Calibration and Component Placement in Structured Light Systems for 3D Reconstruction Tasks

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Nathaniel Davis Bird

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Nikolaos Papanikolopoulos, Adviser

September 2009

Acknowledgements

There is a long list of people who deserve thanks in the creation of this thesis. The most notable are my parents, Ashley and Judith Bird, without whose constant love and support this would not have been at all possible.

My advisor, Nikos Papanikolopoulos, deserves a lot of thanks for somehow managing to put up with me for the six years it took me to meander my way through grad school. My committee, Arindam Banerjee, Vicki Interrante, and Dan Kersten, deserved thanks for providing invaluable input to this thesis. All my lab mates deserve some credit as well, for always being around to pester and bounce ideas off of, over the course of many years—in no particular order, many thanks to Osama Masoud, Stefan Atev, Hemanth Arumugam, Harini Veeraraghavan, Rob Martin, Bill Toczyski, Rob Bodor, Evan Ribnick, Duc Fehr, Ajay Joshi, and the rest in the lab. Of course, the innumerable professors I have taken classes from deserve a place here as well.

The financial support I have received through the years I spent in grad school is very much appreciated. The work presented here was supported by the National Science Foundation through grant #CNS-0821474, the Medical Devices Center at the University of Minnesota, and the Digital Technology Center at the University of Minnesota. Over the course of my time in graduate school, I have also been supported by the National Science Foundation on other grants, the Minnesota Department of Transportation, the Department of Homeland Security, and the Computer Science Department itself.

Many thanks to all.

Abstract

This thesis examines the amount of detail in 3D scene reconstruction that can be extracted using structured-light camera and projector based systems. Structured light systems are similar to multi-camera stereoscopic systems, except that a projector is use in place of at least one camera. This aids 3D scene reconstruction by greatly simplifying the correspondence problem, *i.e.*, identifying the same world point in multiple images.

The motivation for this work comes from problems involved with the helical tomotherapy device in use at the University of Minnesota. This device performs conformal radiation therapy, delivering high radiation dosage to certain patient body areas, but lower dosage elsewhere. The device currently has no feedback as to the patient's body positioning, and vision-based methods are promising. The tolerances for such tracking are very tight, requiring methods that maximize the quality of reconstruction through good element placement and calibration.

Optimal placement of cameras and projectors for specific detection tasks is examined, and a mathematical basis for judging the quality of camera and projector placement is derived. Two competing interests are taken into account for these quality measures: the overall visibility for the volume of interest, *i.e.*, how much of a target object is visible; and the scale of visibility for the volume of interest, *i.e.*, how precisely points can be detected.

Optimal calibration of camera and projector systems is examined as well. Calibration is important as poor calibration will ultimately lead to a poor quality reconstruction. This is a difficult problem because projected patterns do not conform to any set geometric constraints when projected onto general scenes. Such constraints are often necessary for calibration. However, it can be shown that an optimal image-based calibration can be found for camera and projector systems if there are at least two cameras whose views overlap that of the projector.

The overall quality of scene reconstruction from structured light systems is a complex problem. The work in this thesis analyzes this problem from multiple directions and provides methods and solutions that can be applied to real-world systems.

Contents

List of Tables

List of Figures

1 Introduction

The motivation for this thesis is vision-based, full-body patient tracking. This comes from the limitations of the helical tomotherapy device currently in use at the University of Minnesota. The helical tomotherapy device is a radiological treatment method capable of delivering high radiation dosage to certain body areas, while leaving other areas with lower dosage. This type of radiation therapy is known as conformal treatment, as the radiation delivered conforms to the shape of the target, sparing other areas of the body high radiation exposure. To reliably treat patients, it is vitally important to detect when a patient becomes misaligned during treatment and adapt to it. The movement detection tolerances are very tight: movements of just five millimeters out of alignment can adversely affect treatment.

A primary limitation of the helical tomotherapy device is its inability to sense patient movement, making it impossible to adjust the radiation delivery plan to compensate. At this point, detection of *any* movement would be an improvement as the device currently has no feedback mechanism to report on the patient's position while treatment is underway. Thus, problems associated with the precise tracking of a patient's body tracking are the central focus of the work presented in this thesis.

1.1 Helical Tomotherapy

Helical tomotherapy is a recent radiological treatment option for certain cancers. It is shown by Hui, *et al.* [32] and Hui, *et al.* [33] that it can be used to deliver high radiation doses to specific body regions, while leaving surrounding tissues within acceptable doses. Total Marrow Irradiation (TMI) treatment is currently being developed which utilizes this method. TMI is promising because conformal irradiation of select areas can increase the likelihood of successful treatment. TMI is used as a pre-conditioning regimen before bone marrow transplantation. See Figure 1 for an image of the helical tomotherapy device.

Figure 1: Helical tomotherapy treatment device. The patient lies on the platform shown, which moves in 3D through the bore, the circular enclosure that houses the radiation source. It is a largely automated procedure, but currently has no way to ensure the patients are where they are expected to be.

The procedure for treatment using the helical tomotherapy device is as follows. The patient's internal geometry is measured using a high-fidelity kilovoltage CT-scan (kVCT) prior to the first treatment session. This is done using a different machine than the helical tomotherapy device. The scan data is then used by doctors to plan the radiation doses that all interior portions of the body will receive as part of the treatment regimen. At the start of a treatment session, a lower-fidelity megavoltage CT-scan (MVCT) is performed using a scanner built into the helical tomotherapy machine. The MVCT scan is compared, using static registration methods, to the kVCT scan which was used to plan the treatment to make sure the patient is in the same position. After the patient is lined up, the treatment procedure is started. No one besides the patient can be in the room while treatment is underway due to the radiation. With the exception of the emergency stop button, the procedure runs autonomously once started. Although it is monitored by doctors via closed-circuit television. This radiation therapy is often performed in multiple sessions, called fractions, which take approximately 20 minutes each (not including the significant initial placement time, which can double or even triple the time required).

Despite being a state-of-the-art autonomous treatment, the helical tomotherapy procedure is open-loop. That is, there is no feedback to the system while treatment is underway as to the patient's current position and articulation, or whether the

patient's current pose is within treatment tolerances. The system is blind after the initial positioning because the MVCT scan processes cannot be run simultaneously with the treatment, due to the radiation being emitted. Because the patient's position cannot be readily measured, a variety of problems are encountered. For instance, there are movement problems due to the length of treatment. Patients are supposed to lie still for approximately 20 minutes while treatment is underway, but remaining still for that long is very difficult due to shifting, itching, shivering, tapping, and general discomfort. Physical restraints on the patient have been insufficient. Even small movements are very important because a difference of just five millimeters can mean that a high radiation dose is no longer being delivered to cancerous marrow in a rib, but instead to the soft tissue and organs around it. This reduces the effectiveness of treatment.

