Structure from Visibility: Finding low-texture and reflective occluders in outdoor scenes using visibility and occlusion

Student: Martijn Frederik Wouter van der Veen
Supervisor: Gabriel J. Brostow
MSc Computer Graphics, Vision, and Imaging
September 2012

This report is submitted as part requirement for the MSc Degree in Computer Graphics, Vision & Imaging at University College London. It is substantially the result of my own work except where explicitly indicated in the text.

Department of Computer Science
University College London

Abstract

Contents

List of Figures

List of Algorithms

Chapter 1

Introduction

The understanding of physical scenes by machines has been of great interest for researchers over the last couple of decades. Being able to construct the three dimensional shape of objects in a scene can be very helpful in this task. Based on the observation that, for humans, visual clues seem quite useful for understanding the structure of their environment, it is not surprising that researchers in the field of computer vision have put a great deal of effort in trying to build algorithms that construct 3D models out of images. Being able to reconstruct geometry allows for new applications. On a practical level, for example, it allows for easy creation of 3D models for usage in modelling design, computer games and even 3D printing. In addition, geometry reconstruction as an intermediate step opens the door to further processing in advanced street scene understanding, advanced geometry-aware rendering of new or editing of existing footage (such as changing lighting or virtually moving objects in the scene), navigational tasks in robotic systems, and augmented reality applications[1]. Increasing computer power, cheap and omnipresent cameras, and recent developments in the field of computer vision and learning all help making these exciting applications a reality.

There has been published a great amount of research on geometry reconstruction. A variety of approaches has been proposed, each with their own advantages and disadvantages. Most, if not all, of the approaches seem to be based on one or more visual clues that are relatively easy to find automatically, such as silhouettes, corner points or similarly coloured pixels in images. Different visual clues call for various requirements in the input images, resulting in successes and failures specific for the chosen approach. Current approaches often rely on the ability to detect features such as silhouettes or distinctive corner points on *all* geometry, resulting in bad results for objects where the specific features are less present.

However, detected features are not the only pieces of information that can be used for geometry reconstruction; failure of detection can provide useful information too. Detection failures can be detected by using information from features detected in other parts of the data (*e.g.,* other images). In particular, the absence of features in images were they would have been expected due to detection in images taken from surrounding camera poses may indicate the existence of another object. Although this observation might seem obvious, it has not got a lot of attention in the computer vision literature. Bad matches or the

[1] One could even state that geometry reconstruction could be seen as an intermediate feature in advanced computer vision, much in the same way as edge detection is used extensively as a feature in more advanced image processing algorithms.

absence of features have mostly been threated as outliers or as an indication for not using certain data in a particular step in the reconstruction; it has seldomly been used actively for reconstructing geometry[2]. Often camera poses are thrown away after triangulating the feature locations, continuing with the positive information (features) only. Intuitively, this seems like throwing away useful information. It is this intuition that motivates the approach taken in this thesis.

In this thesis, a framework is proposed that uses both *visibility* (detected features) and *occlusion* (undetected but expected features) in order to find, and reconstruct, objects not rich in features. The proposed method uses features detected on other parts of the scene to find objects difficult to detect directly. Based on the visibility and occlusion information of features from given observation locations, space is 'carved' in a probabilistic way. The method has been tested on a variety of outdoor scenes and, given enough detected features on objects in the background, gives high probabilities for regions containing objects difficult to detect with widely-used alternative methods. While space-carving based on silhouettes has difficulties with high-textured and complex outdoor scenes, and multi-view stereo is challenged by low-textured and reflective objects, the suggested method targets at scenes containing both and aims at combining strong properties of both methods.

In the course of the project that resulted in this report, most of the time has been used for the practical implementation and evaluation, in addition to a comprehensive literature review and documentation. Various software libraries have been tested and used to develop not only the method described here, but also several visualisation tools helpful for the visual thinking process necessary for developing geometry algorithms. Since our results are inherently three-dimensional and this report is not, the reader is encouraged to view the supplemented digital material - either using the provided tools or pre-made videos - alongside reading the report.

[2]It is almost as if researchers are afraid of negative information.

Chapter 2

Background and Related Work

Geometry reconstruction has drawn a lot of attention in the computer vision community (Hartley and Zisserman [10], Seitz et al. [26], and Prince [24]). The goal is to reconstruct the three dimensional structure of a particular scene given the data, usually multiple images (Multi-View Stereo). The images can come from different cameras (unordered set), or from the same camera (sequence) that potentially has been moved around. For many algorithms the relative locations from the cameras are required. They are either obtained by calibration between the different cameras or images (*e.g.,* by some identifiable pattern visible in all the images, or measured using external sensors), or estimated based on features naturally occurring in the images. When the positions are known, all the pixels in an image are known to be a projection from a particular point somewhere on a line through the camera centre and the pixel on the image plane, with the location of the point on the line being the only ambiguity. The next task now is to estimate the geometry of the scene pictured in those images, given this ambiguous information. Different approaches exist to tackle this problem, of which we will discuss four categories: layers (Section 2.1), space carving (Section 2.2), structure from motion (Section 2.3), and Multi-View Stereo using Depth Maps (Section 2.4). We like to note, however, that often multiple approaches are combined and overlap exist. Not all the techniques being discussed are directly relevant to the proposed algorithm, but they give a sense of the variety of approaches being published and try to give a general overview of the geometry reconstruction research literature.

2.1 Layers in Video

One of the earlier, and in some aspects simpler, attempts to represent geometry in images is the use of layers. Images are assumed to be a composition of multiple, possibly overlapping, surfaces that can move relative to one another from frame to frame in a sequence. Objects pictured in images can be at different distances to the camera, resulting in overlapping surfaces when projected onto an image (*i.e.,* captured). Successfully identifying coherent surfaces, including their mutual occlusions and movements from one frame to another, allows for a scene representation consisting of *layers*. Such a representation usually consists of one background layer plus one or more foreground layers representing objects moving in front of the background layer.

As one of the first to suggest representing sequences of images with layers, Wang and Adelson [29] decomposed images into 'motion layers' by analysing the motion of extracted segments in the subsequent images using optical flow. Segments with similar motions are combined, and grown and depth-ordered by tracking them over the sequence, resulting in more extensive layers. The layers can then be used in combination with estimated simple motion models to reconstruct the sequence in a memory-efficient way. Multiple extensions on this scheme exist. For example, Jojic and Frey [14] subdivided layers into flexible sprites that can vary their shapes according to a probabilistic model. The sprites and their probabilistic models are learned using the Expectation Maximisation algorithm on a representation consisting of one appearance and one mask pixel vector per sprite. Theoretically, the unsupervised EM learning approach even allows for finding sprites in related but unordered photo collections. Another approach, suggested by Smith et al. [27], is to extract and track edges over the sequence in order. Motion models are then fit to the edges using the EM algorithm. The edges, which form a better criterion for segment borders than the locally-smoothed optical flow vector field, are used to find a good segmentation of the frames and segments with similar motion models are again merged, thereby resulting in layers.

Although layered representations do have the notion of depth-ordering - and thus occlusion - they do not aim at recovering the exact depth of the layers, and inherently do not reconstruct three-dimensional geometry. In the next section, we move on to three alternative approaches that aim at true geometry recovery.

Layers are relevant for the current work because initially the intuition was raised that footage of an outdoor scene can be segmented into layers by tracking feature points and noticing when occlusions happen. However, for these feature points we can estimate the three-dimensional location by using Structure from Motion techniques (Section 2.3), allowing to reconstruct more than relative order. Hence, the step from layers to real three-dimensional reconstruction was made and the proposed method as described in Section 3 originated.

2.2 Space carving

While layers in videos try to represent geometry by their projections onto the image planes, thereby ignoring the original three-dimensional geometrical shapes, space carving aims at estimating the actual surface of objects in a scene. To reconstruct 3D from flat projections, space carving assumes the camera poses are known, *i.e.,* we have calibrated cameras. The poses are either known due to explicit calibration (using a calibration object or external sensor), or estimated based on the natural content of the images.

Usually, space carving is based on finding silhouettes [10, 19], thus segmenting each of the images into a binary background and foreground 'layer' or mask. These techniques are also known as Structure from Silhouettes. The pixels in the images are known to be projections of points somewhere on lines from the camera centres passing through the pixels, but the locations on the lines (*i.e.,* depths) are unknown. Space carving based on silhouettes[1]. does not estimate this depth; instead, it uses the background-foreground masks to categorise each line as either intersecting with an object, or not. The foreground mask therefore represents a cone of space in which the projected object must lie, at unknown distance.

For reconstruction, usually a 3D volume is chosen that lies in between the camera poses, and the volume is discretised and represented as a *voxel grid* - although polygonal representations have been proposed too. Space carving starts with all the voxels marked as 'object'. For each camera pose, all the voxels are projected onto the image plane. Voxels projecting to pixels not set as foreground (*i.e.,* onto the background mask) are set to 'non-object', finally resulting in the intersection of cones in which the object must lie (see Figure 2.1). This intersection is called *visual hull* or *convex hull*. It can be used as the final reconstruction, or used as a useful prior for further processing as the maximum (convex) boundary in which search can continue.

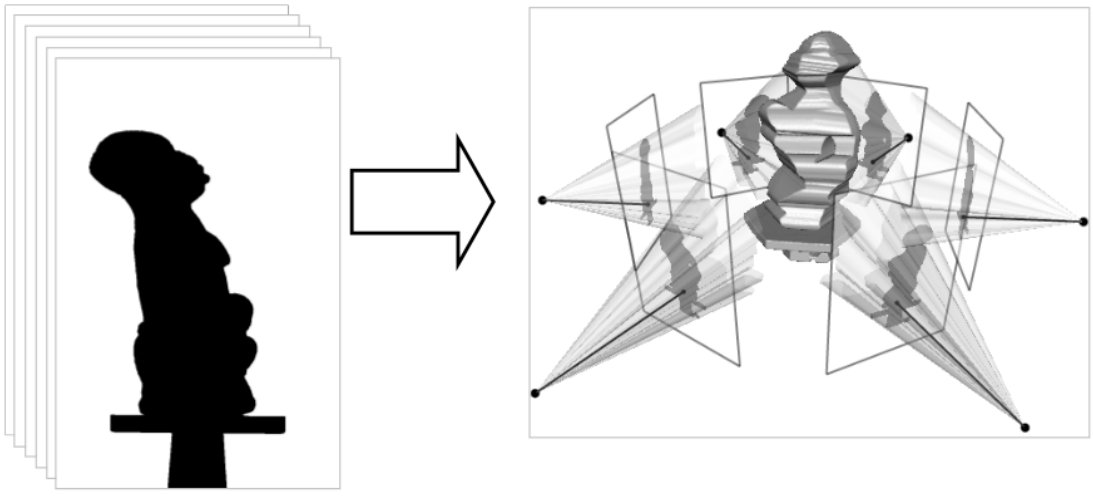


Figure 2.1: Carving using silhouettes. Image taken from 'How to prepare and deliver a presentation' [ppt] by R. Cipolla, who adopted it from unknown source

Segmenting images in foreground and background masks can be accomplished by a simple threshold, or by learning more advanced colour models for both. This can be user-guided or automated. For

[1] Part of the literature reserves the term Space Carving for space carving based on photo-consistency as described later on; here, it is used for both photo-consistency and silhouette-based space carving iteratively refining a model represented by a voxel grid.

example, Campbell et al. [3] automatically learned foreground colour models using pixels at the centre of the images (requiring the user making the footage to fixate at the object of interest) and background colour models using pixels near the image corners. The initial colour models are used together with the calibrated camera poses to make an initial binary segmentation in a voxel grid. The segmentation is then used to refine the colour models, and the process is repeated until the solution converges. As is common in algorithms involving voxel grid representations, Campbell et al. [3] use a Markov Random Field (MRF) prior for the segmentation step, promoting local smoothness for obtaining a global solution. A min-cut/max-flow graph-cut algorithm like the one presented by Boykov and Kolmogorov [1] is used to find the solution. This process is also known as regularisation.

Using more pictures in the space carving approach will give better approximations. However, concavities are not visible from silhouettes, so they will not turn up in the final reconstruction. The acquired visual hull therefore is only a broad approximation of the object's surface. Space carving is very suitable to do on a turn-table, for which the background model is known and also the relative camera poses are known due to controlled turn-table rotation. However, a turn-table is only feasible for objects small enough to put on the device. Placing a green screen behind objects is also a possibility, resulting in the same easy silhouette segmentation. Green screens have their limits too, making it an unsuitable remedy for outdoor scenes, where the background often is not controllable. Furthermore, space carving is not suitable for dynamic scenes. As a last disadvantage, multiple objects are not easy recognisable in a single silhouette and will result in ghosting objects [9].

Extensions have been proposed to overcome the limited applicability of carving based on silhouettes. For example, Guan et al. [9] added time-dependent Bayesian probability to the voxel variables, allowing space carving to be used in dynamic scenes. In their representation, voxels represent probabilities on occlusions before and after the voxel on a line passing through the centres of a set of calibrated cameras, filming from fixed locations. Using time-dependencies, their algorithm is able to follow dynamic objects over time, *e.g.,* a human walking through a temple. To find silhouettes, a per-pixel Gaussian background colour model is made for every camera before the dynamic object enters the scene. The estimates of the silhouettes are used as a prior for the next time frame, allowing to spot the object moving behind static occluders. When the object moves extensively through the scene, static occluders will turn up when they occlude the object. Even concavities in static occluders can be detected if an object moves, in fact allowing to carve away all human reachable locations by walking through the scene. In a later publication, Guan et al. [17] extended their algorithm to detect and label multiple dynamic objects. Each time a new, unknown silhouette (that differs enough from the colour models learned so far) enters the scene, a new appearance colour model is learned. Although the algorithm only works with fixed cameras and dynamic objects, it does have notion of visibility and occlusion.

An alternative to space carving with silhouettes is space carving based on photo-consistency, as introduced by Kutulakos and Seitz [16] . Hereby, voxels are projected onto all image planes that are able to see the voxel according to the current voxel grid. The set of pixels a voxel projects to are then checked for photo-consistency in compliance with a reflectance model (BRDF). The photo-consistency

check returns the probability of this set of pixel colours originating from the same surface point. For a low photo-consistency score, it is unlikely that this voxel is part of a solid object, and so it is carved away. The process continues until all the remaining uncarved voxels have a high photo-consistency. Although results look promising, the algorithm has some problems with low-textured (*i.e.,* uniformly coloured) objects and objects with repetitive textures. Furthermore, the method does not scale very well to bigger environments with lots of same coloured objects, and assumes Lambertian surfaces for all objects. In practise, pixels usually have no unique colour and reflective objects are not uncommon.

