Large scale semantic 3D modeling of the urban landscape

Esteban Lopez, I.

Citation for published version (APA):
Introduction

Creating and understanding digital 3 dimensional models of large urban areas is becoming an important topic both in science and industry. A semantic and accurate representation of a city can be used in many applications such as assisted navigation, autonomous operations, city surveying, forensic investigation or urban planning, to name just a few. There are many possibilities with respect to how to obtain the necessary information to build a model. Naturally, one could approach the problem by using simply user input. This is however a painstaking job that requires advanced knowledge of modeling software along with knowledge of the environment. Alternatively, some source of information could be used to guide the reconstruction. This information could exist as a set of measurements or geometric constraints that describe in some way the layout of the environment. Some sources of information include data recorded by lasers, ground images, aerial or satellite imagery and Geospatial Information Sources (GIS).

Laser data is an accurate source of information, where the device records distance measurements between itself and the objects in the environment. These measurements can then be translated into points in 3 dimensions. If we obtain sufficient number of points, we have a point cloud in 3D space. Laser devices are however expensive machines, difficult to use and slow in recording. Also, the measurements typically consist only of a set of distances to points, though some state of the art laser models also record color data for each point. These sets of points are merely geometric representations and cannot be considered a model. We define a semantic model as a 3D representation that
1. INTRODUCTION

contains information about the environment that is not purely geometrical, namely: the
direction of gravity, the shape of the ground and a geometrical model of buildings.
Images can also be used to obtain information. We, as humans, are used to creating,
understanding and interacting with a world in 3D. Based mostly in our ability to see
our surroundings, we create a mental representation of the scene or object in front of us
where we can establish spatial relationships such as depth, position or shape. Recording
images is also cheaper and easier than operating laser devices. Therefore vision based
methods for reconstructing a scene in 3D have received much attention in the current
research. Single-camera, or monocular, systems are particularly popular given their
flexibility and low cost. Some of the first attempts try to mimic the stereoscopic nature
of the human vision.

GIS data typically consist of 2 dimensional information of urban areas. It contains
information about buildings, streets, public services, etc. Professionally made GIS data
is expensive since it is manually created by experts employing accurate GPS data.
Some freely available and user-contributed initiatives exist such as the Open Street
Map projects where users manually annotate information about the city. GIS data is
typically geolocalized though only 2 dimensional.

This thesis deals with the problem of large scale city-size reconstruction and modeling
from a monocular camera. The goal of the research was twofold. Firstly, we wanted
to obtain an accurate, fast and inexpensive method to perform 3D reconstruction. Sec-
ondly, we wanted to obtain a semantic model of the reconstructed environment.

Given the flexibility, availability and low cost of consumer cameras, we focus on the use
of monocular off-the-shelf cameras to model urban environments. This will render our
methods flexible and inexpensive but also bring certain algorithmic issues that need to
be addressed. The characteristics of the monocular urban setting are:

- An urban environment consist of man-made structures such as roads and build-
ings.

- We are interested in modeling buildings and roads, avoiding movable objects like
cars or pedestrians and ignoring smaller structures such as mailboxes or lamp
posts.

- Sequences of images are recorded while the camera is moving.
1.1 From 3D to 2D, the Camera Paradigm

• In order to capture as much of the buildings and roads as possible, while maintaining overlap between consecutive images, these are recorded pointing diagonally towards the buildings.

• The camera therefore moves along the streets, with a mostly forward motion.

• Facades and ground consist mostly of planar structures.

• Given the large overlap between consecutive images, the reconstruction process is likely to produce a large amount of points.

Given the extent of the field, there are many important questions for which we want to find an answer. 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?

In the remaining of this chapter, we first explore the camera paradigm and explain the basic idea of an imaging process. During this procedure, we reduce the world from 3 dimensions to just 2 dimensions. Then we introduce the basic problem of using the recorded data to obtain a 3 dimensional representation, effectively restoring the third dimension. Finally, we state some of the challenges along the way and give an outline of this Thesis.