Vision-based tracking methods are useful here as they can be used to monitor the patient throughout the treatment process. Besides detecting movement while treatment is underway, vision can be useful for the initial positioning of the patient within the device. This is due to the small changes in a patient's pose which occur from one fraction to the next. Taking a CT-scan is a time consuming process, making this constant reimaging of patients not only unwelcome by the patients themselves, but costly due to staffing and scheduling constraints at clinics. Using vision, the initial positioning for each fraction could be performed faster and more cost effectively.

A computer vision system to precisely monitor the patient's body position should be able to detect movement past acceptable parameters. This work is important, because closing the loop on the control of helical tomotherapy treatment should improve the safety of the patient, reduce the considerable treatment time, and improve the cure rate.

1.2 Structured Light Systems

Structured light refers to vision systems which utilize projected light patterns for reconstruction tasks. It functions much like stereopsis, with the difference being that instead of using multiple cameras, a set of cameras and a set of *structured light sources* of some type are used. In practice, projectors are typically used as the structured

light source. In these systems, the projectors act mathematically like cameras, only instead of detecting features inherent in the scene, artificial features are projected out onto the scene, which the cameras then detect. These artificial features are easier to detect reliably than naturally occurring features in the scene, which is why structured light systems are used.

One limitation of structured light systems is that they are usable only in situations in which the lighting can be strictly controlled. If the light can change haphazardly, locating the projected patterns in the camera images can be very difficult. While this is an issue for outdoor use, lighting can be strictly controlled within medical facilities.

1.3 Contributions

This thesis makes two main contributions:

1. Metrics to judge the quality of element placement for structured light systems are created and explored. In this context, elements include cameras and projectors in a structured light system. This is explored in depth in Chapter 6.

2. A calibration technique for structured light systems that requires no *a priori* information about the scene structure, and which is optimal in terms of reprojection error. This is explored in depth in Chapter 7.

These contributions, while motivated by the underlying problem of reconstructing a patient's body surface, are broadly applicable to all structured light research. Beyond medical vision, the placement metrics described herein can be used for any reconstruction task for which a structured light system is being considered. Similarly, the calibration method presented is useful for *any* structured light system, not just those that deal with patient body surface modeling.

2 Related Literature

The creation of a system of the type described here requires knowledge from several diverse fields of study. One example field is medical tracking, which typically focuses on tracking movement or deformation of a localized region of the body instead of the entire body. The computer vision community does work on tracking the entire human body but usually from far away. This lowers the fidelity of measurements to a point where it is not useable here. Stereopsis is necessary because the only way to determine the 3D structure of the scene using imaging sensors is to use several of them. Structured light can be thought of as an extension of stereopsis which replaces at least one camera with a projector. Structured light is useful here because it lets us use stereoscopic techniques but deals with some associated difficulties. Camera placement is an important issue as well because poor placement can lead to measurements containing higher error than necessary. Also, some form of manifold modeling must be used as well to build up a representation of the patient's body surface. Finally, augmented reality is useful because it provides us with a means to provide feedback to the medical doctors using the system.

2.1 Tracking from the Biomedical Field

In the medical field, there has been various work on tracking patient positions. Meeks, *et al.* [55] present a brief survey of vision-based techniques for patient localization in a clinical environment. Their survey discuses tracking active markers, such as infrared emitters, as well as passive markers attached to the patient's body. They also discuss how to accurately track points within the body using x-ray data by tracking implanted fiducial markers.

Surgically implanted fiducial markers are one of the most accurate ways to track organs within the body. Essentially, a fiducial marker is a small item that shows up well in X-ray photographs. By implanting these markers near an organ of interest,

which does not show up well under X-ray, the organ can be localized better. These provide good tracking but the implantation procedure is time consuming, invasive, and costly. It has further been reported that these fiducial markers can migrate within the body over time [52].

MVCT is very useful for imaging the body to localize areas of interest within the patient before treatment. Moore, *et al.* [59] present a method to determine the distance of 3D surface scans obtained from a stereoscopic camera pair to this type of data. Their fit is used to adjust the patient's pose to match that of the CT-scan for treatment. Unfortunately, many of the algorithms used in practice to correlate patient body scans are considered to be industry secrets, and are not represented in the literature.

Structured light is another method by which a patient can be monitored. Mackay, *et al.* [52] use a structured light technique for monitoring a portion of a patient's body surface. They showed that body surface measurement can be used to correct patient body position to better target cancerous growths of the prostate during treatment. This is not used to track the whole body.

An example of patient tracking using attached markers is the work presented by Linthout, *et al.* [48]. In this work, a set of stereo cameras tracked a pattern of infrared markers attached to a head immobilization device on a set of patients for radiation therapy treatment. This was used to correct the patients' poses between treatment fractions and to monitor movement during treatment. Only translational movement was monitored.

Finally, an example full-body differencing algorithm was demonstrated by Milliken, *et al.* [57]. In this work, a method for video-image-subtraction based patient positioning using multiple cameras was presented. Stereo vision algorithms were not used, just an image difference in two images. This is another method used to position the patient prior to treatment. One interesting comment from this work points out that there is significant variability in the quality of patient repositioning, which seems to be dependent on the skill of the physician. This leads to an intuition that treatments will likely turn out better if the device adapts to the patient instead of the other way around.

2.1.1 Breathing Analysis

The work of Fox, *et al.* [26] demonstrates a method, termed free breathing gated delivery, to monitor the breathing cycle of a patient undergoing conformal radiation therapy targeting the chest cavity. They use a set of two infrared reflective markers on the patient's chest and an associated camera array to triangulate the position of the markers. The patient's breathing cycle is monitored by measuring the location of these two reflective markers, and the radiation source is activated only when they are within a threshold range, determined by the doctor when the treatment is planned.

The work of Wang, *et al.* [79] utilizes a camera-only system to monitor breathing frequency. Infrared cameras are positioned to give a good view of the patient's upper torso while the patient sleeps. Interframe image differencing is used to create a count of how many pixels are different from one frame to the next. By examining the periodicity of this count, the breathing period can be determined. This is monitored to discover when it becomes interrupted due to sleep apnea. The distance of the camera in this system is too far for fine measurements, however.

The work of Chihak, *et al.* [19] describes a system to monitor breathing which uses a grid of infrared reflective markers adhered directly to the skin. They distinguish between three types of breathing based on which portions of the torso move as part of the breathing motion.

Outside of medical breathing monitoring, there are methods that could be useful for breathing classification. Agovic, *et al.* [1] present a comparison of Isomap, LLE, and MDS embedding for separating anomalous shipping trucks from regular ones using a sparse set of information about the trucks in question. In general, such embeddings seem to show merit. However, the authors point out that as with many machine learning techniques, the embeddings arrived at by the methods for identifying potential anomalies in the data may not indicate anomalies as human operators recognize them. Extending this to apply to a data stream as opposed to regular data points could be quite interesting.

Another work from the vision community that might be useful is that of Cutler and Davis [21] which identifies periodic motion in tracked areas of video images. The focus in this work is spotting human movement from long distance, specifically from

robotic aircraft. Several of the heuristics utilized may be useful for application to breathing analysis.