As a last example in the space carving literature, we discuss the work by Hernández et al. [11], who further explored the concept of visibility using the photo-consistency criterion. The algorithm introduced by Kutulakos and Seitz [16] is extended to include a probabilistic voxel grid: voxels are not carved one by one, but assigned the photo-consistency probability for being part of the surface of an object [2]. A second voxel grid represents the visibility evidence, containing probabilities of voxels being visible by at least one camera. This will give a penalty to voxels that are presumed not to be visible from any camera, and thus not part of the surface of an object. As a final step, graph-cut is used to combine the photo-consistency and visibility voxel grids, obtaining the assumed surface. Note that occlusion is being used, but only to exclude voxels inside objects, not for refining the shape itself.

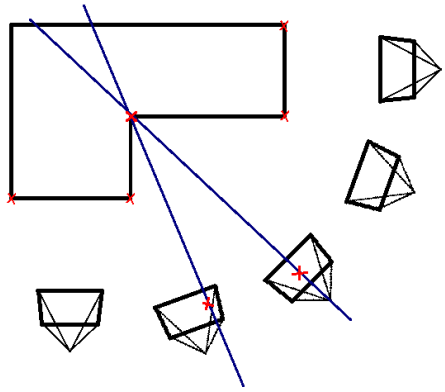
2.3 Structure from Motion

Geometry reconstruction usually involves measurements in the form of images, that is, flat projections of the three-dimensional world at specific locations (and times). If these locations are known, the process of reconstructing the original scene becomes easier. Therefore, much effort has been put in reliable pose estimation. Structure from Motion, as described extensively by Hartley and Zisserman [10], tries to accomplish this. Structure from Motion (SfM) uses natural features in images in order to estimate the relative translation and rotation between the camera poses of different images. It is based on the intuition that moving your head (motion) and noticing differences of the different views allows for a rough estimation of objects and distances (structure) in a scene.

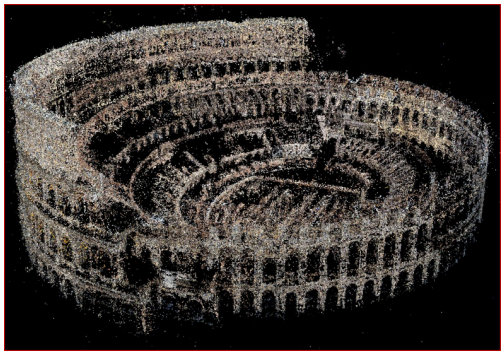
For SfM techniques, the camera poses are unknown a priori and need to be estimated based for a set of given images. As a first stage, features such as interest (corner) points are found in all images. One of the most popular features is SIFT, as introduced by Lowe [18]. SIFT features are found by searching for maxima in scale-space, constructed with Derivative of Gaussian filters. For the detected locations, a unique SIFT descriptor is constructed based on a histogram of gradient directions of points nearby. SIFT features are scale and rotation invariant and partially intensity and contrast change invariant. Alternative features can be used too, as long as they can be matched uniquely between images. In the second step, the interest points are matched between images using their descriptor. Using the matching pairs between images allows for estimating the relative transformation from one camera to another by satisfying projective geometry constraints (Prince [24], Ch15-16, Hartley and Zisserman [10], Ch10-12). For every image pair with estimated transformation, 3D locations of matched interest points can now be estimated

[2] In fact, a fast depth-map (Section 2.4) computation is done and photo-consistency is calculated only for locations with 3D points nearby

by triangulation (Fig. 2.2(a)). The depth ambiguity for *some points* is now eliminated in each image. The image pair point clouds - possibly containing feature points seen in multiple image pairs - can be combined using Bundle Adjustment (*e.g.,* , Wu et al. [30], Hartley and Zisserman [10], Ch18), which minimises noise and estimation errors in point and camera pose locations and obtains a global solution. The result is a *sparse point cloud* (Fig. 2.2(b)), camera models (including pose), plus visibility lists with camera-points pairs. As an example of using alternative features, Chandraker et al. [4] developed a Structure from Motion system based on lines, arguing that indoor environments and scenes containing much human-made objects usually lack distinctive feature points, but a rich of straight lines. Line geometry constraints are used to estimate camera translation and rotation in a similar way as in the case points are used.



(a) 3D point location estimation by triangulating two matched feature points in a pair of images



(b) Sparse point cloud. Image taken from http://grail.cs.washington.edu/rome/dense.html (1st of March 2012)

Figure 2.2: Structure from Motion

After reconstructing both the sparse point cloud and camera poses, usually one of the two is thrown away and the other is used in a subsequent step. Given the sparse point cloud, one can attempt to reconstruct geometry based on the sparse, but reliable points. Different techniques have been used [26]. One simple way of doing this is fitting a polygon mesh to the point cloud. More advanced smooth surface fitting is also possible, as has been done for example by Izadi et al. [13], but it eliminates strong edges and flat surfaces which are overrepresented in real scenes. Another method is fitting primitive shapes like planes and lines to the point cloud, which works well in human-made environments, but not so well in outdoor scenes. In all cases, assumptions are made on possible interpolation between points close to each other, and lack of texture results in gaps and sometimes missing objects in the reconstruction. In practise, sparse point clouds are rarely used for surface reconstruction. More often, the camera poses are kept and used in combination with the original images using a subsequent algorithm (*e.g.,* space carving (Section 2.2) or depth maps (Section 2.4)), obtaining a more dense representation before surface reconstruction is attempted.

2.3.1 Visibility and Occlusion

An important observation that did not got a lot of attention in the literature is the fact that the visibility of the points over time (*i.e.,* over the image sequence) can not only help selecting and fine-tuning points in either sparse or dense point clouds (as noted in the survey of Seitz et al. [26] and used in [21, 11]), but also directly used for surface reconstruction. In other words: occlusion rarely has an essential role in geometry reconstruction. However, visibility tracks over time can help both finding empty space (because we can see points at a certain location from a certain location) and finding occluder objects (because we lost track of a point for a while). Structure from Motion - in addition to pose estimation - therefore does not only deliver sparse points on surfaces, but also gives information about the occupancy state of space between the camera poses and reconstructed sparse points, without the need to find any visual clues at that locations. This observation has an essential role in the proposed approach (Section 3).

Despite the scarcity of occlusion applications, recently there have been some attempts worth a brief discussion. For example, Zach et al. [31]extent the structure from motion pipeline by an outlier check by using 'missing correspondences'. Their method collects triplets of images connected by fair amounts of correspondences. Two of the images are used to triangulate the (sparse) correspondences which are projected into the third image. A probabilistic measure based on the number of found and missing correspondence points in the third image determines whether to discard the third image as a match or not. Pan et al. [22] propose geometry reconstruction (called ProFORMA) of single, well-textured, objects filmed with a fixed-position camera by calculating a sparse point cloud, converting the point cloud into a mesh using Delaunay tetrahedralisation, and carving away tetrahedra who violate visibility constraints. The carving consists of using intersections of rays from camera poses to points to determine - probabilistically - whether to carve away a intersected tetrahedra, finally obtaining a surface mesh. Lovi et al. [5] independently developed an almost identical system (called free-space carving) that can be used on more complicated scenes. Both methods use positive visibility information to carve away tetrahedra; however, negative occlusion information is not being used as tetrahedra are assumed to be occupied by default. Furthermore, the tetrahedralisation is based on triangulated feature points, requiring texture to be omnipresent and determinative for quality and resolution of reconstructed models. Recently Mcllroy et al. [20] (2011) explored the possible use of both visibility and occlusion information with the aim of constructing high-level scene structure. In their approach, every (sparse) point is assigned a discretised 'view sphere' whereby every solid angle bin saves the distance to the furthest camera within the solid angle that detected the point, plus the distance to the closest camera that did not detect the point. Using the view spheres primitives like planes and spheres are fitted to obtain probable locations of scene structure. However, results are shown for simple and well-textured scenes only, modelled by simple planes. Furthermore, since camera poses are used for the fitting, one needs to move extensively through the scene to provide the algorithm with enough camera poses. In contrary, the proposed method (Section 3) focusses not on the specific camera and point locations but on the rays in between, and thence indirectly finds low-textured and reflective objects using both visibility and occlusion information. In

addition, a probabilistic voxel grid is used for discretisation allowing more general shape reconstruction, the implementation is tested on complex real-world scenes, and compared to state-of-the-art Multi-View Stereo results.

2.4 Multi-View Stereo using Depth Maps

As a last category of geometry reconstruction methods, we will discuss reconstruction based on dense depth maps, often called stereoscopic. Usually, two steps are involved. In the first step, pairs or bigger sets of images are combined to form dense depth maps. The second step consists of carefully combining the depth maps in order to reconstruct scene geometry globally, for instance - as with SfM - by using bundle adjustment [30]. Where space carving is mostly defined in scene-space by re-projecting voxel centres onto image planes, multi-view stereo (MVS) using depth maps applies its photo-consistency measure in image-space by comparing pixels [26]. Furthermore, as do structure from motion techniques, MVS using depth maps tries to estimate the depth of pixels explicitly. Depth maps, however, try to estimate this depth for every pixel, giving a dense depth or disparity map. However, estimating the depth for every pixel makes it necessary to match with less confidence than possible with SfM points, which can rely on stable and easy comparable feature points such as SIFT. Finally, combining the dense depth maps results in a *dense point cloud*. We will now look into two different ways of creating the depth maps.

2.4.1 Active lighting

High precision in solving the depth ambiguity can be reached by actively illuminating part of the scene and searching for the illuminated part in images being taken. Depth map techniques recently received increased interest due to affordable active lighting sensors such as Kinect [13] reaching the commercial market, making depth vision suitable for a wider audience, but active lighting systems have been out there for a while. The illumination can consist of some fixed pattern (such as used by the Kinect and exploited by Izadi et al. [13]), a sequence of different patterns (structured light), or a single (laser) point or line that moves over the surface (either for triangulation or time-of-flight). For example, a laser range-finder can be used to accurately map an environment by sending bright laser pulses and measuring travel times. Huang et al. [12] released a database containing range images obtained this way, together with a statistical analysis of natural depths in images. Although pattern-based devices like Kinect are becoming affordable, their active lighting patterns are weak and hence only suitable for indoor scenes. In general, accurate active lighting devices are less affordable and less practical than omnipresent cheap cameras. Therefore, much research effort has been put into passive depth vision based on regular images pairs.

2.4.2 Passive depth vision

If one would know, for every pixel in an image, the particular pixel displaying the same point in scene-space in a second image, the relative displacement allows for calculating the distance between that point in space and the camera ([24], Ch.14). The depth is estimated by triangulation using displacement and relative camera transformation. Stereoscopic methods often attempt to find pixel pairs or matching patches by using a particular photo-consistency measure and exhaustive search. However, pixel colours

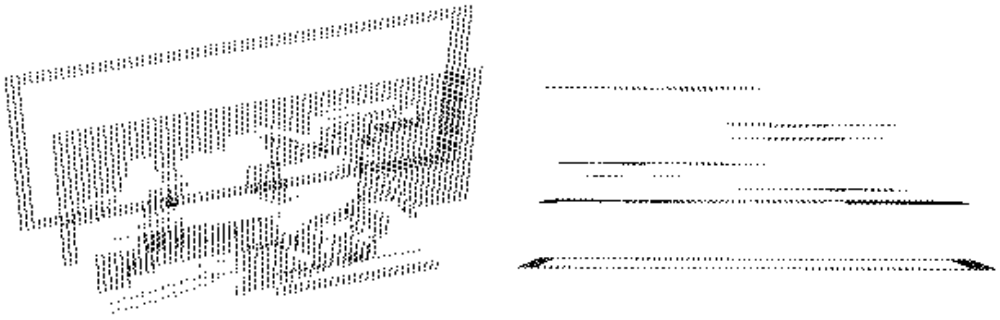


Figure 2.3: Two views of a depth map with few discretised depths, created by graph-cut algorithm. Image taken from own second years project, bachelor artificial intelligence, University of Amsterdam

are less unique than interest point descriptors and also in general not all the points are visible in both images (occlusion), resulting in ambiguous or missing matches, especially for scenes with low-textured surfaces (*e.g.,* painted walls). To simplify the search process, the images are often rectified: morphing them as if they were taken from cameras aligned on a common axis; the search for each pixel now only needs to occur on particular lines of pixels in the second image. Other geometry priors, such as order preservation or minimum and maximum depths, can be used too. Various techniques have been proposed for matching pixels or other primitives in order to obtain smooth depth maps without too many wrong matches or gaps. Hereby, almost all algorithms assume Lambertian surfaces as photo-consistency prior, causing problems for reflective surfaces.

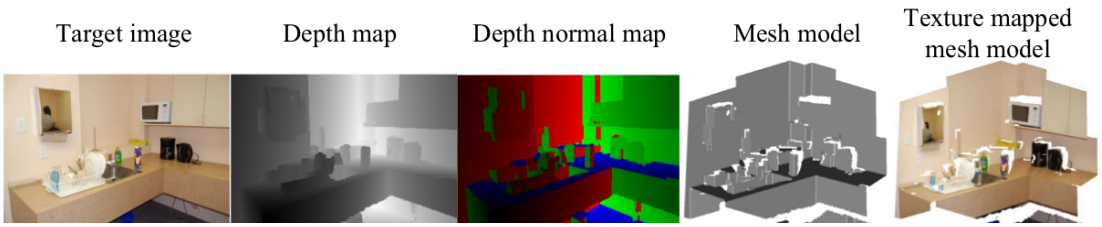


Figure 2.4: Pipeline of Furukawa et al. [7], including a typical depth map. Images taken from [7].

Most algorithms are based on matching pixels based on a photo-consistency measure [26]. After rectification, only pixels on a particular line need to be tried. Those algorithms often use a prior promoting local smoothness, such as a Markov Random Field (MRF), while using a graph-cut algorithm to find a solution. The number of possible depths can be kept low, lowering the number of computations and increasing smoothness while allowing jumps in depth (*e.g.,* Figure 2.3). Effectively, the image is segmented in layers for which depth is estimated. Plane-sweeping algorithms iterate between assigning pixels to planes, and refining the plane equations (which do not necessarily need to be parallel to the image plane). One fast but inaccurate plane-sweeping stereo algorithm is used by Merrell et al. [21] to quickly construct rough depth maps for further processing. Although the depth maps are imprecise, images can be processed very fast and depth maps are merged, obtaining reliable high-resolution dense point clouds. An example using an even stronger planar prior is the work by Furukawa et al. [7], who search for points with strong confidence depths and normals, and find three dominant axes that are more

or less perpendicular. Hypothesis planes are proposed based on the points with strong confidence, and a MRF is used to assign pixels to hypothesis planes, obtaining a planar (Manhattan) world consisting of the most confident planes. Example images from their pipeline are shown in Fig. 2.4.