1.1 From 3D to 2D, the Camera Paradigm

The origin of the word camera is the Latin camera obscura, referring to an optical device to project images on a screen. It was originally used for entertainment or, when projecting onto a piece of paper, as an aid to trace the scene easily. The device, in its outmost simplicity, consist on a dark chamber with a tiny hole in one side. The light rays pass through the hole projecting on the opposite wall. This produces an image upside down and it is the origin of modern cameras and the imaging process.
1. INTRODUCTION

The next step of these simple camera devices consists of placing one or more lenses in front of the hole. The word lens derives from the Latin word for lentil because of its shape. The oldest known artificial lens is dated more than three thousand years old. Lenses change the trajectory of light rays. This can have a magnifying effect on the resulting image, or a focussing effect where the light rays are concentrated in a small point (which can used, for instance, to start a fire). When placed in front of the camera hole, the effect is translated in the way the image is displayed on the projecting surface. Effects such as zooming or distorting the image can be obtained. Modern cameras use lenses mostly to improve the image quality by allowing more light while keeping the image in focus. They are also used to zoom in/out.

A complete camera device consists of a lens placed in front of the hole where the rays of light pass through, projecting on a surface beyond. This procedure, called projection in mathematical terms, creates a representation of the original object where the number of dimensions is reduced from 3 to 2. The projection procedure is explained in Chapter 2.

1.2 From 2D to 3D, the Reconstruction Process

Understanding the imaging process and the dimensionality reduction from the real scene to the image is only the beginning of the path towards 3D reconstruction. The true and intriguing puzzle starts by wondering if this imaging process can be reversed. By reversing it we expect to recover the local depth of the image so we can obtain a 3 dimensional representation of the scene, or as we described it before, a 3D reconstruction. We humans create a mental representation of the things that we see. This representation allows us to infer spatial relationships such as depth, shape or geometrical relationships. Since our eyes work in a similar way as a camera, we can almost be certain that the process must be somehow reversible. There is however, a small difference with respect to a camera as we have described it. We have two eyes, or two cameras that record the scene at the same time from two different points of view. This makes us stereoscopic camera systems. So even when we look at a static scene without moving our head, we perceive and understand depth, which is the basis of a 3D reconstruction.

The human ability to understand and process depth information is out of the scope of this thesis, partly because our mighty brain does not only use images but also details
1.2 From 2D to 3D, the Reconstruction Process

stored in our memory and based on our previous experiences. However, if we explore
the very basic principles of human vision geometry, we hope to use this information to
infer methods and techniques to retrieve the 3 dimensional shape of imaged objects.
Complete details on camera models are given in Chapter 2. However, simple intuition
goes a long way. The object projection onto the camera is simply a ray of light bouncing
from an object and then onto the camera recording sensor. So in principle, if we
backtrace the trajectory of the ray, we can be certain that we will find the object at
some point. The evident problem is that we cannot know the distance to the object,
we can only be certain of the direction in which it can be found. If we now consider the
fact that we have two cameras mounted in our head, we could backtrack the light rays
that bounced off the same object point and impacted on both camera sensors. This
principle is shown in Figure 1.1. Then we can be certain that the backtracked rays will
intersect in space where the object is. This process is called triangulation and a basic
principle in 3D reconstruction.

![Figure 1.1:](image)

We now know that if we have two cameras, for which we know the position and orienta-
tion, and if we are able to identify an image point in both cameras that corresponds to
a ray of light that bounced off the exact same spot in an object, we can then backtrack
those ray trajectories to find the object and therefore estimating its position in a 3
dimensional space. Even though this is only the intuition behind the reconstruction
1. INTRODUCTION

process, it serves well as an introduction to the problem of obtaining a 3 dimensional meaningful representation of an imaged scene.