2.2 Vision-Based Human Behavior Tracking and Recognition

There is a lot of work from the computer vision community concerning tracking of the human body. Although in these works, the application areas are very different than the precise, medical domain area we are examining here. For instance, methods dealing with human activities recognition do not require the location of a subject's arm to the nearest millimeter. Thus, methods created for these applications often involve fundamental assumptions that preclude them from working well for precise body tracking.

Some methods rely upon background segmentation to detect silhouettes of humans of interest for classification of their location, orientation, and activities. This type of research focuses on the middle field, where a standing human is approximately 100 pixels tall. This is good when dealing with people from a security standpoint, but does not provide the level of precision needed in medical applications. Examples of these techniques include the Pfinder system by Wren, *et al.* [80], and the W4 system by Haritaoglu and Davis [30].

Limb tracking is another active area of research, often for novel computer interfaces. One recent example is Siddiqui and Medioni's work [68], in which edge features coupled with skin tone detection is used to track the forearms of subjects wearing short-sleeved shirts. The arm's 3D pose and orientation are found based on the assumption that the limbs act as rigid bodies. Methods like this which use skin tone as a cue would not work in the presence of more skin-tone than just the arm, and self occlusions would throw it off. This is because areas of interest are no longer separated from each other. Another limb tracking method is presented by Duric, *et al.* [22], in which an optical flow method is used to locate an arm moving through space. In low motion, this type of solution is also unlikely to work well.

Human pose estimation is another area of intense research. Lee and Cohen [45] present a method which uses skin-tone detection to pick out salient features from

static images representing the head, arms, and legs. They assume T-shirts and shorts are worn, making such segmentation much easier. A data-driven Markov chain Monte Carlo method is used to find the pose of a standard kinematic human model with the maximum likelihood of generating the observed data. A recent example using video would be the work of Bissacco, *et al.* [10], which uses appearance-based features and interframe differencing to locate people within a scene. A boosting method uses these input features to learn classifiers for human joint angles.

The author has also done work on vision-based human behavior detection, tracking, and recognition as well. For instance, methods to detect loitering individuals at public bus stops are presented in [28, 9]. Methods for differentiating between safe and unsafe motorist behavior are presented in [76, 77]. Finally, a system for detecting abandoned objects in public areas is presented in [8]. These methods deal with tracking humans and identifying their behaviors, but at a different scale than is needed in this work.

2.3 Stereopsis

The stereoscopic vision literature covers techniques for combining data from multiple views to extract 3D information about the scene in question. Hartley and Zisserman [31] provide the definitive tome on the computational geometry underlying stereoscopic vision systems. The coverage of calibration and 3D structure estimation is excellent. However, methods for matching points of interest across different views are not discussed. This issue is known as the correspondence problem, and it is very difficult to solve in a general sense. Many of the problems encountered when using stereoscopic systems can be attributed to the correspondence problem.

One way to try to solve the correspondence problem in stereo vision is to do patch matching. Unfortunately, the patches will not be aligned in both images in the general case. This can be addressed by rectifying the scene images, as per the work of Fusiello, *et al.* [27]. Another way to address this issue is to use features designed to be invariant to camera view, such as SIFT [51], or SURF [6]. These features work well for matching, but deciding which area of the scene to use for matching is not a straightforward matter. For low-contrast areas like skin or regular clothing, any

features will not work very well because everything looks roughly alike.

An interesting probabilistic treatment of the stereoscopic camera calibration problem is presented by Sundareswara and Schrater [69]. Essentially, instead of simply choosing the most probable camera calibration for use in the system, a distribution of camera calibrations is used instead. Thus, the calibration estimation covariance is propagated into the triangulation calculations, providing better estimates.

2.4 Structured Light Vision

Structured light is a vision-based method to perform 3D reconstruction of a given surface, much like stereo vision. The difference is that instead of using multiple cameras, a set of cameras and projectors is used instead. Mathematically, projectors act as cameras, only instead of detecting features inherent in the scene, they project artificial features out onto the scene for the actual cameras in the system to detect. Thus, while the correspondence problem is a major issue for stereo vision systems, it is significantly less important for structured light systems. A major problem with structured light though, is that it requires an environment in which the ambient lighting can be controlled. In uncontrolled environments, the projected features may not be visible.

Batlle, *et al.* [5] offer a good survey of the structured light literature up to about 1996. They introduce the geometry involved and relate structured light to stereoscopic camera systems. They also do a good job explaining where structured light methods can succeed where stereoscopic methods cannot–the primary reason being the ease at which structured light techniques can solve the correspondence problem. Salvi, *et al.* [64] offer a newer survey of the structured light literature, providing useful classifications of existing methods. They also provide a more intuitive explanation of the types of problems the different classes of approaches are applicable to.

The motivating application of this thesis is conceptually similar to the application presented in the Ph.D. thesis of Livingston [49]. In his work, he proposed a two projector, one camera structured-light system to measure the surface of the body for unspecified treatments. His structured light system was very simple, projecting only a pixel at a time, which simplifies the correspondence problem to be solved, but

at the cost of requiring a lot of time to completely scan a surface. In addition, no registration of the patient body surface is performed. The only results presented were on the calibration of this system and the expected error from it.

Caspi, *et al.* [14] present a colored structured light method that compensates for the underlying color of objects within the scene. Their method expands on the visual Gray codes, which were first presented by Inokuchi, *et al.* [36]. Both of these methods project patterns onto the scene which, after capturing many frames, can be used to solve for many points at a time.

Koninckx, *et al.* [42] propose a solution to the problems of aliasing (due to foreshortening) and over exposure of the structured light in the scene by iteratively adapting the projected patterns for better visibility. They present an interesting idea for better scene reconstruction—use an estimate of scene geometry to project a pattern that is as easy as possible for the camera to determine.

Raskar, *et al.* [61] present a hypothetical system that uses structured light to keep track of an office environment. In particular, they propose using *imperceptible* structured light, in which the projector swiftly alternates between two different projected patterns. If the rate of change is fast enough, all a human perceives is a uniform light but a synchronized camera can detect each image individually. This would theoretically allow a structured light system to be used where it would otherwise visually disturb human users. A synchronized camera and projector system is demonstrated.

Lanman, *et al.* [44] present and interesting system utilizing a camera, projector, and a set of mirrors to project and detect structured light on multiple sides of an object in a single scan. This address the problem of occlusion that plagues all vision systems by cutting down on the portions that cannot be seen. However, this method seems designed to work for freestanding objects where mirrors could be placed to cover the entire object. This would not work on patients lying on beds, especially where the movement that needs to be detected is so small.

Moiré patterns form an interesting subset of the structured light literature. The classic work by Takasaki [70] introduced the use of moiré topography to project contour lines onto an object of interest. Essentially, a planar grating between the object and a point light source is used to project contour lines onto the object. The result is very similar to a topological map. The work by Kim, *et al.* [41] applies the

moiré pattern idea to the wide scale detection of scoliosis, a disease that leaves the spine deformed. The focus in this work is on using learning methods to discriminate between the moiré patterns projected onto healthy versus those projected onto the backs of children suffering from scoliosis. This is done using a linear discriminant function on region patch density in the moiré image. In addition, Reid [62] provides a (rather dated) survey on analyzing fringe patterns such as those in moiré images.