Although planes are a useful prior in indoor scenes and scenes with a lot of human-made architecture, the smoothness prior is more useful in natural environments (*i.e.,* general outdoor scenes). The approach of Hernández et al. [11] uses photo-consistency in scene-space, whereby projection lines of pixels (*i.e.,* a set of possible depth locations) are projected onto camera planes nearby, and the depth candidates giving the best overall photo-consistency are picked. The estimated depths are then used in their carving algorithm using 3D graph-cut on a voxel grid.

Larger areas can be reconstructed using video and reliable location estimation (for example, using SLAM techniques). Pollefeys et al. [23] developed a system for real-time urban 3D reconstruction. It uses GPS and inertia sensors in addition to image-based pose estimation for reliable localisation, and uses a plane-sweeping dense stereo algorithm with a prior preferring locations containing points from the SfM sparse point cloud. Again, depth maps are created at a high rate and then merged into more accurate models. A similar large-scale reconstruction method is described by Frahm et al. [6]. Building on the success of the Photo Tourism system [28] they collect publicly available images on the internet; however, large amounts of images make extensive matching of image pairs intractable, therefore their system finds clusters of images with use of their short image descriptor before structure from motion is applied on each cluster individually. For each cluster, dense stereo is used and clusters are merged based on Geo-tag location and inter-connected clusters, finally obtaining a dense reconstruction for a big part of a city, within a day on a powerful PC.

Chapter 3

Method

3.1 Overview

Where space caving based on silhouettes (Structure from Silhouettes) has difficulties with high-textured environments and multi-view stereo techniques using depth maps have difficulties with low-textured or reflective environments, our method targets at scenes containing both. Specifically, it aims at reconstructing low-textured and reflective objects without relying on the ability to find *any* features on the objects themselves - provided enough texture-rich surrounding objects are present. To get statistical reliable reconstructions, enough data needs to be provided; therefore, the proposed method works best with video footage shot while moving through the scene. The method is based on space carving and iteratively improves a model represented with a voxel grid. Carving is accomplished by shooting finite rays between camera poses and locations of triangulated features, affecting voxels hit by the ray segments only. This is different from photo-consistent space carving, which affects single voxels each step, and differs from silhouette-based space carving in the sense that carving stops at a specific point, allowing concavities to be reconstructed. Furthermore, the voxel grid is probabilistic, allowing confidence estimation for reconstructed three-dimensional structures depending on the amount of data used for each voxel.

Two algorithms are proposed based on visibility and occlusion information. Central to the taken approach is the intuition that the ability to see a feature with known location means it is unlikely that there exists an objects in between the observer and the feature, and not seeing an earlier seen feature means there could well be such an occluding object. The algorithms require as input calibrated cameras, a reliable feature cloud, and visibility lists describing which cameras successfully detecting which points. The input can be provided by standard Structure from Motion techniques. Output of the algorithms is a probabilistic (or thresholded) voxel grid representing occupancy probability, and can be used for further surface reconstruction if desired.

The first algorithm, called **Visibility Space Carving**, assumes all space is occupied by default, unless proven otherwise. Space between camera poses and corresponding detected features is then carved, that is, marked as free. In case the features are points without clear size (such as obtained by standard SfM), in practise means carving tubes the size of voxels. Output are those voxels which do re-project in at least one image plane, but for which could not be proven to not occlude anything, either due an actual occluder object or due to absence of information.

The second algorithm, called **Visibility-Occlusion Space Carving**, starts with unknown occupancy state voxels. The occupancy probability is then altered by both visibility and occlusion information. Occupancy probability of voxels in space between camera poses and detected features is increased, and space between camera poses and undetected features is made more likely to be occupied. A slightly altered version is proposed for cases where visibility information is much more reliable than occlusion information, which can be caused by bad matches, as is often the case for publicly available SfM systems.

Collectively, the proposed algorithms could arguably be called **Structure from Visibility**. More detailed descriptions follow in Sections 3.2 and 3.3. For clarity reasons, a high-level overview of the complete pipeline from footage to reconstructed model is given next.

3.1.1 Reconstruction pipeline

1. **Shoot footage and extract images**

2. **Structure from Motion** (Section 2.3) - Stable features (*i.e.,* SIFT) are found, matched over the sequence of images, and relative camera transformations are estimated. The features are triangulated and the resulting 3D feature clouds are combined (*i.e.,* Bundle Adjustment) to obtain global camera poses, features poses and visibility information of the features for all camera poses. Existing tools can be used for this step, which we review in Sect. *implementation*.

3. Optional: **Extend visibility lists** (Section 3.4) - Many SfM systems are conservative with claiming visibility of features. It may help to review the visibility lists and attempting to extend them.

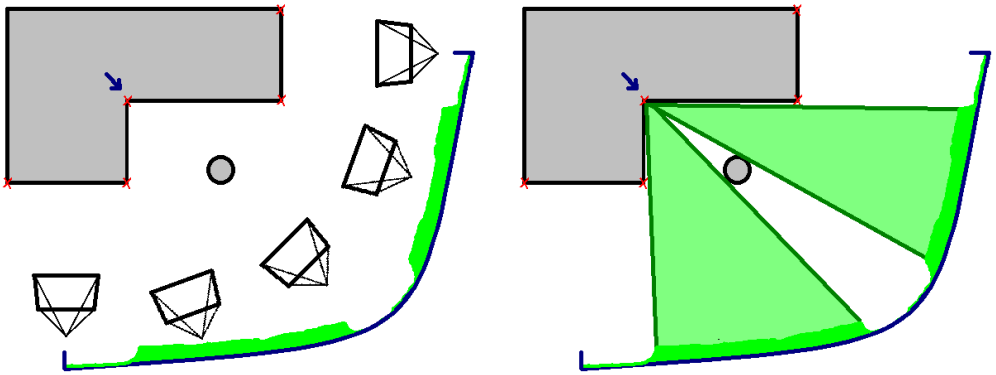
4. **Space carving** (Sections 3.2-3.3) - Space is partitioned into a voxel grid. Using the camera poses, feature locations, and visibility information, the voxel grid is carved, obtaining occupancy probabilities for all voxels.

5. **Regularisation** (Section 3.5) - The probabilistic voxel grid - containing independent voxels - is feed to a 3D graph cut algorithm to obtain a global solution where local smoothness is promoted.

6. Optional: **Surface reconstruction** - Raw voxel grids can be processed further to obtain alternative surface representations such as a polygon mesh (*i.e.,* using Marching Cubes) or fitted surface (*i.e.,* Poisson Surface), followed by texture mapping. Surface reconstruction is well-established and will not be covered in this thesis. Our final result will be a voxel grid.

7. **Visualisation** (Section 5) - Depending on the final representation, results can be displayed in a appropriate way. They can be rendered from original camera poses, or new positions. They can be rendered as 3D voxel grid, depth map, annotated on original images (original poses only) or texture-mapped (useful for new poses). Being able to interact with a 3D model usually greatly improves the ability to estimate geometry.

3.2 Visibility Space Carving



(a) Probabilistic visibility track / list (curved blue plot) over image sequence for one interest point (blue arrow)

(b) Carving away space based on one visibility track

Figure 3.1: Graphical example of Visibility Space Carving. Cameras in (a) represent a dense sequence of camera poses, hence the 'interpolation' in between.

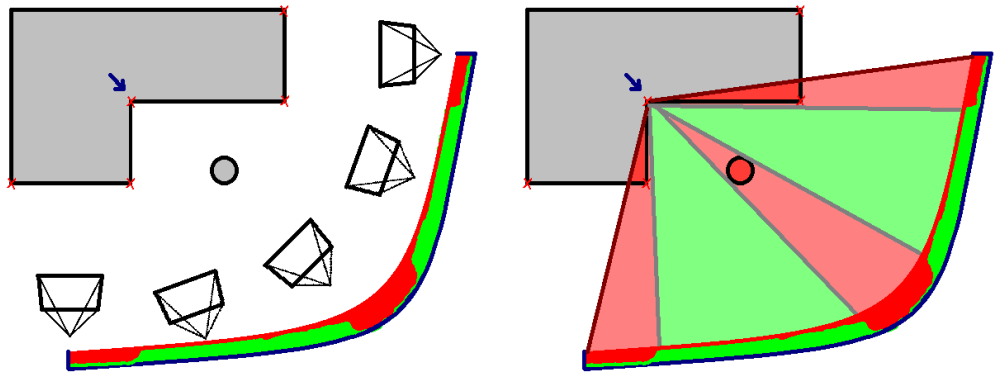
Algorithm 3.1 Visibility Space Carving

function VISIBILITYSPACECARVING(c, f, v, r) ▷ Arguments: cameras c, features f, visibility lists v, voxel grid resolution r
  $G \leftarrow G_{'unknown'}(r)$ ▷ Initialise voxel grid G with label 'unknown' and finite size
  for all Feature f_i in f **do**
    for all Camera c_j in v_{f_i} **do** ▷ Get all camera poses that detected this feature
      $X \leftarrow \text{getVoxelsBetween}(G, c_j, f_i)$ ▷ Get voxels in between camera and feature
      for all Voxel x_i in X **do**
        $x_i \leftarrow \text{free}$ ▷ Mark voxel as 'free'
      end for
    end for
  end for
  $H \leftarrow H_{'free'}(r)$ ▷ Initialise voxel grid H with label 'free' and same indexes as G
  for all Voxel G_i in G **do**
    $visible \leftarrow \text{false}$
    for all Camera c_j in c **do**
      if $\text{reprojectsInsideImage}(G_i, c_j)$ **then**
        $visible \leftarrow \text{true}$
      end if
    end for
    if $visible$ **then**
      $H_i \leftarrow G_i$ ▷ Only export the state of those voxels that could have been visible
    end if
  end for
  return H
end function

The Visibility Space Carving algorithm has its origins in Space Carving based on Silhouettes. It starts with a fully occluded world-view and carves away space that is not occluded. It does not carve space based on silhouettes until infinity, but carves until visible triangulated features (*i.e.,* SIFT interest points). Voxels between camera poses and features are found (*i.e.,* by shooting rays) and set to free. Since this is a binary process, the algorithm uses two labels instead of probabilities. The process for

one interest point is shown in Figure 3.1. Note that Structure from Motion can give an indication of the chance a feature is present (*i.e.,* how good the match is), but here we use thresholded, binary visibility lists and standard SfM tools often export only these. Also note that the example is shown in 2D and one interest point will only carve away one slice of space. If enough features are used, space will be carved away in a substantial part of the free space in the scene. After the carving step, a subspace of the grid is exported as output. Since voxels outside the camera views will always be occluded, we project the voxels onto the camera planes and export only those that project on at least one camera window. The algorithm is described more formally in Alg. 3.1.

3.3 Visibility-Occlusion Space Carving



(a) Probabilistic visibility track / list (curved blue plot) over image sequence for one interest point (blue arrow)

(b) Increasing (green) or decreasing (red) occupancy probability based on one visibility track. Space that is likely to get high occupancy probabilities after processing all data are coloured darker red.

Figure 3.2: Graphical example of Visibility-Occlusion Space Carving.

Assuming the world is occupied by default means that information is needed for every free part of space. In practise, footage often turns out to contain insufficient detected features to recover whole the scene. During initial tests with an implementation of Alg. 3.1 it became clear that reconstructed scenes typically contain a few regions which were free but for which there was simply insufficient positive information (*e.g.,* detected features on the background). Here, we present an algorithm that tries to overcome this problem. Regions for which no information is available are marked unknown, while negative information (*i.e.,* missing matches) needed to increase the probability of occluded regions. Two versions are presented: the general algorithm, and a slightly altered version for practical reasons.

The presented algorithm starts with a fully unknown world, with some prior probability value for occupancy. Similar to Alg. 3.1, voxels representing space between camera poses and features. However, both space towards detected and undetected features are affected. To this end, all features are projected onto all image planes. For the features that project inside the window borders of a particular camera (*i.e.,* all features that would have been visible if no occluders existed and matching was perfect), we check if they were detected or not. Features that were indeed detected contribute positive information, and so the occupancy probability of space between the particular camera and detected features is *decreased* by

a pre-defined constant factor Pr_{decr}. Features that did re-project inside the window borders, but were *not* detected, contribute negative information, and so the occupancy probability of space between the camera and undetected features is *increased* by a pre-defined (although potentially different) constant factor Pr_{incr}. The process for one interest point is shown in Figure 3.2. The algorithm returns the final probabilistic voxel grid. An occupancy threshold Thr_{occ} can be used to extract binary space partitioning where only voxels with high occupancy probability are set to 'occupied'. If desired, a second 'free-space' threshold Thr_{free} can be used for visualising voxels that are highly probable to be free. The general version of the algorithm is formalised in Alg. 3.2.

In practise, occlusion information is often less reliable than (positive) visibility information. Publicly available Structure from Motion solutions often act conservative while creating the visibility lists, and no probabilities are exported (Sect. 4). In the end, claimed visibility of features is very reliable, but claimed invisibility can well be a weak, but correct, match. To solve this balance equality, one could set the positive visibility information factor Pr_{decr} to a high value and set the occlusion information factor Pr_{incr} relatively low. Instead, we propose a modified version of the general algorithm whereby visible features cause a *veto* or absolute vote against occupancy of the space between camera pose and feature. Voxels within this space are set to Thr_{free}, which also prevents further increasing of occupancy probability. The modified version is shown in Alg. 3.3.

Note that all three algorithms require camera models c, features f, and (binary) visibility lists v, which are all provided by standard Structure from Motion techniques. The resolution of the voxel grid r needs to be set too. The two Visibility-Occlusion algorithms also need occupancy prior $Pr_{unknown}$ and probability increasing constant Pr_{incr}. The general version needs probability decreasing constant Pr_{decr} too, where the veto version needs the threshold for free space Thr_{free}. For visualisation purposes, an occupancy threshold Thr_{occ} can be used too. Appropriate values for these probability and threshold parameters need to be chosen. Parameters $Pr_{unknown}$, Thr_{free} and Thr_{occ} can be kept constant in most cases, and indeed that is the case in the rest of this thesis. However, Pr_{incr} and, if necessary, Pr_{decr}, are dependent on the data provided. Relative displacement between subsequent frames, amount of texture in the scene, stability of the features in the dataset, and voxel grid resolution r are all factors that can influence the choice of these parameters. In practise, often a few parameter values can be tried and results can be inspected visually.