We will use only one camera; fortunately, under many circumstances the same procedure could be applied in this case. If we set the constraints and assume the scene is static but we are able to move the camera and then retrieve its position, we can still perform the reconstruction by the backtracking method.

This intuition already sets the requirements of a 3D reconstruction using single camera systems:

- A camera that is able to record the scene by projection of rays of light
- The means to move the camera around a static environment
- A method to estimate the position of the camera in space
- The ability to identify corresponding projected rays (or corresponding image points)
- A backtracking method to find the object in space

Having understood the camera recording process and the intuition behind the reconstruction process, in the following section we detail the most important challenges that we will face in 3D reconstruction.

1.3 The Challenges Ahead and Our Contribution

In the previous section we sketched a simplified version of the steps of a reconstruction pipeline where cameras are employed to record a scene and the resulting images are in turn used to obtain the reconstruction. However the exact details on the different steps that need to be performed are yet unspecified. Variations in both the pipeline design and choice of algorithms for each of the steps will have a definitive impact on both the quality of the model and the computational complexity. The reconstruction procedure we envisioned would find an imaged object in space, however no meaningful information is obtained.

The challenges ahead are focused on those two areas: the accurate estimation of the position of cameras in space, and the modeling of reconstructed scenes.
1.4 Overview of this Thesis

State-of-the-art methods approach the problem of reconstruction and camera position estimation as a single procedure. Computationally expensive numerical optimization methods are typically employed that aim at solving a set of geometrical constraints. Our contribution for the first challenge focuses on the analysis of the different steps of the reconstruction process, and the tailored design of an urban modeling pipeline where accurate camera positions and reconstructed structures can be found.

Obtaining a reconstruction in 3 dimensions is however not the end of the line. If these models are ever to be used in other tasks besides visually appealing multimedia, some information needs to be inferred. The process of semantic modeling clearly represents the door to highly profitable uses of the reconstructed data, specially when dealing with the city-size reconstructions. State-of-the-art methods applied to urban scenes aim at modeling mostly stand alone buildings. The modeling process is usually done by means of a set of predefined shapes that are fitted to the reconstruction. Topological information is rarely and barely modeled. Our contribution in the area of semantic modeling focuses on obtaining certain topological and physical characteristics of the urban scene, and on modeling the city as a collection of individual buildings with unique properties.

1.4 Overview of this Thesis

The remaining of this Thesis is organized in seven chapters.

In Chapter 2 we specify the visual geometry concepts presented in the previous sections and put them into a mathematical context. We formalize the camera model and the details of the use of lenses. We also define our choice of geometrical reference frames and present the mathematical details of the projection process.

In Chapter 3 we present an extensive literature review, exploring the field of 3D reconstruction categorized based on different aspects such as user driven versus automatic procedures. This review allows us to put our research into context and points to the goals of each of the following chapters.

In Chapter 4 we design the 6-step reconstruction pipeline for large scale urban environments. We introduce the mathematical concepts, motivating our decision for each step of the algorithm. We focus on the development of inexpensive and fast methods to obtain a 3D reconstruction from sets of consecutive images.
1. INTRODUCTION

In Chapter 5 we focus entirely on the mathematical details of the estimation of the scale of the camera motion. We describe the least square solution and our optimal method for the computation of the local scale.

In Chapter 6 we focus on large scale semantic modeling of a city. We explore the different sources of information that are used to create the semantic models and devise a method for the correct registration of the data. We propose a method to accurately estimate the direction of gravity (or vertical direction). This is then used to estimate the isosurface over which the city is settled. Finally a parametric model of the building and the rooftops is obtained.

In Chapter 7 we introduce our Matlab toolbox FIT3D, which implements most of our work in 3D reconstruction. We analyze the motivations and goals of this publicly available toolbox and compare with state of the art alternatives.

Finally, in Chapter 8 we summarize the contents of this Thesis, present the overall conclusions and point to some directions for future work.