2.5 Camera and Sensor Placement

Most work in camera and sensor placement is built around variations of the art gallery problem [66]. The general formulation is that for a given polygon-shaped room, the minimum number of cameras or security guards needed to watch the gallery must be found. Solutions to this and related problems are used to find the number of cameras or other sensors are required for complete operability.

Exterior visibility is the problem of determining the minimum number of cameras to deploy around the outside of a polygon or polyhedron such that the object of interest is completely visible. It is essentially the inverse of the art gallery problem. The work of Isler, *et al.* [37] presents a theoretical work on exterior visibility in which they show that in a two dimensional case, a polygon must to have five or fewer vertices to be guaranteed to be viewable by an unbound number of cameras.

On the more practical side, the work of Bodor, *et al.* [11], [12] presents a method to determine the placement of a set of cameras for best visibility based upon an example distribution of trajectories. This work places the cameras by minimizing a cost function which penalizes foreshortening and poor resolution. The work of Olague and Mohr [60] introduces a genetic algorithm-based method to place cameras around virtual objects to minimize the scene reconstruction error. The work of Mittal and Davis [58] deals with camera placement in the presence of randomly positioned occlusions in the environment for which an example distribution is known. Cameras are placed such that the probability of occlusion is minimized. Chen and Davis [18] present a camera placement parametrization which considers self-occlusions. Their work also provides a metric for analyzing error in the 3D position of a point seen by several cameras. Tekdas and Isler [73] present a method for optimally placing a

stereoscopic camera pair to minimize the localization error for ground-based targets in 2D. Krause, *et al.* [43] present a method utilizing Gaussian processes for 2D sensor placement.

The two dimensional error regions discussed by Kelly [40] provide justification for considering the real-world projection of detection error, as the area of this error changes for different sensor placements.

The work presented here seeks to expand upon this existing literature by expanding the placement question beyond merely cameras and into camera and projector systems.

2.6 Calibration of Structured Light Systems

Research in the area of calibration of structured light systems tend to have the same general problem setup and solutions. They all deal with the problem of calibrating a structured light system consisting of one camera and one projector. In all the methods discussed in this section, specific geometric constraints on the scene must be known a priori for calibration of the structured light projection device. The nature of this a priori information is the central element that changes from method to method.

A method to calibrate a camera and laser-stripe projector system by scanning objects in the environment and then measuring the exact location of these points by moving a robotic manipulator to their positions is presented by Chen and Kak [16]. A similar work where a laser-stripe projector and camera are calibrated with respect to each other by scanning a known object and then identifying specific points when the laser-light plane strikes them is presented by Theodoracatos and Calkins [74].

The work by Wakitani, *et al.* [78] presents a calibration method which utilizes a robotic calibration rig. The rig moves a plane of known dimensions around the scene specifically. The camera is calibrated by noting the location of printed landmarks on the plane while the stripe projector is calibrated using the intersections of the light plane with the calibration rig plane. Similarly, the work by Che and Ni [15] presents a calibration method for a mechanically-moving laser-stripe based structured light system based on scanning a tetrahedron target with known dimensions.

The method presented by Reid [63] performs a homography calibration between

a plane illuminated by a laser-stripe projector mounted to a mobile robot with an attached mounted camera. This calibration is not Euclidean, and cannot be used for 3D reconstruction, but is sufficient for localization on a 2D map. This method requires a precisely known scene to calibrate.

The work of Bouguet and Perona [13] presents a creative take on low-accuracy structured light using nothing more than a desk lamp and a hand-moved pencil to cast essentially structured shadows onto a small scene. In this system, the calibration parameters for the lamp are found by manually measuring the location of the shadow of a pencil that is placed upright in the scene. This can be thought of as "structured shadow" as opposed to structured light.

A method is presented by Huynh [34] and Huynh, *et al.* [35] which calibrates a laser-stripe structured light system by calculating a projection matrix for each individual light stripe. Not only does this method require painstaking measurement of where each individual light stripe intersects the scene, but it represents the projector with a set of projection matrices while the ideal solution is a single projection matrix for the projector.

The method presented by Jokinen [38] finds the extrinsic calibration parameters for a laser and camera structured light system based on structure-from-motion techniques. The intrinsic parameters are considered known a priori. This method requires a movable structured light rig and is not for calibrating fixed systems.

The work in Sansoni, *et al.* [65] presents a calibration method that constrains the environment by including only a parallepipedic block of known dimensions, in addition to requiring that the camera be positioned perpendicular to the reference frame associated with the block.

The work in Fofi, *et al.* [24] and Fofi, *et al.* [25] presents a method for calibrating a single camera single projector system without the use of a calibration pattern. However, they require the scene to have a very specific setup–a pair of orthogonal planes. They recognize distortions in a known pattern to estimate which points detected by the camera belong to each plane and calculate a reconstruction based on this. This method does not produce the Euclidean calibration, as there is no way for the method to determine the scale.

The work of Chu, *et al.* [20] presents a method in which a mobile camera can be

used with a fixed light-plane projector mounted to a frame with identifiable LEDs attached to it. The camera can be moved freely. The LEDs on the frame are used to recover the camera's extrinsic parameters at any point in time, which can then be used to perform reconstruction of objects within the frame. This method requires the intrinsic parameters of the camera and all parameters of the light-plane projector to be known a priori.

The work presented by McIvor [53] and McIvor [54] demonstrates a method to calibrate a camera and light-plane projector. In this case, the scene is restricted by requiring a specific object to be present in the scene while calibrating. Positions on this object are each identified by a human operator, and the laser stripe that intersects each is used to find the calibration parameters of the system.

The work in Chen and Li [17] and Li and Chen [46] addresses the problem of finding the calibration of a camera projector pair that has previously been calculated, but in which the relative placement of the camera and projector has been changed by movement on the z-axis and rotation about the y-axis. The mechanism used provides many constraints to relative movement, which are used to estimate the new calibration parameters based on the image appearance of a projected pattern on a plane.

The research by Shin and Kim [67] is a work in which a single camera is calibrated off of a known pattern first, followed by a period in which the projector shines another known pattern onto the scene. A human must then painstakingly measure the world location of each of the points in this projected pattern, as they are being projected, which are then used to calibrate the projector.

The method of Zheng and Kong [84] calibrates the camera prior to the projector, and then calibrates the projector using geometric constraints imposed by manually placing a plane in the scene such that projected patterns appear correctly on it. This plane must be moved around to several specific locations in order to collect sufficient data to calibrate the scene.

The work of Zhang and Li [82] starts with a calibrated stationary projector, which projects a known pattern onto the scene. Using this information, a method is formulated by which a single moving camera is calibrated based upon the image of the pattern.

