Large scale semantic 3D modeling of the urban landscape

Esteban Lopez, I.

Citation for published version (APA):
Esteban Lopez, I. (2012). Large scale semantic 3D modeling of the urban landscape

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Conclusions

In this thesis we have treated the problem of large scale city-size reconstruction and modeling using a monocular camera. The proposed methods were designed to work in urban environments in order to obtain a 3D city model including the ground, the buildings and the rooftops. In order to record as much as possible of the buildings, the camera was moving at the street level, pointing towards the facades and moving in the direction of the streets. The resulting sequence of images pictured man-made objects, mostly facades and ground, which are typically flat. The images were recorded by either hand-holding the camera while walking or employing a vehicle mounted camera and driving around the city. In both cases, the camera would travel parallel to the ground with a mostly forward motion.

Most state-of-the-art reconstruction approaches rely heavily on numerical optimization in order to solve a set of geometrical constraints and to obtain a 3D reconstruction. Even though these methods have been shown to work well, they are computationally expensive and the details of the reconstruction process stay hidden behind the optimization procedure. These methods typically produce a set of points in space, the point cloud. Some techniques exist to model urban areas with 3D shapes, moving up from simple points. These methods analyze point clouds, aerial images or GIS data in order to obtain models of the buildings. These techniques have been shown to work on suburban areas where buildings are separated from each other, though they are not suited for tightly packed cities where the boundaries of buildings are sometimes indistinguishable.

A large scale semantic model of a city has potential applications in the fields of forensic investigation, assisted or autonomous navigation, city surveying or urban planning. The
goal of our research was twofold. Firstly, we wanted to obtain an accurate, fast and inexpensive method to perform 3D reconstruction. Secondly, we wanted to obtain a semantic 3D model of urban areas.

These two goals attempt to answer the questions we posed in Chapter 1, namely: What are the best algorithms? How many steps are in the reconstruction procedure? How do the choices that we make in each step affect the overall result? Is it possible to obtain accurately the positions in space of the images used for the reconstruction? Do we need an optimization procedure to refine the estimation of the model? Which representation to use? Are colored points in space a sufficiently rich descriptor for a scene? And finally, how do we bridge the gap between representation and understanding?

We approached the first of our goals by taking a detailed look into each of the required steps for 3D reconstruction. In order to do so, we designed a 6-step reconstruction pipeline (section 4.1). We analyzed the available algorithms and made the choices that were best suited for urban reconstruction (sections 4.2 - 4.7). We motivated the use of the 8-point algorithm given its better performance for typical motions of the camera in the urban environment (section 4.3.2.2). We proposed an inexpensive local optimization for every pair of cameras (section 4.5.1). By assuming the previous camera position and parameters had already been optimized, we proposed a chain optimization in which each camera was optimized sequentially. This procedure improved the estimated camera position and parameters at a low computational cost. Since cameras were referenced locally with respect to the previous camera, the procedure could be easily parallelized to obtain a fast and accurate motion estimation.

The motion estimation procedure revealed the scale problem (section 4.3.3): when estimating the translation direction of a camera using Image-to-Image approaches, the scale of the translation could not be recovered. Image-to-World methods implicitly solve the scale, though we showed the lower overall performance when estimating the rotation and translation. Instead of employing Image-to-World approaches to estimate the camera position, we employed the 8-point algorithm to estimate the rotation and a scale free translation. Then we proposed a method for estimating the scale with only one 2D-3D correspondence (section 5.3.2). Our method provided a closed form Least Squares solution when more than one correspondence was available. Furthermore, we provided a first order error analysis (section 5.4) and derived an optimal Maximum Likelihood method to estimate the scale very accurately (section 5.5.2). Our proposed method
performed significantly better (section 5.6) than alternative approaches in estimating both the camera positions and the scale of the translation.