3.4 Extending visibility lists

Since visibility lists provided by standard SfM tools are often conservative, it is worth exploring possibilities to extend those lists. This is especially true for the proposed Visibility-Occlusion Space Carving algorithms, which rely on missing entries in visibility lists. Early experiments using Alg. 3.2 and Alg. 3.3 indeed showed small uncarved regions caused by incorrectly missing matches. Therefore, a short excursion in extending the visibility lists has been made.

Missing matches are either caused by features being truly occluded, or bad similarity scores by

Algorithm 3.2 Visibility-Occlusion Space Carving - General formulation

function VISIBILITYOCCLUSIONSPACECARVING(c, f, v, r, $Pr_{unknown}$, Pr_{decr}, Pr_{incr})
  $G \leftarrow G_{Pr_{unknown}}(r)$ ▷ Initialise voxel grid G with prior of 'unknown' space
  for all Camera c_j in v_{f_i} **do**
    for all Feature f_i in f **do**
      if reprojectsInsideImage(f_i, c_j) **then** ▷ Get features projecting inside image borders
        $X \leftarrow$ getVoxelsBetween(G, c_j, f_i) ▷ Get voxels in between camera and feature
        if $c_j \in v_{f_i}$ **then** ▷ Feature is visible
          for all Voxel x_i in X **do**
            $x_i \leftarrow x_i - Pr_{decr}$ ▷ Thus, decrease occupancy probability
          end for
        else ▷ Feature was not detected
          for all Voxel x_i in X **do**
            $x_i \leftarrow x_i + Pr_{incr}$ ▷ Thus, increase occupancy probability
          end for
        end if
      end if
    end for
  end for
  return H
end function

Algorithm 3.3 Visibility-Occlusion Space Carving - Veto version

function VISIBILITYOCCLUSIONSPACECARVINGVETO(c, f, v, r, $Pr_{unknown}$, Pr_{incr}, Thr_{free})
  $G \leftarrow G_{Pr_{unknown}}(r)$ ▷ Initialise voxel grid G with prior of 'unknown' space
  for all Camera c_j in v_{f_i} **do**
    for all Feature f_i in f **do**
      if reprojectsInsideImage(f_i, c_j) **then** ▷ Get features projecting inside image borders
        $X \leftarrow$ getVoxelsBetween(G, c_j, f_i) ▷ Get voxels in between camera and feature
        if $c_j \in v_{f_i}$ **then** ▷ Feature is visible
          for all Voxel x_i in X **do**
            $x_i \leftarrow Thr_{free}$ ▷ Thus, use veto and set to free
          end for
        else ▷ Feature was not detected
          for all Voxel x_i in X **do**
            if $x_i >= Thr_{free}$ **then** ▷ Only if veto has not been used ...
              $x_i \leftarrow x_i + Pr_{incr}$ ▷ ... increase occupancy probability
            end if
          end for
        end if
      end if
    end for
  end for
  return H
end function

other reasons than invisibility. To discriminate between the two, we project the features, in our case SIFT interest points, into the cameras that did not detect them. Descriptions of the regions around the projected features are extracted, and compared to descriptions of the same points projected onto image planes of the nearest camera that did detect the particular feature. For simplicity, we use patch matching where patches are centred around the projected points, but other descriptors such as SIFT descriptors can be used too. The patches are projections of pieces of geometry with unknown orientation; however, since sequences are used and nearest cameras (in absolute distance) are searched for, it is reasonable to assume

that relative orientation with respect to the cameras would not differ too much, and thus projections would not show many differences, if indeed the same geometry is projected. Since we use SIFT features - which do not have a clear size when SfM tools do not export their scale - and we therefore space carve tubes the size of voxels, we chose the size of the patches to be equal to the size of the projected voxel containing the SIFT feature. The descriptors (*i.e.,* patches) are compared and thresholded, obtaining possibly new matches. New matches are added to the visibility lists. After extending visibility lists, we can apply one of the Visibility-Occupancy Space Carve algorithms described before. Since the assumption of conservative, thus reliable, visibility lists is now lost and some of the new matches will be incorrect, the veto version is less suited and the general formulation (Alg. 3.2) is preferred.

3.5 Regularisation

The voxel grids returned by the algorithms described all contain voxels that are individual solutions: their occupancy probability or state is independent of their neighbouring voxels' probabilities. To obtain a better global solution we can introduce a prior promoting local smoothness. Together with the proposed algorithms, they form joint voxel-carving algorithms. The voxel grid can be converted to a 3D graph which can be solved by standard min-cut/max-flow graph-cut methods [1]. Voxels are thereby represented as nodes and nodes for neighbouring voxels are connected. Nodes have unary costs and pairwise costs. The unary costs U_i for connection with the Sink node are set equal to occupancy probabilities (after truncating to the range $0.0 - 1.0$), and connection with the Source is set to $1 - U_i$. Pairwise costs could have been based on some photo-consistency comparing appearances of pairs of re-projected voxels, but this would introduce dependencies on appearances again, which were tried to be dropped in the first place. Therefore, pairwise costs are based on differences in occupancy probability weighted with a constant factor, as $P_{i,j} = \gamma * (1 - |U_i - U_j|)$.

Chapter 4

Implementation

In this chapter, details on implementation of the proposed method are discussed. Code is listed in Appendix D, supplied on DVD, and made available online[1]; this chapter will complement the code and will serve as a general introduction, therefore no code is listed in this chapter. The pipeline given in Section 3.1.1 will be used as a guideline, allowing the reader to follow the process in execution order. This order, though, is not necessarily the order of implementation; in fact, most applications developed for visualisation (last step in the pipeline) were written in an early stage, allowing for following progress and viewing intermediate steps. Developing geometry reconstruction algorithms greatly benefits from the ability to quickly shoot footage, run suggested and implemented algorithms, *visualise* intermediate steps, and suggesting modified algorithms. Details on collecting datasets (first step in the pipeline) are discussed in the next chapter; implementation details of the remaining steps in the pipeline are discussed after an overview of the libraries that are being used.

4.1 Libraries

All tools are written using platform-independent and open-source libraries, and all code should compile and run on all major platforms (although extensive tests have only been done under Linux, Ubuntu 12.04). The same holds for all tested external tools (SfM step), except Voodoo which is only free for research-purposes. All tools developed for this thesis are written in C++ and a cmake config file is provided.

The well-established computer vision library **OpenCV**[2] Bradski [2] is used for general image and video i/o, and 2D image and feature visualisation. Very recently, a sister project called **Point Cloud Library**[3] (PCL) was announced by Rusu and Cousins [25], providing common data structures, algorithms, and tools useful for 3D computer vision. Although PCL still lacks extensive documentation and has a rapidly changing API, it does provide useful data structures and tools for handling point clouds, including advanced filter, segmentation and surface algorithms. Regrettably, it has (yet) no data structures for visibility information nor a commonly-agreed camera representation, reflecting the rare interest in visibility and occlusion (Sect. 2.3), so they have been developed for this thesis. In addition, the

[1] Code available on Google Code: http://code.google.com/p/martijn-msc-thesis
[2] OpenCV homepage: http://opencv.willowgarage.com
[3] PCL homepage: http://pointclouds.org

provided voxel grid representations are mostly useful for space partitioning and searching. Instead, we used the more suitable and feature-rich library **OctoMap**[4], recently released by Wurm et al. [15]. OctoMap was chosen because it makes use of probabilistic nodes, implements a memory-efficient octree instead of same-size voxels, and allows for ray shooting. It also includes an octree viewer. Lastly, a set of open-source C++ files released by Boykov and Kolmogorov [1] has been used for efficient graph-cut regularisation. Needless to say, writing methods to convert between representations of these four libraries was inevitable.

4.2 Pipeline implementation

4.2.1 Structure from Motion

Camera models and a feature point cloud can be obtained by standard Structure from Motion algorithms (Sect. 2.3). Structure from Motion has been implemented oftentimes already. Nowadays, reliable software tools are available for processing image sequences and outputting camera poses and point clouds. Both commercial and open-source tools exist. Since the proposed algorithms use SfM output as is and do not innovate on the SfM process itself, we experimented with three freely available Structure from Motion tools instead of building our own. The tools tested are Bundler, Voodoo, and VisualSfM. All tools are used with default settings.

Originating as part of the Photo Tourism work [28] and released under the GPL open-source license, **Bundler**[5] became an authority and is still one of the most cited SfM systems. Originally developed for unordered photo collections taken from the internet, it claims to work on normal sequences too. The command line tool takes a directory with images as input, extracts and matches SIFT features and descriptors by default, and optimises estimated parameters incrementally using sparse bundle adjustment. Bundler outputs an ASCII file containing camera models, triangulated points with colours and visibility lists. The latest binary version (0.3) was tested.

Voodoo[6] was developed as non-commercial alternative to software tools such as Boujou and is free to use for research purposes. Finding and matching SIFT descriptors is possible, although by default Harris edge detection is used and points are tracked over the sequence. Voodoo is developed for sequences only for which the small displacement assumption is reasonable. The GUI allows for manually fine-tuning of estimated feature tracks (*e.g.,* linking a lost feature point to its rediscovery a few frames later) and includes simple modelling tools for improving the results, but those are not used during testing. Voodoo outputs various ASCII file formats, but unfortunately none of them includes visibility information. The latest version (1.2.0 beta) was tested.

Recently, a new graphical tool called **VisualSfM**[7] by Wu et al. [30] was released that constitutes a graphical shell around a GPU implementation of both SIFT (SiftGPU) and Bundle Adjustment by Wu et al. [30] for SfM, and the CMVS/PMVS tools by Furukawa and Ponce [8] for patch-based dense multi-view stereo. VisualSfM and its components are all released under the GPL license. VisualSfM

[4] OctoMap homepage: http://octomap.sourceforge.net
[5] Bundler homepage: http://phototour.cs.washington.edu/bundler
[6] Voodoo homepage: http://www.digilab.uni-hannover.de/docs/manual.html
[7] VisualSfM homepage: http://www.cs.washington.edu/homes/ccwu/vsfm/

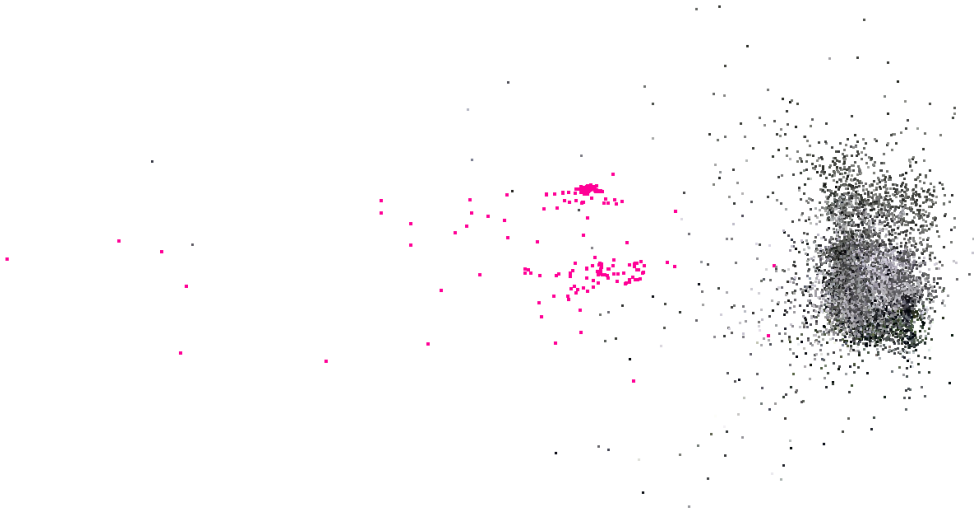
outputs an ASCII file similar to Bundler's format, and a binary file after the optional dense reconstruction step. The interface offers visual feedback during incremental bundle adjustment, and provides quite a few other useful visualisations. The latest version (0.5.17) was tested.

Example outputs are shown for two datasets; one example frame for each is pictured in Fig. 4.1). The chess dataset gives typical results and are shown in Figure 4.2; good results are obtained for the houses1 dataset, shown in Figure 4.3. More details on all datasets used in this thesis are listed in Appendix C.

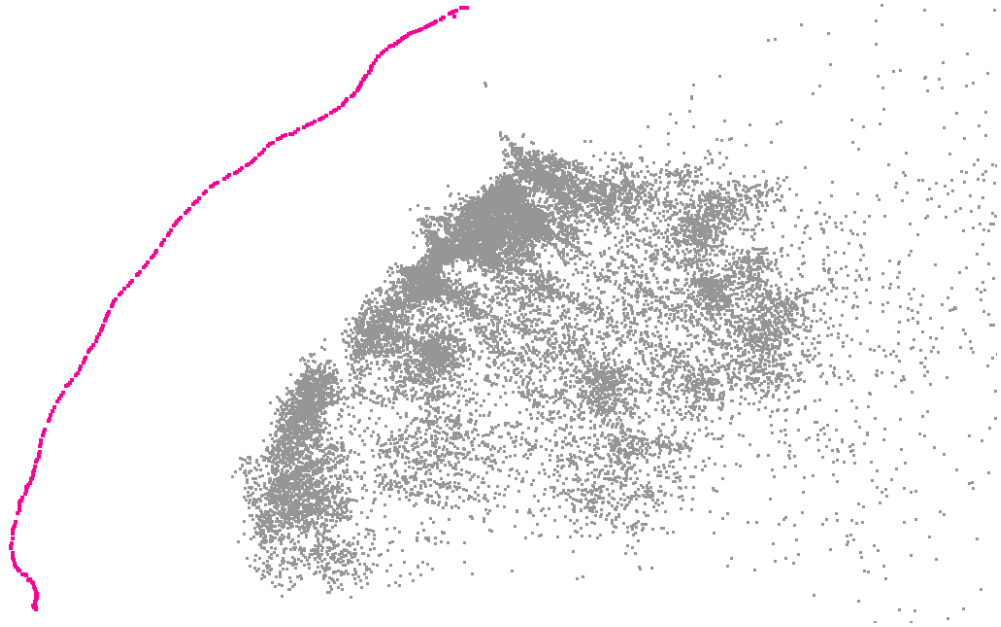
In general, Bundler performs poorly on the supplied datasets. Output usually consists of a point cloud without identifiable structures. An explanation for the bad results can lie in the fact that Bundler is developed for unordered sequences and therefore uses no prior for camera displacement nor for possible identical camera models. Indeed, estimated camera locations often have a variety of distances (and intrinsics) from the point cloud, where the other two tools place cameras in a string. Voodoo often exports reasonable results with, indeed, trustworthy looking camera paths and point clouds. It does not export colours and does not automatically rediscover points lost during tracking earlier on in the sequence, and no bundle adjustment is used to improve results. Using SIFT matching does not improve results, and matching only occurs between nearby frames. Voodoo results often contain a fair amount of points triangulated very far away from the scene. More importantly, none of the variety of export formats contains visibility lists. VisualSfM gives good results on almost all tested sequences. Due to its GPU implementation, it is faster than Bundler (and Voodoo in SIFT matching mode): two hours on sequences of 500 frames against half a night. Visual feedback during reconstruction and bundle adjustment is useful for monitoring progress. Due to Bundler's bad results, Voodoo's lack of visibility lists, and VisualSfM's good performance over all, VisualSfM is used for all further experiments.

