this kind of difficulty.

5. Fitting Parametric Models to Images

The techniques discussed in the last two sections are used, among other things, to create parameterized planar approximations of objects such as roofs, walls and pavement. A last major technical hurdle in our methodology is to use the input images to perform high-precision refinement of the geometrical parameters and the segmentation of these planar approximations. Fitting 3D planes directly to images is an idea that has been used with a great degree of success both for motion recovery [25, 13] and for shape recovery [5, 22] — although in the later case we are not aware of any effort towards refining the segmentation of the planar patches during the reconstruction process.

The common basis of all such techniques is the fact that any two perspective projections of a single plane are related by a very simple geometric transformation — a *homography* that can be represented as a 3×3 matrix [22]. If a plane is described by the general equation $\mathbf{n}^T \mathbf{p} = d$, where \mathbf{n} and d are, respectively, a column vector representing the plane’s normal, and a scalar representing its distance from the origin, then under the projective transformation performed by a camera with a 3×4 projection matrix $[\mathbf{R} \ \mathbf{t}]$, each 3D point \mathbf{p} in this plane is mapped to a pixel \mathbf{q} by:

$$\mathbf{q} = \mathbf{H} \mathbf{p}, \quad \mathbf{H} = d \mathbf{R} + \mathbf{t} \mathbf{n}^T, \quad (1)$$

where \mathbf{q} is expressed in homogeneous coordinates, and all equalities are up to a homogeneous scaling factor. Since the 3×3 matrix \mathbf{H} is, in general, invertible, a direct mapping between corresponding pixels \mathbf{q}_1 and \mathbf{q}_2 in any two projections of a plane is given by

$$\mathbf{q}_2 = \mathbf{H}_2 \mathbf{H}_1^{-1} \mathbf{q}_1. \quad (2)$$

Thus, a Maximum-Likelihood Estimator for the plane’s parameters, given the two images, can be obtained by minimizing in a least-squares sense the error vectors \mathbf{e}_i formed by the differences between the band intensities of the i -th pixel in the first image and its match in the second image under Equation 2.

A problem with this approach is that points that do not belong to the object’s plane, or that are occluded in one of the images, can not be matched, generating large errors in the alignment between the two images and potentially reducing the accuracy of the final solution obtained. A way of overcoming this difficulty is to use a robust error metric, which imposes a pre-defined maximum limit on how much individual pixels can influence the alignment process [3]. This allows not only a more reliable registration between the two images, but also the automatic identification of the unmatched pixels and, consequently, the automatic

refinement of the object's segmentation, as a side effect of reconstruction.

More specifically, we refine the initial description of the object's plane, obtained from the scene's segmentation and from the depth map computed with dense stereo, through the minimization of the following robust error metric [3]:

$$e(\mathbf{n}, \mathbf{t}) = \sum \rho(e_i), \quad \rho(e_i) = \frac{\mathbf{e}_i \cdot \mathbf{e}_i}{\sigma + \mathbf{e}_i \cdot \mathbf{e}_i}, \quad (3)$$

where σ is an empirical parameter that controls the threshold of influence of individual pixels — the smaller σ is, the smaller is the influence of outliers on the error metric.

After the robust error metric defined in Eq. (3) is minimized — using Levenberg–Marquardt's method — a pixel is identified as an outlier if its contribution to the total error is greater than $\sigma/\sqrt{3}$ [3]. Pixels identified as outliers are then excluded from the object's boundaries and their contribution to the error function is ignored.

6. Conclusion

The general methodology presented in this paper is a road map for our ultimate goal of obtaining accurate reconstructions of large-scale architectural scenes in an affordable and efficient way. In addition to this general plan, we have discussed in detail how the major technical hurdles in this task can be overcome. Although this paper reports work that is still in progress, one of the three major hurdles between us and our goal has already been overcome, as evidenced by the results that we presented for the area-based segmentation, and we are optimistic about our ongoing work on the other two.

This research has been supported by CNPq, by Fapemig and by PRPq-UFMG (Fundo Fundep RD).