The 8-point algorithm is known to suffer from degeneracies, in particular when all points lie in a plane on the scene. For these special cases, we proposed an optimization procedure that works with at least three images (section 4.5.2). We reconstructed corresponding matching points from two pairs of images. This produced two sets of corresponding 3D points. The optimization of the camera positions and parameters worked by minimizing the distance of the corresponding points in 3D space. We showed how this procedure could overcome the points in a plane problem and how it could be extended to optimize globally with a lower computational expense than standard Bundle Adjustment methods.

In our setting, the camera was moving mostly forward, parallel to the ground. If the camera would not move mostly forward, the 5-point algorithm should be employed to estimate the camera rotation and translation, instead of the 8-point algorithm. If the camera would not move parallel to the ground, methods 3 or 4 from Chapter 5 could be employed to obtain an optimal solution for the local scale. Local and global refinement techniques do not depend on the urban assumptions and only on the visibility of image features across different images. If the overlap between three consecutive images would be greatly reduced, the global optimization would suffer since it relies on matching features across 3 frames. In the same way the performance of local optimization would be reduced if the overlap between consecutive images were small.

We provided a theoretical basis for our proposed methods and showed experimental results to support them. We recorded an urban scene with a handheld camera while taking a stroll (section 5.6.2). For that 200+ meters trajectory the error was 0.14%, and more importantly, no global optimization was required. Similar results were shown with the wellknown P11 dataset. Our proposed six step reconstruction pipeline provided an alternative method to standard numerical optimization procedures in order to obtain fast 3D reconstructions of urban areas from sequences of images. Additionally, we implemented all methods in Matlab and created our freely available Toolbox called FIT3D (Chapter 7).

The second goal of our research was to model the 3D reconstructions by means of semantics. In particular, we were interested in estimating the direction of gravity (or the vertical direction), the topological map of the city, and a watertight model for each
building. These semantics provide fundamental information for machine understanding of the city landscape. We proposed a basic modeling method to fit and texture planar patches to the point clouds (section 4.6.2). These patches, while providing a visually appealing model, did not introduce any semantics. The fast and robust plane fitting method however was employed to find the facades in the point cloud. By assuming that, on average, facades are built perpendicular to gravity, we were able to find very accurately the direction of gravity (section 6.3). This direction is potentially useful for city scale surveying: forensic investigation, trajectory analysis, flooding scenarios, rain fall analysis, etc. Then we proposed a set of prototype methods to estimate the remaining semantic elements. We employed the direction of gravity to model the topological map of the city (section 6.5). By establishing a grid under the point cloud, we grew the grid in the direction of gravity so that it would fit the point cloud. Smoothing was applied to obtain a surface that represented the topological map of the city. The final semantic element we wanted to introduce was a watertight model of the buildings in the city. For this, we employed GIS data to obtain roughly the outline of each building (section 6.2.1). Then, we analyzed the distribution of the distances between the points in the point cloud and the outline of the building to find the rooflines (section 6.6). This allowed us to estimate the height of the facades to create a building model with a flat roof. Then, employing aerial images of the buildings, we extracted a set of control points for which the height was estimated to obtain a geometrical model of the roof, which was connected to the facades (section 6.7).

Our proposed semantic model represents a significant step towards automatic modeling of city. Our method is more robust than alternative techniques and less computationally expensive. It relies on the assumption that on average, the facades are built perpendicular to gravity. The estimation of the topological map is a true novelty. Related methods either assume the surface to be flat, or estimate ground points by analyzing the height change in the neighboring area, without estimating a complete surface. We showed experiments for the city of Lausanne and compared the resulting topological map with an isoline map. Most relevant ground features appeared to be correctly estimated. Our method, thanks to the smoothing stage, is able to roughly estimate the areas where no points were available. A topological map, together with the direction of gravity, is a powerful tool for large scale analysis and simulation on a city area. Final
building models were estimated. The height of the facades and the shape of the rooftop compared for two buildings.