The method presented by Kai, *et al.* [39] utilizes neural networks to estimate the calibration of a structured light system. This is fascinating from a machine learning standpoint, but an optimal method would be preferable.

The work of Zhang, *et al.* [83] presents a calibration method for structured light systems which requires a planar surface in the scene. Intrinsic parameters are assumed to be known a priori. A single solution is not assured.

The research presented by Liao and Cai [47] is interesting in that they line up a projected pattern with a pattern painted onto an object in the scene. This requires significant human interaction with the system to manually line up a projected calibration pattern precisely with a physical calibration pattern in the scene.

The work of Yamauchi, *et al.* [81] presents a method in which a camera is calibrated first, and then the projector is calibrated with significant human interaction during which reference planes are precisely setup in the camera and projector's view.

Finally, the method presented by Audet and Okutomi [2] is another calibration method for structured light systems that requires the user to move a plane of known dimensions around the scene. This plane has markers painted at specific locations on it. A projected pattern is then aligned with the markers. When this algorithm is run with the board at multiple locations, calibration for the system can be found.

2.7 Augmented Reality

Augmented reality (AR) is a version of virtual reality with the goal to combine virtual with real-world sensory data convincingly and intuitively. For instance, a classical example of an AR system is one in which a user views the world through a head mounted display. A computer recognizes objects, such as buildings, that the user is looking at and adds extra information about them to the user's view. This could be as simple as a label along the bottom of the user's view with the name of the building, or as complex as virtual X-ray vision created by overlaying a rendered view of the inside of the building over the user's view of the building.

To provide meaningful feedback for the initial patient positioning problem in helical tomotherapy, the patient's actual position needs to be synchronized with a previous body scan, with the resulting discrepancy communicated to the doctor. AR

provides an intuitive way to feed this information about where the patient should be back to the user. In fact, projector(s) used by the incorporated structured light methods could be used to display this information directly on the patient and treatment area. Raskar, *et al.* [61] talk about a similar combination of structured light and augmented reality, but did not actually build a system. By projecting this information onto the patient ambiguity in what is being communicated can be eliminated.

A good survey of the augmented reality literature is presented by Azuma, *et al.* [4]. This survey built upon an earlier survey by Azuma [3]. In the nomenclature, the system we are proposing would be a projective system because the virtual data is intended to be projected onto the real-world objects. This is unlike other forms of AR, which typically require the use of head-mounted monitors and cameras. At the time these surveys were published, real-time operation was not present in most AR systems.

A classical work tying computer vision to the medical arena is the work of Grimson, *et al.* [29] in which input from a laser range scanner is registered with an existing model of the patient created from CT or MRI data in a user-guided manner. Once registration is complete, they overlay a rendering of their virtual model over a video of the patient. The work of Mellor [56] is similar, but the registration is performed using circular markers fixed to the patient's head. Other examples of augmented reality in surgical applications include Betting, *et al.* [7], Edwards, *et al.* [23], Lorensen, *et al.* [50], and Taubes [72].

3 Preliminary Attached Marker-Based Experiments

This chapter describes initial research into the problem of tracking the patient for the helical tomotherapy treatment. The vision-based tracking system is set up as follows. First, a stereoscopic camera system is utilized to track near-infrared reflective markers attached to a patient. The initial locations of these points are recorded, and the points are subsequently tracked in 3D space. Boney body structures, such as the head, can be tracked independently of the rest of the body points, providing roll, pitch, and yaw information about their pose.

3.1 Preliminary Stereoscopic Marker Detection

In the preliminary system, near-infrared reflective markers were used as tracking points on the patient's body. This is to show proof of concept that a minimally-invasive vision system can provide sufficiently accurate tracking information. Physical markers that must be attached accurately are still invasive, however. Non-contact vision modalities are vastly preferred for any usable system.

A pre-packaged stereoscopic near-infrared marker detection system built by Motion Analysis Corp.[1] was used to detect the markers placed on a patient. The markers, shown in Figure 3, are small rubber spheres covered with reflective tape. These markers are then attached to the patient using an adhesive on their bases. The cameras built by Motion Analysis Corp. have a ring of near-infrared LEDs positioned around the lens, and the reflective tape on the markers reflects the light from these LEDs back to the camera. The cameras are sensitive to this near-infrared light. The cameras capture intensity images, which are then thresholded, and the centroid of each high-intensity connected region is found. The cameras are calibrated, so the locations of these centers in each cameras image plane are then used to determine

[1] The Motion Analysis Corp. website is located at `http://www.motionanalysis.com`

the 3D location of each detected marker. The entire stereoscopic marker detection system returns a set of 3D marker locations at a rate of 60 Hz.

3.2 Point Tracking

The point tracking algorithm works by correlating tracking identification numbers from one frame to the next. Essentially, the tracking number of the nearest point in the last frame is assigned to each point in the current frame. Defining $l_{(i,n)}$ to be the location of point i in the current frame and $l_{(j,n-1)}$ to be the location of some point j in the previous frame, the tracking identification number associated with $l_{(i,n)}$ will be the tracking identification number associated with the point in the previous frame satisfying $\min_j\{\|l_{(i,n)} - l_{(j,n-1)}\|_2^2\}$.

Consistency checks are performed to make sure no identification number is assigned to more than one point in the current frame. If two or more points in the current frame are associated with a single point in the previous frame, only the closest point in the current frame will be associated with it. If a point does not correlate with any point in the previous frame, it is given a new tracking number. A threshold distance d_{thresh} is set for the maximum distance allowed for a point in the current frame to correlate with a point in the previous frame. This value should be based upon the expected movement of points between frames. Thus, a point $l_{(i,n)}$ will not correlate with any points in the previous frame if $\min_j\{\|l_{(i,n)} - l_{(j,n-1)}\|_2^2\} > d_{thresh}$.

In addition, points that disappear are propagated in their current location for a short duration of time, in the hope that they will reappear shortly. This is done to prevent losing points that temporarily disappear due to a momentary hiccup in the stereoscopic marker detection system. A threshold is manually set as to how many frames a point should be thus propagated since it has last been seen.

3.3 Rigid-Body Tracking

Tracking points on the body is useful, but being able to talk about the articulation of segments is more informative. Using the point tracking algorithm as a base, body segments can be tracked as well. Boney body segments such as the head or the breastplate can be fairly well approximated as rigid bodies individually as they do

not deform much, but not with respect to each other. This allows us to collect information about the location and orientation of individual body parts throughout the treatment process, which provides meaningful feedback for medical personnel to reposition patients.

The rigid bodies tracked must be manually defined by an operator on captured data sets. They choose a set of points for each rigid body in the initial frame, which are then tracked.

In three dimensions, a rigid body contains six degrees of freedom. The three dimensional location of the rigid body, $\mathbf{x} = (x, y, z)^{\mathrm{T}}$, is found by taking the mean location of the points associated with it. The locations of each of the constituent points are then found with respect to this center location. Rotation is described by an angle of rotation, θ, and a unit vector around which rotation takes place, $\mathbf{r} = (r_x, r_y, r_z)^{\mathrm{T}}$, in which $\|\mathbf{r}\|_2 = 1$.