(a) Frame 70 of chess dataset (b) Frame 11 of houses1 dataset

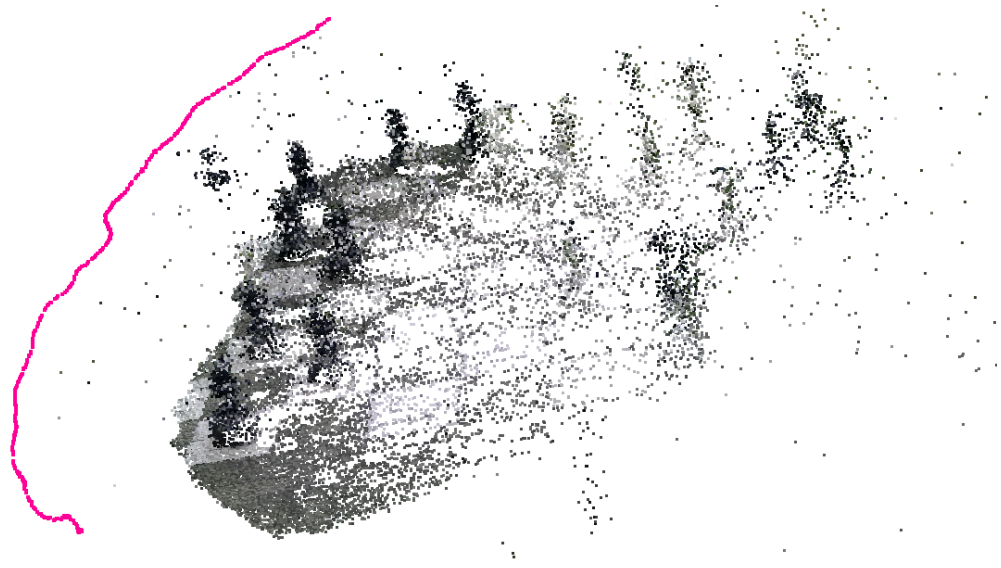
Figure 4.1: Example frames for chess and houses1 dataset (used for SfM tools comparison).



(a) Bundler output



(b) Voodoo output

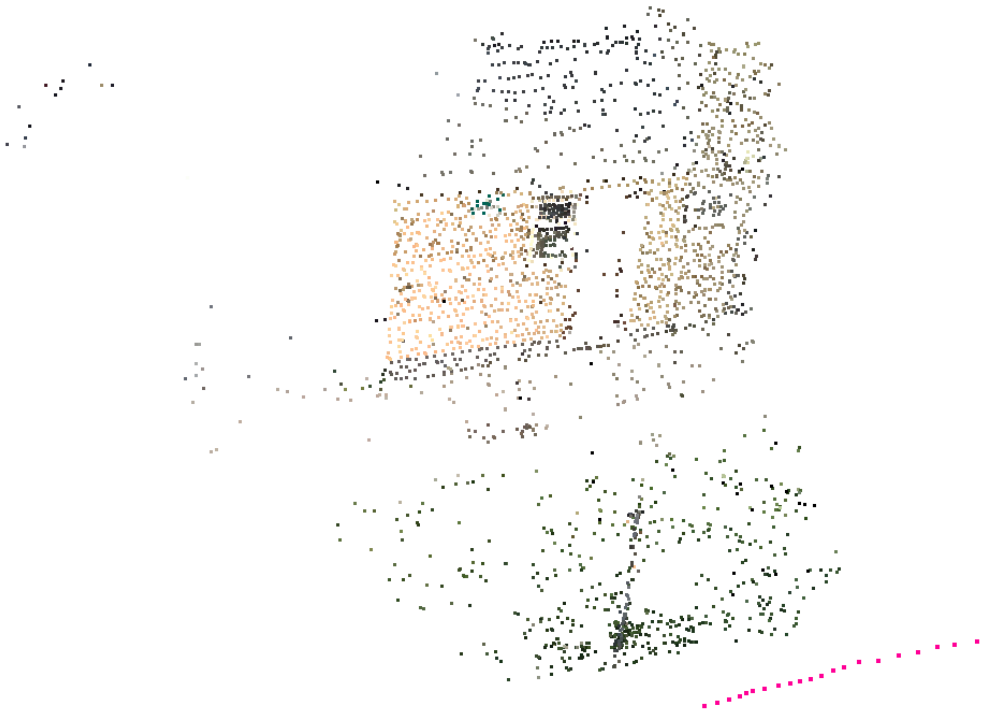


(c) VisualSfM output

Figure 4.2: Comparison of Structure from Motion tools Bundler, Voodoo and VisualSfM; typical example (chess dataset). Points shown in estimated colour (or grey if not given), camera poses displayed in pink. Visualisations by `sfmviewer`.

(a) Bundler output

(b) Voodoo output



(c) VisualSfM output

Figure 4.3: Comparison of Structure from Motion tools Bundler, Voodoo and VisualSfM. Good example (houses1 dataset). Points shown in estimated colour (or grey if not given), camera poses displayed in pink. Visualisations by `sfmviewer`.

4.2.1.1 Reading SfM output

A common interface class `sfm_reader` has been written for reading the various different formats used by different SfM tools. Furthermore, data structures from the C++ standard library are used to represent visibility information. Another reason to discuss this particular class is that it also provides useful methods for the carving algorithm concerning visibility and occlusion.

Four file formats parsers are written. The ply format (Stanford) is shared among SfM tools; however, it only contains points or polygon meshes. Ply files consist of a table (therefore, no variable length lists can be represented) and data can be binary or ASCII. The binary variant can be read by the PCL library, but for the ASCII data variant a parser has been written. Bundler and VisualSfM can both output the ASCII ply format, and VisualSfM outputs the binary ply format after dense reconstruction with CMVS/PMVS. Voodoo can output scripts for various programs (such as Blender). Its native output format has the common extension .txt and contains a list of points and a list of camera models. Bundler's native output format has extension .out and contains camera models, points, and a visibility list for each point. VisualSfM's native output format has extension .nvm and contains the same information as Bundler's, although it contains a different camera model using quaternions instead of R, t matrices, and distortion has one parameter instead of two. For all three native formats a parser have been written.

Points are saved in a `PointCloud` object as provided by PCL. It uses `PointXYZRGB` points which contain a location and an RGB colour. Cameras are saved in two separate structures: camera poses are saved in another `PointCloud` for easy visualisation - no pyramid-like camera visualisation wrappers are provided by PCL - as well as in a separate list of `camera` structures. The user-defined `camera` structure contains intrinsics R and t, focal length, and two radial distortion parameters. All camera models from the different file formats are converted to this structure (*e.g.,* quaternion to R matrix). Each visibility list is saved in a `std::map` object mapping every frame index for which a point was detected to a `visibility` structure. This user-defined structure contains the index of the feature for the particular frame, and the x and y coordinate of detection in the frame. Every feature point now has a global (x, y, z) position in the `PointCloud` object, and a list of local (x, y) camera coordinates for those cameras that detected the feature point. The map object allows for fast checks on visibility given a frame and feature point number, even for larger sets of frames.

Class `sfm_reader` provides re-projection methods for given feature point and camera, including re-distortion. The method `reprojectsInsideImage` returns a boolean indicating whether a given point re-projects inside a camera window or not. This method is also used by `selectPointsForCamera`, which re-projects all points inside a given camera and makes an occlusion list of those points that could have been visible, but have not been detected (*i.e.,* the camera is not present in the visibility list). A list of currently visible points is made as well.

4.2.2 Extending visibility lists

The optional step of extending visibility lists checks all point-camera pairs that are not represented yet in the vector of maps of visibility information. In practise, this is implemented by looping through the points and, for each point, creating a visibility and invisibility list with help of the re-projection

functions. Then, the whole sequence of a particular point is followed and all invisibility entries (re-projects inside image but not detected) are checked. The check consists of finding the nearest camera in absolute distance that did detect the feature point, re-projecting the point into that image and the current image, and comparing patches around the points. The size of the patches is determined by translating the feature point half the size of a voxel (given by the octree) in a direction perpendicular to the vector between the point and camera pose, re-projecting the translated point into one of the images, and calculating the distance to the projection of the original point. This distance is an approximation of the size of a projected voxel near the feature point, and is used to extract rectangular patches from the two images. They are compared by calculating the mean squared distance. A value below the threshold causes the camera index to be added to the visibility list of the feature point. By manually inspecting patch pairs and mean squared distance values, the threshold was set to 0.2 for all further experiments.

4.2.3 Carving

The proposed carving algorithms are implemented using the probabilistic octree of the OctoMap library [15]. The octree is initialised with a single parameter specifying the smallest node size possible. By default, all ray shooting operations are executed on nodes at this smallest scale, that is, the leaf nodes in the tree representation. Various node types are provided, the one containing probabilities is used. Probabilities are stored logarithmically, but conversion to and from normal probability values is straightforward. Two thresholds are used to allow nodes to be labelled with either 'free' ($val <= Thr_{free}$) or 'occupied' ($val >= Thr_{occ}$), or no label (unknown). Those thresholds are $Thr_{free} = 0.2$ and $Thr_{occ} = 0.7$ by default, and are kept at these values for all experiments. For memory efficiency, eight (leaf) nodes can be merged recursively if they have the same label, thereby averaging and thus losing individual occupancy probabilities. By default leaf nodes are not initialised and depth of the tree is kept to a minimum.

For the implemented space carving algorithms, the octree is initialised with smallest node size equal to the largest axis variation in the feature point `PointCloud` divided by a given resolution parameter. OctoMap provides two ray shooting methods: `computeRayKeys` takes as input two points and returns a list of node keys that were hit by the ray between the two points; `insertRay` does the same, but sets the last node to Thr_{occ} and all the others to Thr_{free} instead of returning the list. The latter is used for the binary Visibility Space Carving algorithm (Alg. 3.1), and the former is used for both probabilistic Visibility-Occlusion Space Carving algorithms (Alg. 3.2 and 3.3), there where `getVoxelsBetween` is mentioned. Note that since we carve tubes of space in this implementation, the resolution parameter potentially has a big influence on the result. In practise, there is a range of possible resolution values that give good results, and the range is laid down by the dataset provided.

4.2.4 Regularisation

Regularisation is implemented used the code provided by Boykov and Kolmogorov [1]. For simplicity, the octree is converted to a voxel grid with voxels the size of the smallest possible octree node. First an empty voxel grid is initialised using the occupancy prior for unknown space, which we set to $Thr_{unknown} = 0.2$. Since conversion splits existing bigger nodes and adds nodes where nothing was

initialised before, this increases memory usage quite a lot. This means that only lower resolution settings can be used in the regularisation process. After applying the graph-cut algorithm the voxel grid is converted back to the memory-efficient octree representation.

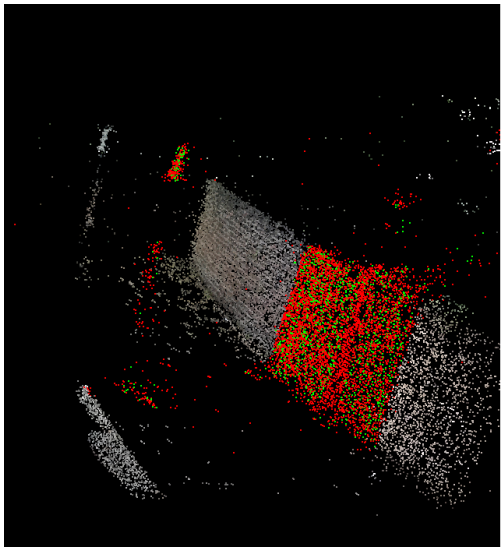
4.2.5 Visualisation

During the pipeline, a few intermediate and final results can be visualised. Three visualisation tools are developed and one provided by the OctoMap library is used. We will discuss each briefly. The first two tools are used for visualising features before carving, one 2D image annotator and one interactive 3D tool. The two other tools are meant for visualising voxel grids after carving, again one 2D image annotator and one interactive 3D tool.

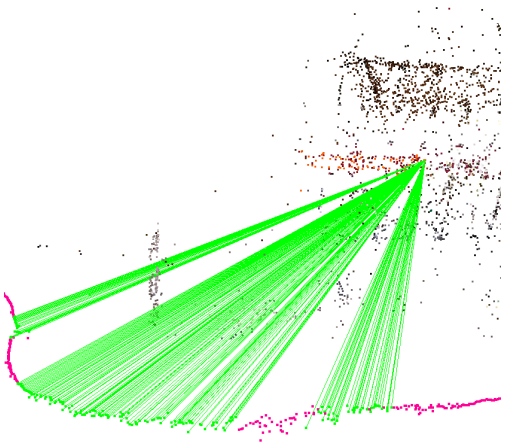
Although external structure from motion tools are being used, a tool has been written for experiments with and visualisations of various feature detectors. It can be used to check the amount of interest points discoverable in a dataset, and the uniqueness of their descriptors during tracking. It is not used in the final pipeline, but is discussed for completeness. The visualiser, `featureviewer`, takes as input either a video, directory with images, or webcam stream, and annotates the images with a given type of interest point (*e.g.,* SIFT, SURF, ORB, FAST, HARRIS) including their size, plus edges (Canny edge-detector). One can interactively click on a feature point, which will be matched over the sequence by a simple ratio test: if the ratio of the two best matches in the next or previous frame (*i.e.,* the L1 distance) is big, the interest point is quite unique and the best match is taken. If the ratio is close to 1, we keep the last known good descriptor and continue to the next frame. Note that this simple heuristic does not use any location prior. We can define the visibility confidence as one minus that ratio, and plot it for the sequence, similar to the visibility tracks shown in Fig. 3.1 and 3.2. One example image processed with `featureviewer` is shown in Figure 4.4(a).