References

- [1] M. C. Andrade, G. Bertrand, and A. A. Araujo. Segmentation of microscopic images by flooding simulation: A catchment basins merging algorithm. In *Proc. SPIE Non-linear Image Processing*, pages 164–175, 1997.
- [2] S. Beucher. Watershed, hierarchical segmentation and waterfall algorithm. In *Mathematical Morphology and its Applications to Image Processing*, pages 69–76. Kluwer, 1994.
- [3] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Comp. Vis. Image Understanding*, 63(1):75–104, 1996.
- [4] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for space carving. In *Proc. 8th Int. Conf. Comp. Vis.*, pages 388–393, 2001.
- [5] R. L. Carceroni and K. N. Kutulakos. Multi-view scene capture by surfel sampling: From video streams to non-rigid 3D motion, shape & reflectance. *Int. J. Comp. Vis.*, 49(2–3):175–214, 2002.
- [6] Q. Chen and G. Medioni. A volumetric stereo matching method: Application to image-based modeling. In *Proc. Conf. Comp. Vis. Pattern Recog.*, pages 29–34, 1999.
- [7] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proc. SIGGRAPH'96*, pages 303–312, 1996.
- [8] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Proc. SIGGRAPH'96*, pages 11–20, 1996.
- [9] O. D. Faugeras and R. Keriven. Variational principles, surface evolution, PDE's, level set methods and the stereo problem. *IEEE Trans. Image Processing*, 7(3):336–344, 1998.
- [10] A. Hanson, M. Marengoni, H. Schultz, F. Stolle, E. Riesenman, and C. Jaynes. Ascender II: a framework for reconstruction of scenes from aerial images. In *Proc. Int. Workshop on Automatic Extraction of Man-Made Objects from Aerial and Space Images*, 2001.
- [11] R. M. Haralick and C. nan Lee. Analysis and solutions of the three point perspective pose estimation problem. In *Proc. Conf. Comp. Vis. Pattern Recog.*, pages 592–598, 1991.
- [12] R. Hartley. In defense of the 8-point algorithm. In *Proc. 5th Int. Conf. Comp. Vis.*, pages 1064–1070, 1995.
- [13] M. Irani, B. Rousso, and S. Peleg. Recovery of ego-motion using region alignment. *IEEE Trans. Pattern Anal. Machine Intelligence*, 19(3):268–272, 1997.
- [14] K. N. Kutulakos. Approximate N-View stereo. In *Proc. European Conf. Comp. Vis.*, pages 67–83, 2000.
- [15] S. Moezzi. Immersive telepresence. *IEEE Multimedia*, 4(1):17–26, 1997.
- [16] P. J. Narayanan, P. W. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. In *Proc. 6th Int. Conf. Comp. Vis.*, pages 3–10, 1998.
- [17] S. Roy and I. J. Cox. A maximum-flow formulation of the N-camera stereo correspondence problem. In *Proc. 6th Int. Conf. Comp. Vis.*, pages 492–499, 1998.
- [18] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets or How do I organize my holiday snaps? In *Proc. European Conf. Comp. Vis.*, pages 414–431, 2002.
- [19] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comp. Vis.*, 47(1):7–42, 2002.
- [20] S. M. Seitz and C. R. Dyer. Photorealistic scene reconstruction by voxel coloring. *Int. J. Comp. Vis.*, 35(2):151–173, 1999.
- [21] C. Silva and J. Santos-Victor. Intrinsic images for dense stereo matching with occlusions. In *Proc. European Conf. on Computer Vision*, pages 100–114, 2000.
- [22] R. Szeliski. Video mosaics for virtual environments. *IEEE Computer Graphics and Applications*, 16(2):22–30, 1996.
- [23] E. Trucco and A. Verri. *Introductory techniques for 3D Computer Vision*. Prentice-Hall, 1998.
- [24] A. Yezzi and S. Soatto. Stereoscopic segmentation. *Int. J. Comp. Vis.*, 53(1):31–44, 2003.
- [25] L. Zelnik-Manor and M. Irani. Multi-frame estimation of planar motion. *IEEE Trans. Pattern Anal. Machine Intelligence*, 22(10):1105–1116, 2000.