The Levenberg-Marquardt minimization algorithm is then used to find the best quantities for the set of values $\{x, y, z, r_x, r_y, r_z, \theta\}$ by minimizing the errors between the expected location of the object points and the detected location of these points within the scene. The Levenberg-Marquardt algorithm is a good choice because the bodies being tracked are not perfectly rigid, but minimum error approximations can still be found.

3.3.1 Rigid Body Description and Initialization

To initialize a rigid body for tracking, we are given a set of points $\mathbf{P} = \{\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_N}\}$. These points should be from a baseline frame and should represent a body that should not deform much throughout the tracking process. The location $\mathbf{x}$ of the rigid body is set to be the mean of all the points in the set $\mathbf{P}$. Thus, $\mathbf{x} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{p_i})$. The initial rotation axis, $\mathbf{r}$, is completely arbitrary, although for our implementation, it was set to $\mathbf{r} = (1, 0, 0)$. The initial angle of rotation, θ, is set to 0.

The points, $\mathbf{P}$, attached to the rigid body need to be transformed so they are expressed in relation to the location of the rigid body, $\mathbf{x}$, as opposed to the world reference frame. This is necessary later when finding the error between the expected locations of the points and their actual locations in the scene. Thus, the representation of the rigid body contains the set of points $\mathbf{P'} = \{\mathbf{p'_1}, \mathbf{p'_2}, \ldots, \mathbf{p'_N}\}$, such that all

$\mathbf{p_i'} = \mathbf{p_i} - \mathbf{x}$.

3.3.2 Rigid Body Error Calculation

The Levenberg-Marquardt algorithm iteratively minimizes a given value determined by an input vector. In the system presented here, the input vector to the system is shown in:

$$\mathbf{v} = (x, y, z, r_x, r_y, r_z, \theta)^{\mathrm{T}}. \tag{1}$$

The value we are trying to minimize is $err(\mathbf{v}, \hat{\mathbf{P}})$, which is the error associated with a given state vector $\mathbf{v}$ and an observed set of points $\hat{\mathbf{P}}$.

For a given point $\mathbf{p_i'}$ in the set of attached points $\mathbf{P'}$, the expected location of the point in the world coordinate system based upon the state vector $\mathbf{v}$ is shown in:

$$\mathbf{p_i}(\mathbf{v}) = \mathbf{R}(\mathbf{v}) * \mathbf{p_i'} + \mathbf{T}(\mathbf{v}). \tag{2}$$

In this case, the translation vector is $\mathbf{T}(\mathbf{v}) = (x, y, z)^{\mathrm{T}}$ and the rotation matrix $\mathbf{R}(\mathbf{v})$ are calculated as shown in Equation (3). Note that before $\mathbf{R}(\mathbf{v})$ is calculated, the values in $\mathbf{r} = (r_x, r_y, r_z)^{\mathrm{T}}$ are normalized such that $\mathbf{r} = \mathbf{r}/\|\mathbf{r}\|_2$ because a unit vector is required for calculating the rotation matrix.

$$
\begin{aligned}
R_{11}(\mathbf{v}) &= \cos\theta + r_x^2(1-\cos\theta) \\
R_{12}(\mathbf{v}) &= r_x r_y(1-\cos\theta) - r_z\sin\theta \\
R_{13}(\mathbf{v}) &= r_x r_z(1-\cos\theta) + r_y\sin\theta \\
R_{21}(\mathbf{v}) &= r_x r_y(1-\cos\theta) + r_z\sin\theta \\
R_{22}(\mathbf{v}) &= \cos\theta + r_y^2(1-\cos\theta) \\
R_{23}(\mathbf{v}) &= r_y r_z(1-\cos\theta) - r_x\sin\theta \\
R_{31}(\mathbf{v}) &= r_x r_z(1-\cos\theta) - r_y\sin\theta \\
R_{32}(\mathbf{v}) &= r_y r_z(1-\cos\theta) + r_x\sin\theta \\
R_{33}(\mathbf{v}) &= \cos\theta + r_z^2(1-\cos\theta) \\
\mathbf{R}(\mathbf{v}) &= \begin{pmatrix} R_{11}(\mathbf{v}) & R_{12}(\mathbf{v}) & R_{13}(\mathbf{v}) \\ R_{21}(\mathbf{v}) & R_{22}(\mathbf{v}) & R_{23}(\mathbf{v}) \\ R_{31}(\mathbf{v}) & R_{32}(\mathbf{v}) & R_{33}(\mathbf{v}) \end{pmatrix}. \qquad (3)
\end{aligned}
$$

If the measured position of point $\mathbf{p_i}$ in the current frame is $\hat{\mathbf{p}}_\mathbf{i}$, the error for this point is then shown in:

$$err(\mathbf{v}, \hat{\mathbf{p}}_\mathbf{i}) = \|\hat{\mathbf{p}}_\mathbf{i} - \mathbf{p_i}(\mathbf{v})\|_2^2. \qquad (4)$$

The total error for the state associated with vector $\mathbf{v}$ is then shown in:

$$err(\mathbf{v}, \hat{\mathbf{P}}) = \sum_{i=1}^{N} err(\mathbf{v}, \hat{\mathbf{p}}_\mathbf{i}). \qquad (5)$$

3.3.3 Levenberg-Marquardt Iteration

The goal is then to find the state vector $\mathbf{v}$ which will minimize the error $err(\mathbf{v})$ in a given frame. At every iteration, the Jacobian of $err(\mathbf{v}, \hat{\mathbf{P}})$ with respect to $\mathbf{v}$ needs to be calculated (as the observation $\hat{\mathbf{P}}$ is constant for a given frame). The Jacobian $\mathbf{J}$ is calculated as shown in:

$$\mathbf{J}(\mathbf{v}, \hat{\mathbf{P}}) = \begin{pmatrix} \frac{derr(\mathbf{v},\hat{\mathbf{p_1}})}{dx} & \frac{derr(\mathbf{v},\hat{\mathbf{p_1}})}{dy} & \cdots & \frac{derr(\mathbf{v},\hat{\mathbf{p_1}})}{d\theta} \\ \frac{derr(\mathbf{v},\hat{\mathbf{p_2}})}{dx} & \frac{derr(\mathbf{v},\hat{\mathbf{p_2}})}{dy} & \cdots & \frac{derr(\mathbf{v},\hat{\mathbf{p_2}})}{d\theta} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{derr(\mathbf{v},\hat{\mathbf{p_n}})}{dx} & \frac{derr(\mathbf{v},\hat{\mathbf{p_n}})}{dy} & \cdots & \frac{derr(\mathbf{v},\hat{\mathbf{p_N}})}{d\theta} \end{pmatrix}. \tag{6}$$