Visualisation of triangulated SfM feature points or dense point clouds can be done with `sfmviewer`. The SfM tools outputs (Fig. 4.2 and 4.3) were made this way. It uses class `sfm_reader` for opening SfM output files and representing their contents. The PCL library provides wrappers around the VTK visualisation toolkit. It is used for easy point cloud visualisation, both the feature point cloud and camera poses. Interaction is implemented using the default PCL visualiser for mouse navigation, plus VTK call-back functions for custom shortcuts, and selection of points and cameras. Selecting a point colours the cameras that detected the point (using the visibility list of the point). Selecting a camera calls `selectPointsForCamera` as described earlier, thereby projecting all the points into the camera. Detected points are coloured green, while undetected points that do re-project within the image borders but are coloured red. To be able to recover the original point colours during the next selection, `sfm_reader` keeps a shadow copy of the original point cloud. Another implemented feature is the ability to visualise those visible (green) and invisible (red) points onto the original corresponding frame. The camera-feature pairs can be visualised even more explicitly by drawing green lines between the camera and visible points or between the point and detecting cameras. Two example visualisation is shown in Fig. 4.4(b) and 4.4(c).

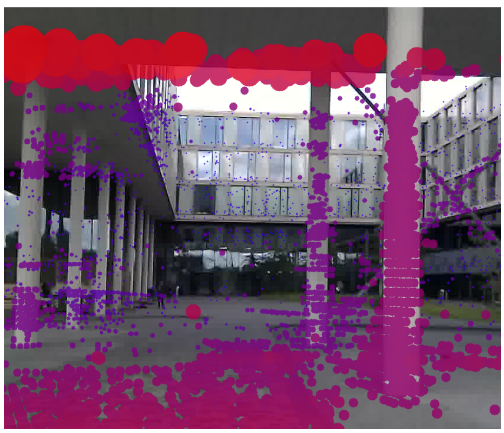
(a) `featureviewer` for frame in chess, including SIFT features and visibility track for the green interest point



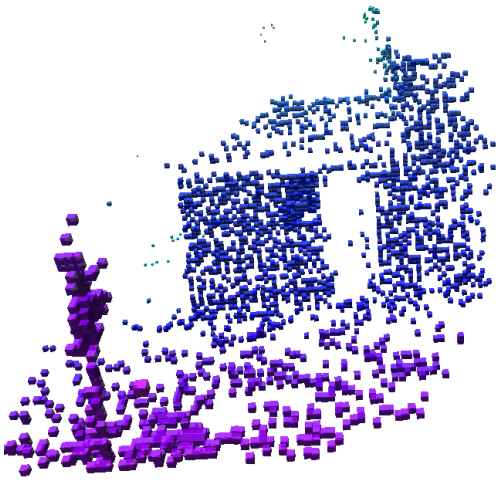
(b) `sfmviewer` for selected camera of lamp-posts_on_wall1; one lamppost causes a red area to emerge on the wall, but the conservative visibility information of VisualSfM is noticeable too in the 'green' areas



(c) `sfmviewer` for selected point of sainsburys1; two reflective poles become 'visible' due to absence of lines



(d) `carveviewer` for frame in sciencepark2 (with incorrect nearby voxels at the top and only few detected features)



(e) `octovis` view for houses1; a partly reconstructed lamppost is visible on the left (purple)

Figure 4.4: Visualisations by three developed tools and one provided viewer

After space carving has been done, a probabilistic octree is the result. For visualisation purposes, we use the occupancy threshold Thr_{occ} to make a binary distinction between occupied and not occupied. The last developed visualisation tool, `carveviewer`, annotates the original images with the carved octree. To prevent axis dependent results, all nodes are imagined as spheres and projected on top of the images. As with the code for extending visibility lists, both the node centres and centres with offsets the size of the nodes are projected into the images, and the projected points are used to estimate the radii, resulting in a list of x, y, r entries. Then, rays are casted from the camera in the direction of the projected nodes, and all rays hitting another, occluding node before the projected one, are removed from the list. The remaining nodes are visible from the given camera and are drawn as semi-transparent circles on top of the image. The circles are drawn with the estimated radii and coloured according to the log distance from the camera (nearby voxels are red, voxels far away blue). One example rendering is shown in Figure 4.4(d). For more dense voxel grids, results remind of depth maps on top of their original images.

An interactive octree viewer is provided by the OctoMap library. It has been used without modification. Nodes are rendered as semi-transparent cubes which are either light blue for occupied and dark blue for very high occupancy probability (threshold undefined), or coloured according to the value of one of the axes. Optionally, free space (occupancy probability below Thr_{free}) can be visualised too as green cubes. One example view is shown in Figure 4.4(e).

The visualisation tools will be used extensively in the next chapter.

Chapter 5

Results and Evaluation

5.1 Data capturing

The proposed method was designed for scenes containing objects occluding other, rich-textured objects. While alternative methods work best for either low-textured objects (silhouette space carving) or high-textured objects (MVS with depth maps), the proposed Structure from Visibility alternative is expected to reconstruct scenes containing both. Specifically, the proposed method has no requirements with regard to the appearance of occluder objects in scenes containing other sufficiently textured objects. Therefore, it should work on complex coloured scenes containing low-textured or even reflective objects, whele existing techniques will have difficulties reconstructing those kind of objects. Unfortunately, datasets containing those kind of scenes are rare, and none are known by the authors that contain ground truth models. For example, the Middlebury Multi-View Stereo dataset[1] [26] does provide ground truths, but only contains statuettes made of approximately Lambertian objects, and background-foreground segmentation is easy. Therefore, we collected our own data and evaluate the proposed and implemented system visually by comparison with state-of-the-art.

Outdoor scenes containing low-textured and reflective objects were captured using two different cameras: a Google Nexus One phone-camera and a Sanyo HD camcorder. In filming mode, the phone-camera gives surprisingly low-quality looking 720x1280 resolution images; the HD camcorder was set to 1080x1920 and all other settings were set manually to obtain clear footage for every scene. Datasets are listed in Appendix C; datasets and results are supplied on DVD and made available online[2] in various native output formats and as screencasts. Here, we show two typical results (Figures 5.1 and 5.3) and two good results (Figures 5.5 and 5.7. As mentioned in Section 1, the interested reader is encouraged to view the supplied three-dimensional results in addition to the 2D snapshots shown here.

Each result consists of one example frame, the sparse point cloud as reconstructed by VisualSfM (`sfmviewer`), the same sparse point cloud discretised as octree (for comparison), result of Visibility Space Carving (Alg. 3.1) and Visibility-Occlusion Space Carving Veto version (Alg. 3.3) as 3D model (`octovis`) and occasionally as annotated image too (`carveviewer`), and the result of the patch-match based dense reconstruction algorithm CMVS/PMVS by Furukawa and Ponce [8] from 2010

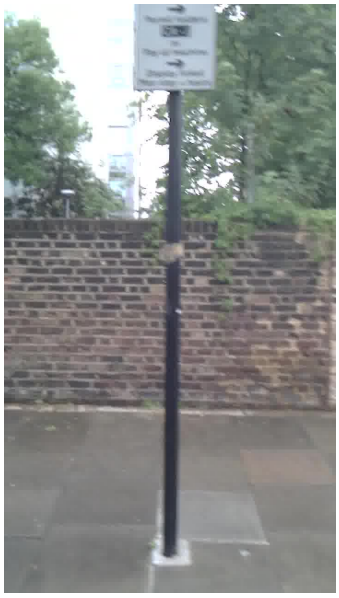
[1] http://vision.middlebury.edu/mview/

[2] Link to datasets available at: http://code.google.com/p/martijn-msc-thesis

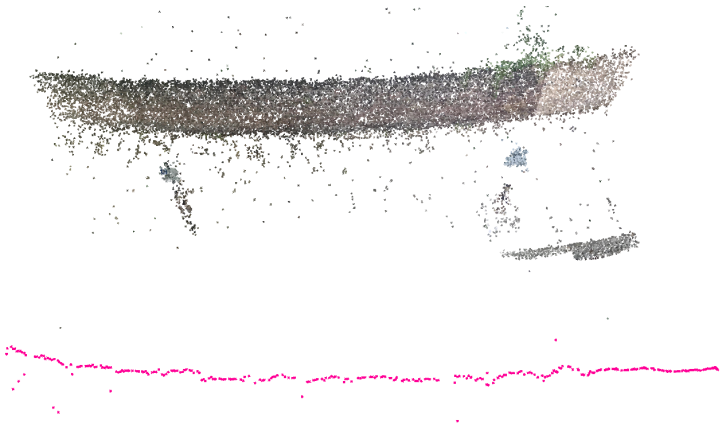
(`sfmviewer`). For the Visibility-Occlusion algorithm, a few values for Pr_{incr} are tried and the best result is shown. Note that our space carving pipeline currently does not include texture mapping; the results will be judged based on shape rather than photo-realistic rendering.

Typical run times (on an AMD Phenom X4 quad-core 3.2 GHz) for sequences of around 400 images are 1.5 hours for structure from motion (VisualSfM), an additional 1.5 hours for dense reconstruction, or 10 minutes for carving with $r = 2500$ (1 minute with $r = 250$).

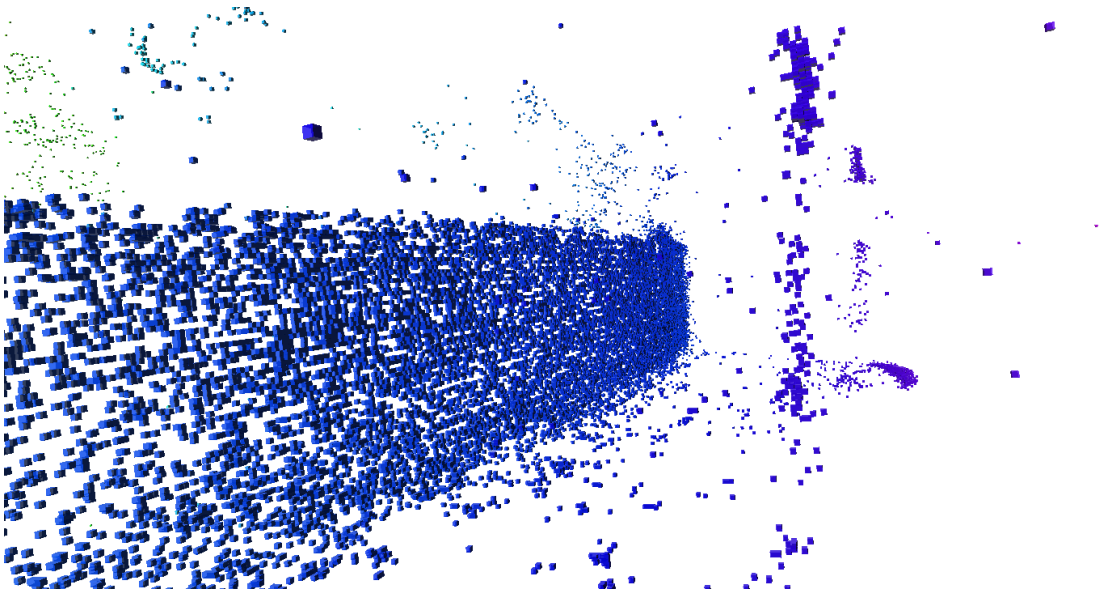
5.2 Visual results



(a) Example frame; notice the poor image quality



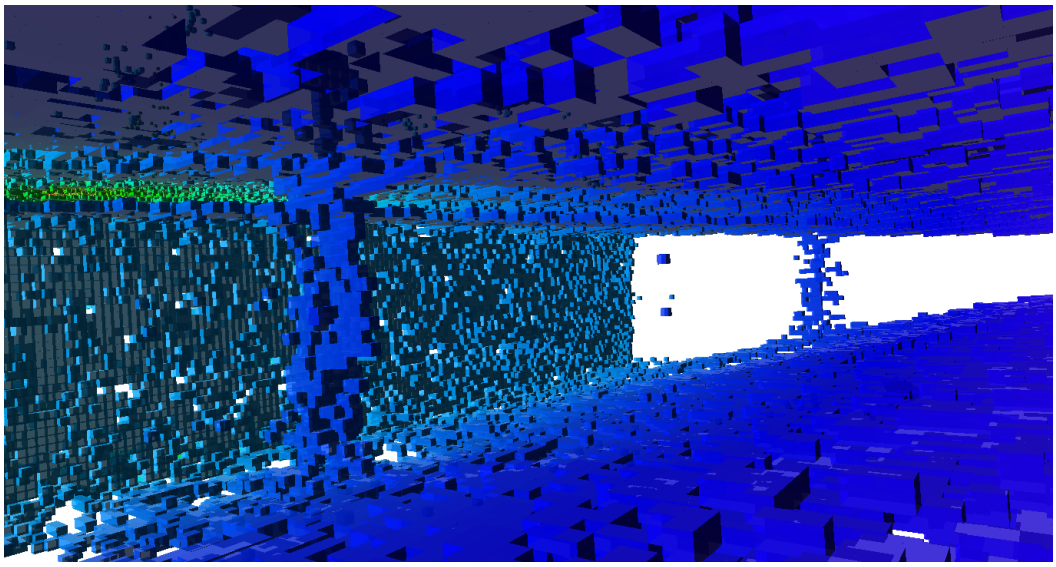
(b) Sparse point cloud and camera poses



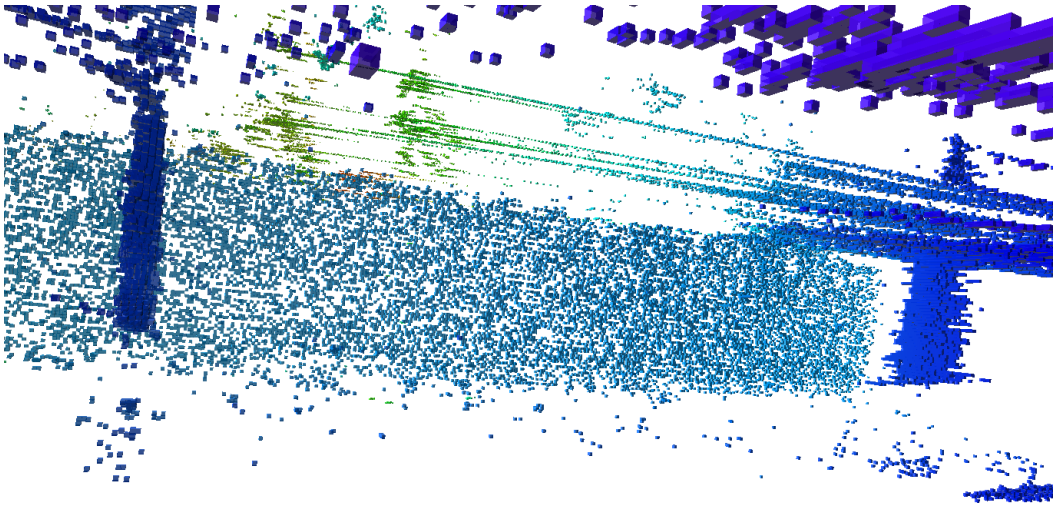
(c) Discretised sparse point cloud ($r = 1000$)

Figure 5.1: Typical result 2: lampposts_on_wall1 dataset

5.3 Evaluation



(a) Result of our Visibility Space Carving (Alg. 3.1) with $r = 500$; Notice the occupied-labelled space above and underneath the carved space, caused by an insufficient amount of feature points in the trees above the wall and on the ground.