The semantic modeling methods rely on point clouds of urban areas obtained from both the ground and the air. We assumed the ground based point clouds contain mostly points of the facades of buildings, which is essential for the estimation of the direction of gravity. These facades are assumed to be on average perpendicular to the direction of gravity, which is assumed to be the same for the complete model. If the ground based point cloud contained more points of the ground than facades or the facades were not perpendicular to one vertical direction, the gravity could not be estimated reliably (our proposed method for the estimation of gravity would not work on the Death Star in outer space). The preliminary methods presented to obtain the watertight models of buildings rely on the density, distribution and accuracy of the points. For the estimation of the height of facades, the outline of the building is assumed to roughly match the points in the area. If the outline in the GIS were oversimplified and would not fit the points, the height could not be estimated. The accuracy of the rooftop geometry depends on the estimated facade height and the accuracy of points that belong to the roof.

In this thesis we have provided the theoretical and experimental grounds to support our contributions in 3D reconstruction and modeling. We designed a reconstruction pipeline that can accurately recover the structure of an urban scene from sequences of images without the need for expensive numerical optimization. We provided a valuable insight into the steps required for reconstruction that would allow a real time implementation. This allowed us to identify the local scale problem. We developed the theoretical analysis of the error propagation and proposed an optimal method to estimate the local scale that outperforms state-of-the-art approaches. After obtaining a fast reconstruction procedure, we focussed on estimating semantics. We proposed preliminary methods to obtain valuable information at a large scale, obtaining a semantic 3D model of a city. These methods provide the starting steps towards a meaningful representation of a city. Our proposed novel methods for the estimation of the topological map and the height of facades significantly contribute to the tasks of automated city modeling.
8. CONCLUSIONS

8.1 Future Work

One of the most pressing challenges in the methods presented for semantic modeling is evaluation. In order to be able to use these methods across the application fields, it is necessary to understand how accurate they are and under which conditions they will break. We have shown some preliminary evaluation for the estimation of gravity, which we evaluate with the measurements obtained with an INS sensor. A deeper error analysis would be desired involving the sensor accuracy, number of readings, etc. For the evaluation of the ground surface, we showed a topological map of the city of Lausanne. A numerical evaluation would also be desired. A more detailed topological map could be obtained along with the height of isolines. Then the registration could be obtained based on GPS information in order to obtain a numerical evaluation. The evaluation of the methods described to estimate the height of facades and the geometry of rooftops is even more challenging due to the lack of availability of ground truth information. If accurate images are available at the ground level with known camera parameters (position and calibration), the height of buildings could be obtained by manually indicating the ground and rooftop points, and then establishing a real life measure based on some known measurements (for instance the width of a window). This however would be a painstaking job and could only be done for a handful of buildings. More accurate laser readings could be employed, though at this time it is difficult to envision an accurate and reliable large scale ground truth measuring process. The same issue affects roof topology. A hand annotated method could be employed for the 2D geometrical structure of roofs. Applying this at a large scale and obtaining the real 3D structure seems however difficult.

Images are commonly used in the literature for estimating features and computing camera poses. We used then in the methods presented as a source of features to obtain points in space and also features to obtain the geometry of roofs. However, images contain a lot of semantic information: doors, windows, cars, facade outlines, etc. Many of these elements could be extracted employing image analysis and machine learning techniques. Once these elements are found in the images, they could be mapped to the 3D model of the city.

We have employed in our methods for modeling large scale point clouds. However, the original images that were used to obtain the point clouds were not available. An
alternative source of information would be Google Street View images. These images are geolocalized, therefore the registration with our model would be straightforward. The challenge would be the lack of camera parameters and lens information, without which the mapping from image to model would be very difficult.

For many years, state-of-the-art methods have been focused on obtaining accurate camera position and calibration parameters. The semantic information in the images and the models had been somehow forgotten. We are now at a point where camera pose and calibration can be estimated very accurately and very fast. We believe that the emphasis of city modeling is going to be semantics and the ability to use those for large scale applications.