We define the error vector for the current state to be $\mathbf{e}(\mathbf{v}, \hat{\mathbf{P}})$, specified in:

$$\mathbf{e}(\mathbf{v}, \hat{\mathbf{P}}) = (err(\mathbf{v}, \hat{\mathbf{p_1}}), err(\mathbf{v}, \hat{\mathbf{p_2}}), \ldots, err(\mathbf{v}, \hat{\mathbf{p_N}}))^{\mathrm{T}}. \tag{7}$$

If λ is the heuristically chosen damping factor used by the Levenberg-Marquardt algorithm and $\mathbf{I}$ is the identity matrix, the state adjustment for this iteration then shown in Equation (8). For this application, the damping factor was initialized to $\lambda = 0.5$.

$$\mathbf{q} = -(\mathbf{J}(\mathbf{v}, \hat{\mathbf{P}})^{\mathrm{T}}\mathbf{J}(\mathbf{v}, \hat{\mathbf{P}}) + \lambda\mathbf{I})^{-1}\mathbf{J}(\mathbf{v}, \hat{\mathbf{P}})^{\mathrm{T}}\mathbf{e}(\mathbf{v}, \hat{\mathbf{P}}). \tag{8}$$

The new state vector based on this adjustment is $\mathbf{v}' = \mathbf{v} + \mathbf{q}$. The current and adjusted residuals, $err(\mathbf{v}, \hat{\mathbf{P}})$ and $err(\mathbf{v}', \hat{\mathbf{P}})$ respectively, are calculated. If $err(\mathbf{v}', \hat{\mathbf{P}})$ is below a minimum error threshold, or $\|\mathbf{q}\|_2$ is below a minimum change threshold, the state $\mathbf{v}'$ is a good final approximation of the position of the tracked body, so the process is stopped for this frame. Otherwise, the iteration must be performed again. If $err(\mathbf{v}, \hat{\mathbf{P}}) > err(\mathbf{v}', \hat{\mathbf{P}})$, the overall error has increased, meaning $\mathbf{v}'$ is *not* an improvement over $\mathbf{v}$. In this case, the damping factor λ is reduced such that $\lambda = \lambda/2$ and the state vector $\mathbf{v}$ is left unchanged. Otherwise, the error has decreased so the damping factor λ is increased such that $\lambda = 2\lambda$, and the state vector $\mathbf{v}$ for the next iteration is updated such that $\mathbf{v} = \mathbf{v}'$.

Figure 2: An anthromorphic phantom mannequin. These are used in place of real people in radiation therapy research. A phantom's internal electron density closely matches that of a real person.

Table 1: MVCT versus stereoscopic translation detection (mm)

Platform	MVCT μ	MVCT σ	Vision μ	Vision σ
0.0	0.1	0.0	0.3	0.3
10.0	10.0	0.1	10.1	0.4
30.0	30.1	0.0	30.1	0.5
50.0	49.3	0.6	50.4	1.0

3.4 Experiments and Results

3.4.1 Stereoscopic Marker Detection Experiment

In this experiment, the accuracy of the near-infrared stereoscopic marker detection system is compared to the MVCT scan-based patient positioning system currently being used for patient repositioning. The MVCT scan collects volumetric data about the person or object scanned. A current MVCT scan can then be matched to the previous kVCT scan using closed-box image registration methods to determine how much movement has taken place from one scan to the next. The software produced by Tomotherapy Inc. includes several registration methods to choose from, which seem to vary based on the weights given to dense bony material versus less dense

(a) (b)

Figure 3: (a) Close-up of the reflective markers used by the stereoscopic marker detection system. (b) Reflective markers positioned on the phantom's head for an experiment.

fleshy material. The results for these different registrations are typically measurably different.

For the experiment, a anthromorphic phantom is positioned on the moving platform on the helical tomotherapy machine. An anthromorphic phantom as shown in Figure 2 is essentially a mannequin whose internal radiation interaction closely matches that of a real person. They are used extensively in the testing of radiation therapy methods. Near-infrared reflective markers are placed on the phantom as shown in Fig. 3. The platform was then moved a set amount, which in turn moves the phantom the same amount. Typically, initial positioning is checked using an MVCT scan. There are three different registration methods currently used that are performed on the MVCT scans to determine translation: Bone, Bone and Tissues, and Full Image. These three methods are unfortunately proprietary and closed source. For the stereoscopic system, three markers were used, positioned on the phantom's head.

The difference recorded by the MVCT image registration methods as well as the stereoscopic marker location differences are then recorded. See Table 1 for the results.

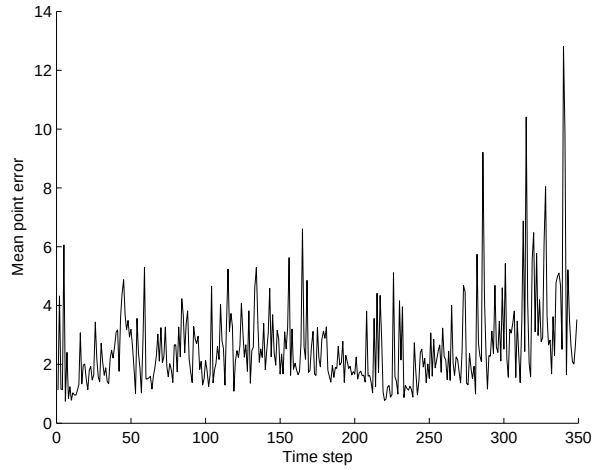


Figure 4: Plot of the mean point error on per frame for a normal run of the rigid-body tracking experiment.

This table records how far the platform was instructed to move (Platform), the mean translation detected by the MVCT methods (MVCT μ), the standard deviation of these values (MVCT σ), the mean translation detected by the vision system of all markers (Vision μ), and the standard deviation of these values (Vision σ). As can be seen in the table, the stereoscopic marker detection system detects roughly the same translation as the MVCT methods, although the variance is higher. These results are encouraging as they suggest that a vision system should be able to perform under the tolerances necessary to detect small changes in a patient.

It is worth noting that the stereoscopic marker detection system can be used while treatment is ongoing. This by itself is a benefit over the MVCT scan, which cannot be used simultaneously with treatment. The markers can also be positioned 60 times per second, versus the 10 to 20 minutes it takes to perform a single MVCT scan on a 10 cm portion of a patient.

3.4.2 Rigid-Body Tracking Experiment

To test the rigid body tracker, simulated data was used so that there would be ground truth data with which to compare the tracking. A spherical object, with 10 random, statically located points on it is simulated rotating and translating through space.

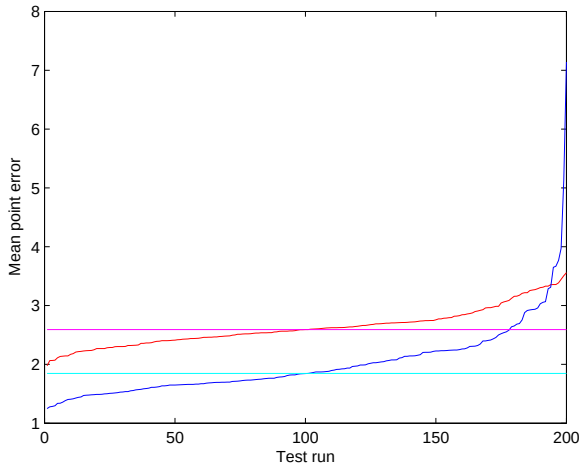


Figure 5: Plot of the mean point error per run from the rigid-body tracking experiment. The red function shows the mean point error per run using noise corrupted data, while the median value of this function is shown in magenta. The blue function shows mean point error per run using non-corrupted data, while the median value of this function is shown in cyan.