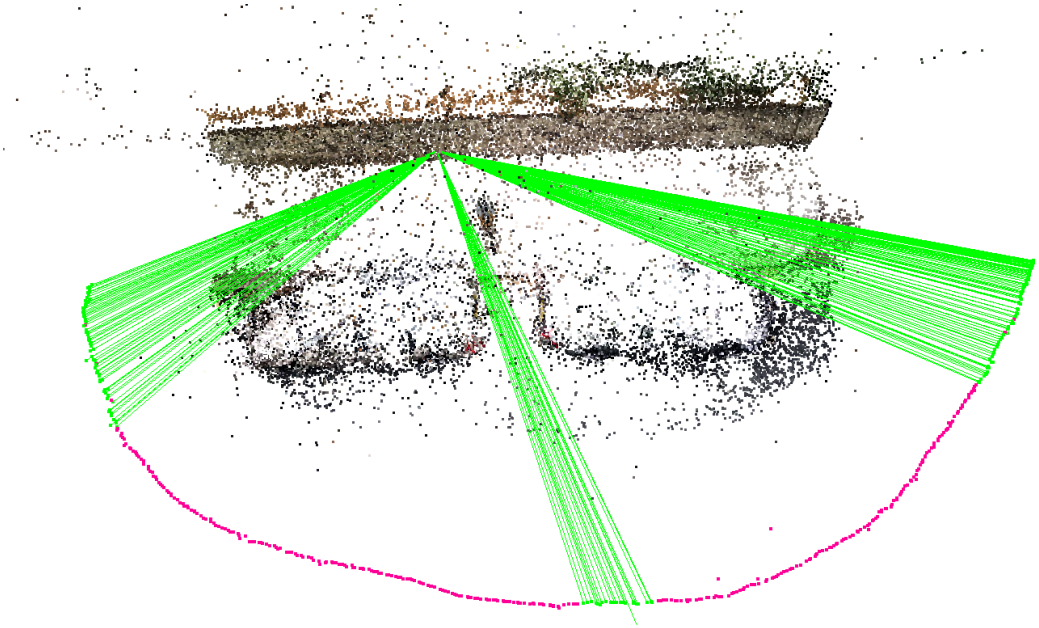


(b) Result of our Visibility-Occlusion Space Carving Veto (Alg. 3.3) with $r = 1000$, $Pr_{incr} = 0.005$

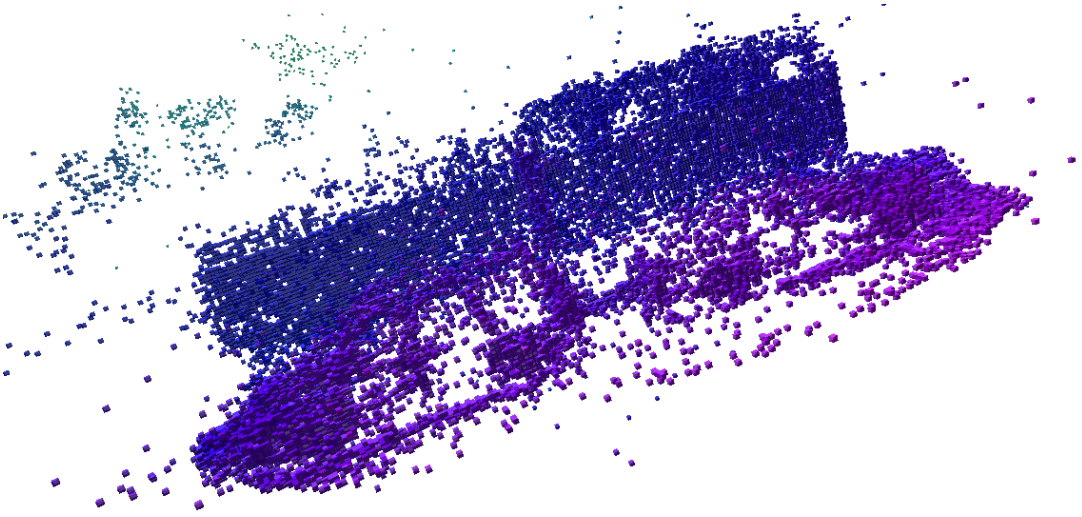
(c) Result of CMVS/PMVS [8]

Figure 5.2: Typical result 2: lampposts_on_wall1 dataset (cont.)

(a) Example frame

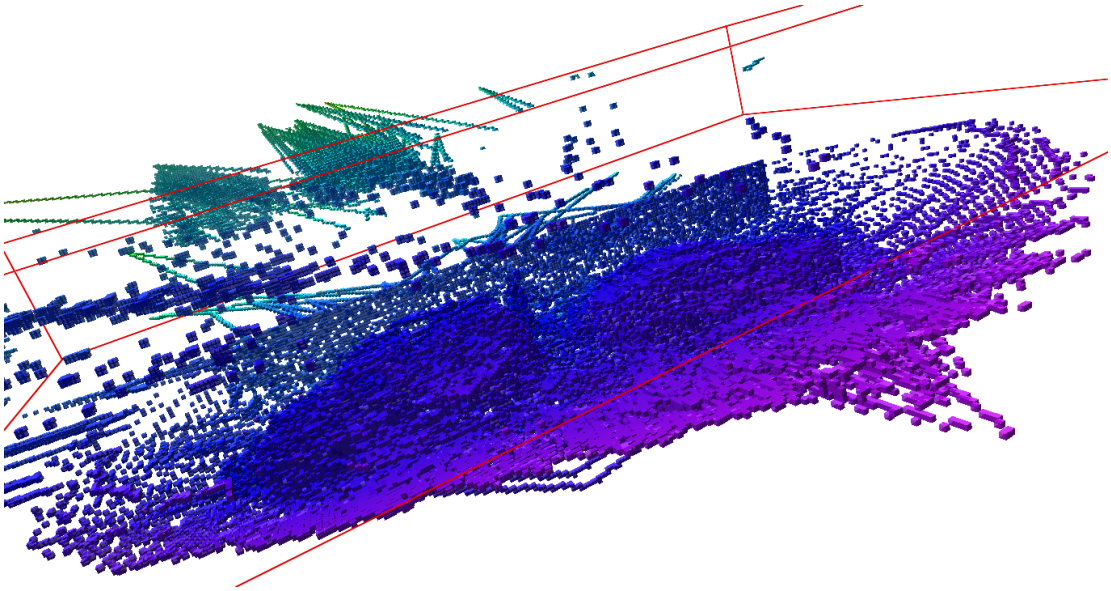


(b) Sparse point cloud and camera poses, annotated with visibility lines from one point (hinting at possible occluding object locations)

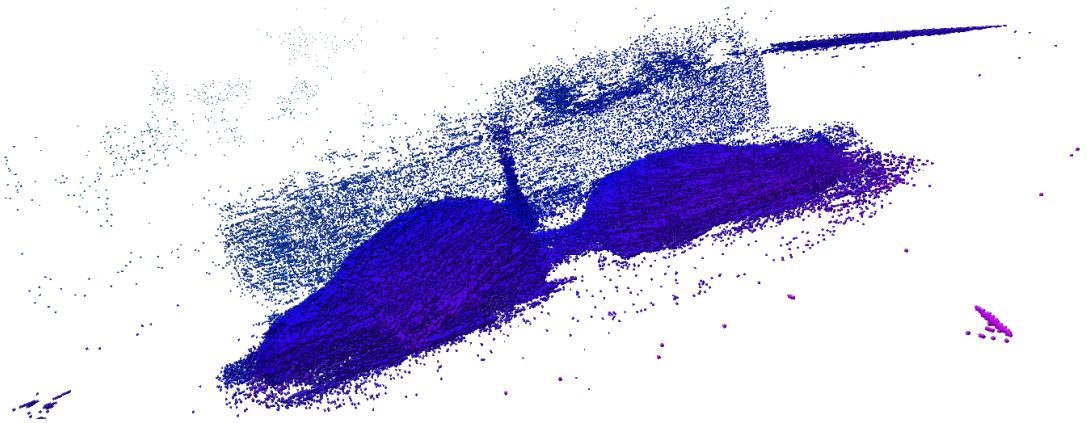


(c) Discretised sparse point cloud ($r = 1000$)

Figure 5.3: Typical result 2: car_and_wall1 dataset



(a) Result of our Visibility Space Carving (Alg. 3.1) with $r = 1000$; Incorrectly occupied-labelled space inside the red 'box', caused by an insuffient amount of feature points above the wall (*e.g.,* on trees), was removed for visualisation purposes.

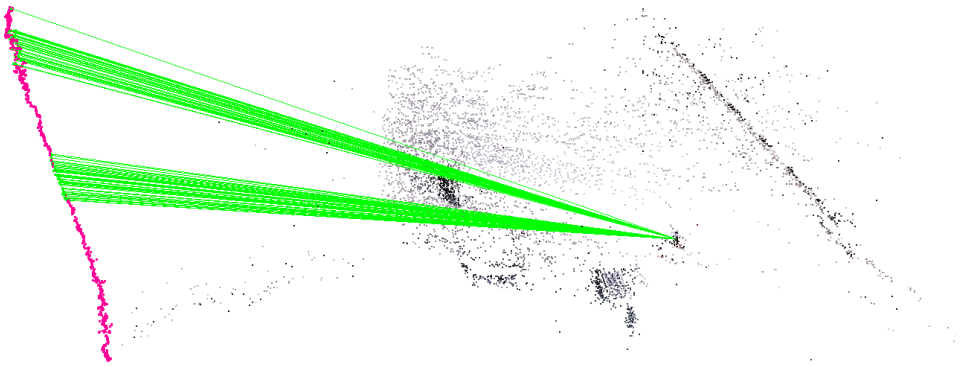


(b) Result of our Visibility-Occlusion Space Carving Veto (Alg. 3.3) with $r = 2500$, $Pr_{incr} = 0.005$

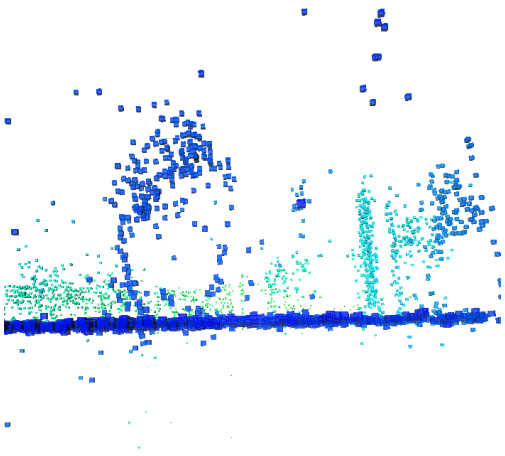
(c) Result of CMVS/PMVS [8]

Figure 5.4: Typical result 2: car_and_wall1 dataset (cont.)

(a) Example frame



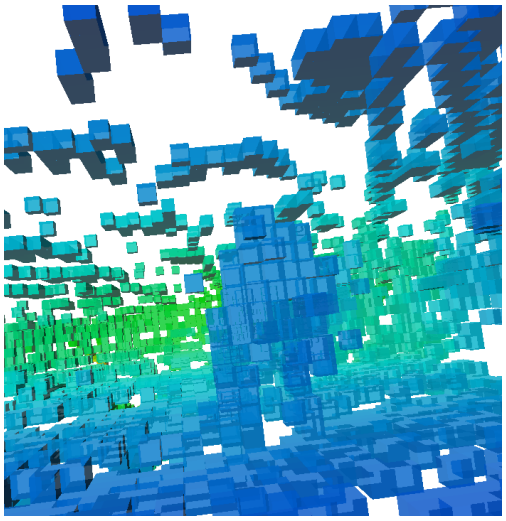
(b) Sparse point cloud and camera poses, annotated with visibility lines from one point



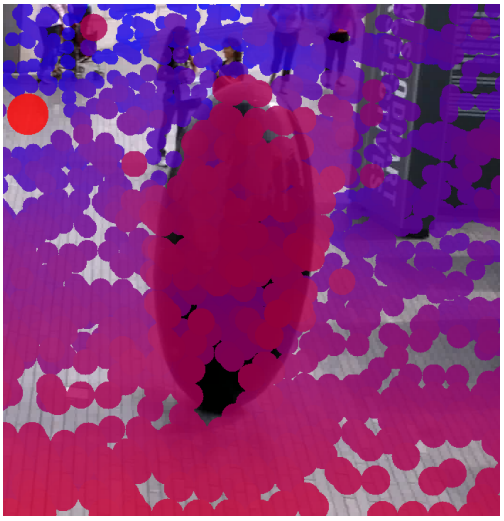
(c) Discretised sparse point cloud ($r = 1000$)

(d) Discretised sparse point cloud ($r = 1000$), carveviewer; notice the incorrectly visible ground voxels (blue-ish) on the lower half of the sculpture

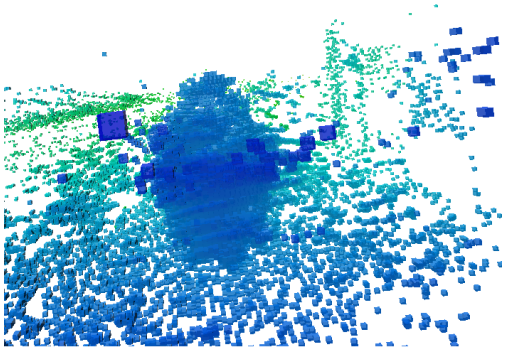
Figure 5.5: Good result 2: memorial dataset



(a) Result of our Visibility Space Carving (Alg. 3.1) with $r = 250$



(b) Result of our Visibility Space Carving (Alg. 3.1) with $r = 250$, carveviewer



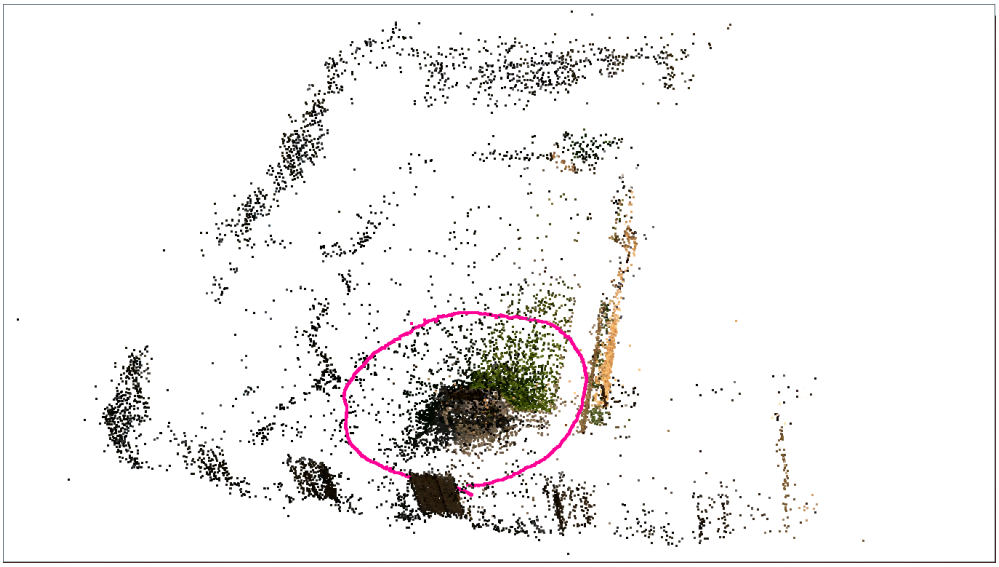
(c) Result of our Visibility-Occlusion Space Carving Veto (Alg. 3.3) with $r = 1000$, $Pr_{incr} = 0.001$

(d) Result of our Visibility-Occlusion Space Carving Veto (Alg. 3.3) with $r = 1000$, $Pr_{incr} = 0.001$, carveviewer

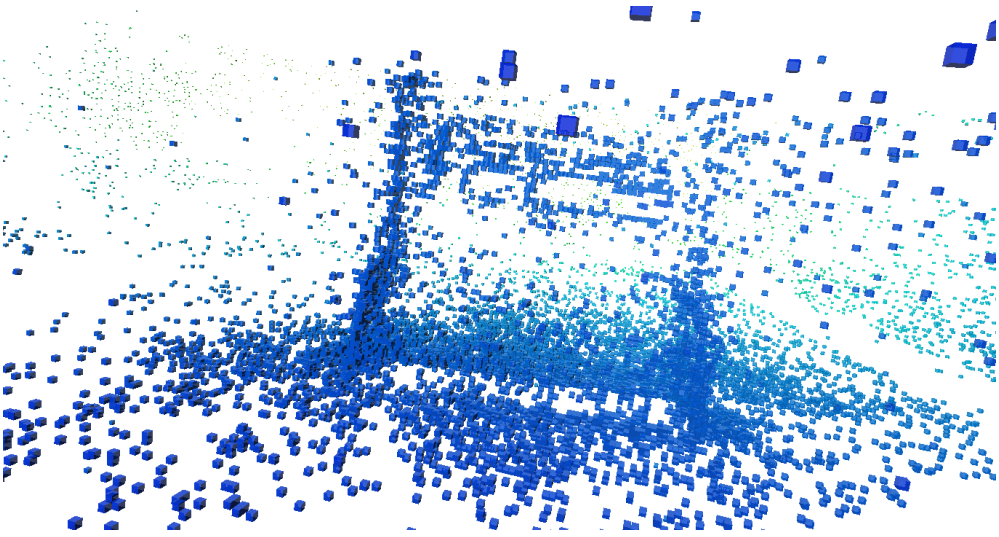
(e) Result of CMVS/PMVS [8]. Notice the similarity with the sparse point cloud as seen from this angle, shown in 5.5(c); lack of texture clearly influences both sparse and dense point cloud reconstructions.

Figure 5.6: Good result 1: sculpture1 dataset (cont.)

(a) Example frame

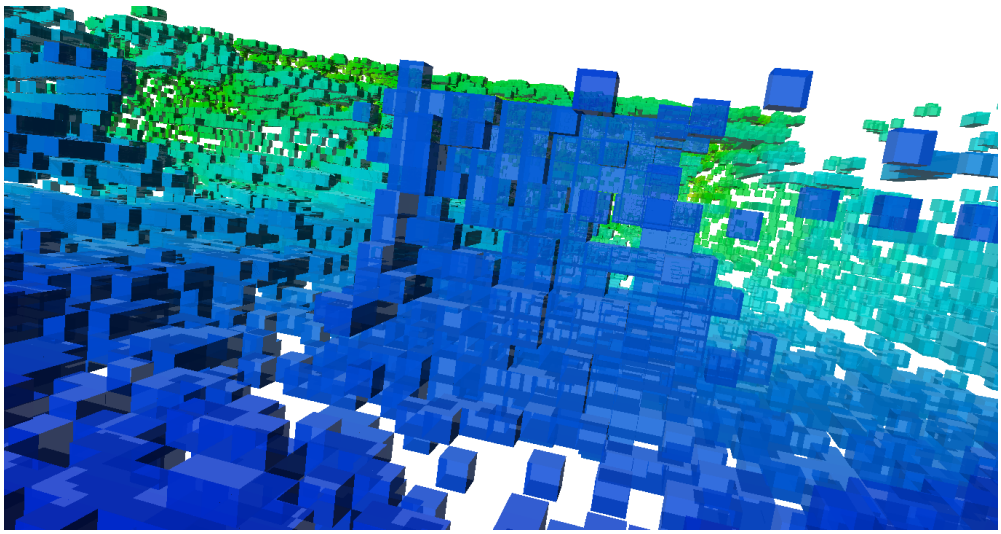


(b) Sparse point cloud and camera poses

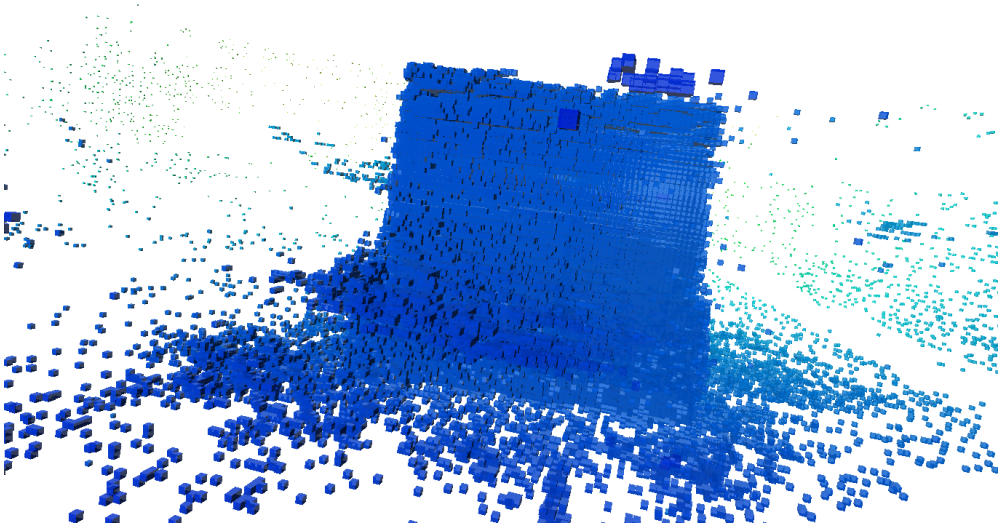


(c) Discretised sparse point cloud ($r = 2500$)

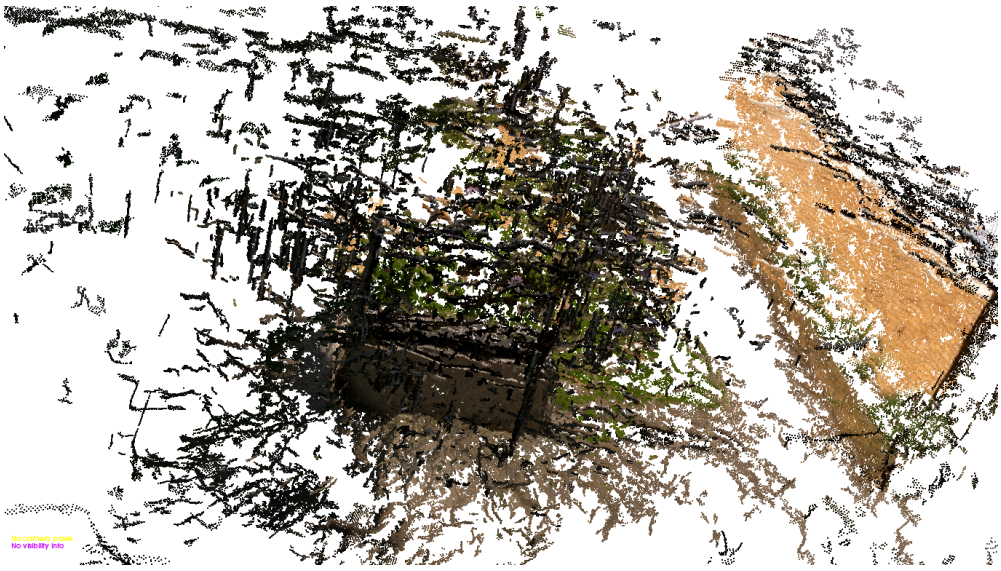
Figure 5.7: Good result 2: memorial dataset



(a) Result of our Visibility Space Carving (Alg. 3.1) with $r = 500$



(b) Result of our Visibility-Occlusion Space Carving Veto (Alg. 3.3) with $r = 2500$, $Pr_{incr} = 0.001$



(c) Result of CMVS/PMVS [8]

Figure 5.8: Good result 2: memorial dataset (cont.)

Chapter 6

Conclusion

Bibliography

[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[3] N. Campbell, G. Vogiatzis, C. Hern, and R. Cipolla. Automatic 3D Object Segmentation in Multiple Views using Volumetric Graph-Cuts. *Proc. 18th British Machine Vision Conference*, 28(1):14–25, 2007.

[4] M. Chandraker. Moving in Stereo : Efficient Structure and Motion Using Lines. In *Computer Vision, 2009 IEEE*, pages 1741–1748, 2009.

[5] M. J. D. Lovi, N. Birkbeck, D. Cobzas. Incremental free-space carving for real-time 3d reconstruction. In *Proc. of 3DPVT 2010*, 2010.

[6] J. Frahm, P. Fite-Georgel, and D. Gallup. Building Rome on a cloudless day. In *Computer VisionECCV 2010*, volume 6314/2010, pages 368–381, 2010.

[7] Y. Furukawa, B. Curless, S. Seitz, and R. Szeliski. Manhattan-world stereo. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1422–1429. Ieee, June 2009.

[8] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–76, Aug. 2010.

[9] L. Guan, M. Pollefeys, and U. N. C. C. Hill. 3D Occlusion Inference from Silhouette Cues. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference*, pages 1–8, 2007.

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[11] R. Hernandez, C. and Vogiatzis, G. and Cipolla. Probabilistic visibility for multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (2007)*, pages 1–8, 2007.

[12] D. Huang, J. and Lee, A.B. and Mumford. Statistics of range images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000*, volume 1, pages 324–331. IEEE Comput. Soc, 2000.

[13] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. KinectFusion : Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera. In *Proc. of the 24th annual ACM symposium on User interface software and technology (UIST '11)*, pages 559–568, 2011.

[14] N. Jojic and B. Frey. Learning flexible sprites in video layers. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–199–I–206. IEEE Comput. Soc, 2001.

[15] W. B. Kai M. Wurm , Armin Hornung , Maren Bennewitz , Cyrill Stachniss. OctoMap: A probabilistic, flexible, and compact 3D map representation for robotic systems. In *Proc. of the ICRA 2010 workshop*, 2010.

[16] S. M. Kutulakos, Kiriakos N. and Seitz. A Theory of Shape by Space Carving. *International Journal of Computer Vision*, 38(3):199–218, 2000.

[17] M. Li Guan and Franco, J.-S. and Pollefeys. Multi-Object Shape Estimation and Tracking from Silhouette Cues. In *IEEE Conference on Computer Vision and Pattern Recognition (2008)*, pages 1–8, 2008.

[18] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov. 2004.

[19] W. Matusik, C. Buehler, and R. Raskar. Image-based visual hulls. In *Proceedings of the 27th . . .*, pages 369–374, 2000.

[20] P. McIlroy, R. Cipolla, and E. Rosten. High-level scene structure using visibility and occlusion. In *Proc. of the British Machine Vision Conference*, pages 1–11, 2011.

[21] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-m. Frahm, R. Yang, and D. Nist. Real-Time Visibility-Based Fusion of Depth Maps. In *IEEE Conference on Computer Vision (2007)*, pages 1–8, 2007.

[22] T. Pan, Q. and Reitmayr, G. and Drummond. ProFORMA: Probabilistic feature-based on-line rapid model acquisition. In *Proc. 20th British Machine Vision Conference (BMVC)*, pages 1–11, 2009.

[23] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles. Detailed Real-Time Urban 3D Reconstruction from Video. *International Journal of Computer Vision*, 78(2):143–167, Oct. 2008.

[24] S. J. D. Prince. *Computer vision : models, learning and inference*. Cambridge University Press, 2012.

[25] R. B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). *Robotics and Automation (ICRA), 2011 IEEE International Conference*, pages 1–4, May 2011.

[26] R. Seitz, S.M. and Curless, B. and Diebel, J. and Scharstein, D. and Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 1:519 – 528, 2006.

[27] P. Smith, T. Drummond, and R. Cipolla. Layered motion segmentation and depth ordering by tracking edges. *IEEE transactions on pattern analysis and machine intelligence*, 26(4):479–94, Apr. 2004.

[28] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics (TOG)*, 25(3):835 – 846, 2006.

[29] E. Wang, J.Y.A. and Adelson. Representing Moving Images with Layers. *Image Processing, IEEE Transactions on*, 3(5):625–638, 1994.

[30] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *Cvpr 2011*, pages 3057–3064. Ieee, June 2011.

[31] C. Zach, A. Irschara, and H. Bischof. What can missing correspondences tell us about 3D structure and motion? In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, volume 1, pages 1–8, 2008.

Appendix A

Installation manual

Appendix B

User manual

Appendix C

Datasets

Appendix D

Code listing