The sphere is 150 mm in diameter. This size was chosen as it is roughly the size of the human head, and thus should provide a good analog. It translates approximately 1500 mm, and rotates 3π radians in the course of 350 frames.

The locations of the points at every frame are corrupted by Gaussian white noise with a 0.5 mm standard deviation before being tracked by the rigid body tracker. This standard deviation was chosen based on the expected standard deviation for reasonable small measurements from the stereoscopic marker detection experiment in Section 3.4.1. For this experiment, the points on which the point tracker is initialized are corrupted as well. Figure 4 shows the mean point error per frame between the predicted object points and the real, non-corrupted points for a single run of this test. A total of 200 runs were performed. Figure 5 shows the total mean point error for each of these 200 runs in red. For comparison purposes, the total mean point error for a second tracker, which tracks non-corrupted data, is presented in blue. The sets of values for each have been arranged in ascending order. Overall, the median mean error per point across all experimental runs for the noise corrupted data is 2.6 mm, and the median error per point across all experimental runs for the non-corrupted

data is 1.8 mm.

In Figure 5, it can be seen that there are tails on either end of the distribution, with the tails at the high end being the greatest in magnitude. This is due primarily to places when the tracking strayed off and did not get back for many frames. Examples of this can be seen in Figure 4 where there are some periods where the error is high before coming back down. Because the rigid body tracking is seeded by the previous known location of the rigid body, tracking in the current frame can be thrown off by inconsistencies encountered in previous frames. As the only differences between runs of the experiment are the locations of the selected points on the sphere, it is likely that the selection of point locations on the body being tracked has a great different deal of influence on the ability to reliably track the object.

It can also be seen in Figure 5 that the median corruption error for tracking the rigid bodies is 2.6 mm. Thus, the desired 5 mm motion detection threshold for the helical tomotherapy system is likely realistic for rigidly tracked portions of the body.

3.5 Preliminary Work Final Thoughts

In this chapter, we have laid the groundwork for a vision-based patient tracking system for use with the helical tomotherapy treatment. We have shown that a near-infrared stereoscopic marker detection method can detect changes on par with those of the currently used MVCT scan based method. In addition, we have shown that boney portions of the body can be accurately tracked in three dimensions. The results of our experiments indicate that a vision-based system should perform well enough to reliably detect 5 mm movement in the patient.

The benefits of this type of system are numerous. First, a vision-based patient tracking system can be used while treatment is underway, and issue alerts if a patient moves significantly—a capability that is not currently present. Second, the vision-based acquisition can be performed very quickly compared to the MVCT scan method currently used for initial patient positioning. Third, portions of the body such as the head and chest can be tracked individually of each other, aiding doctors by providing them with information about how much each portion of the body is out-of-alignment as opposed to a single measure for the patient's entire body.

4 Breathing Investigation

Breathing is a periodic motion of the torso as a person inhales and exhales as part of respiration. In vision terms, breathing can be described as a periodic motion of a nonrigid surface. The deformation in the body surface from the breathing is non-negligible. Breathing cannot be stopped for extended periods of time, making it an important phenomenon to track in terms of whole-body tracking.

In this chapter, a single scan-line laser rangefinder surface scan of a person breathing is used to show that non-marker based sensing apparatuses can differentiate between normal and abnormal breathing. Anomalies in breathing include a complete shifts of body position, either gross or slight, coughing, and other unnatural motion of the body.

Monitoring breathing is useful for conformal radiation treatments. By monitoring where in their breathing cycle a patient is, radiation sources can be turned on and off as the breathing motion puts portions of their body in and out of range. This idea is discussed in the work of Fox, *et al.* [26], although there are limitations to the work they present. The motion model they used is somewhat crude and only two points on the patient are tracked, using physically attached markers. Simple thresholding is what they used determine if the markers are out of ideal position. Denser sensing should be able to provide more nuanced and complete data about the patient's body movement.

4.1 Equipment Setup

The equipment is setup as follows. A Hokuyo laser rangefinder is attached to a support rig approximately 18 inches above the patient such that it's field-of-view is across the chest of the patient. The laser rangefinder has a scanning frequency of 10 Hertz. The range of the laser rangefinder is between 0.2 and 4.0 meters. The raw data from the laser rangefinder is a vector of distance values from the center of its

emitter, in half-degree increments. It has a 270 degree field of view. In short, the laser rangefinder is adequate for collecting a single scan line across the chest of a patient ten times per second.

4.2 Analysis of Breathing Data

There are several steps in analyzing the data collected from the laser rangefinder sensors, as outlined below.

1. Smooth the raw laser rangefinder data by convolving it with a 1D Gaussian kernel to remove noise.
2. Convert the raw laser rangefinder data to cartesian world coordinates.
3. Determine the base from which the Hausdorff distance function for the scanning period will be computed.
4. Compute the Hausdorff distance function for the scanning period. See Figures 6 and 7 for example Hausdorff distance functions from real data.
5. Smooth the Hausdorff distance function by convolving it with a 1D Gaussian kernel to remove noise.
6. Compute the power spectrum of the Hausdorff distance function if desired. See Figures 8 and 9 for example power spectra from the Hausdorff distance functions in Figures 6 and 7.

4.2.1 Converting Scanner Data to World Coordinates

Converting raw laser rangefinder measurements for a single point is simply the straightforward conversion from polar to cartesian coordinates, as shown in Equation (9), where ϕ is the angle of the current scan point and d is the distance it registers.

$$\begin{aligned} x &= d\cos\phi \\ y &= d\sin\phi. \end{aligned} \tag{9}$$

4.2.2 Hausdorff Distance

The Hausdorff distance is a metric used to measure the distance between two curves. It is very useful in this domain as it can be used to compare the position of the chest scan lines between a frame at any given time and the position of a base frame. This, in turn, can be used to gain useful information about what is occurring with the patient.

Given two curves, the Hausdorff distance can be stated as the maximum of the set of minimum distances from every point in the first curve to every point in the second curve. It can be described mathematically as follows. Consider two curves, denoted $\mathbf{F}$ and $\mathbf{G}$. Assume $\mathbf{F}$ has m elements and $\mathbf{G}$ has n elements. The minimum distance from a single point in $\mathbf{F}$, denoted F_i, to all n points in $\mathbf{G}$ can be denoted $d_{min}\left(F_i, \mathbf{G}\right)$. Thus, the Hausdorff distance, d_H, is described in:

$$d_H = \max_i \left\{d_{min}(F_i, \mathbf{G})\right\}. \tag{10}$$

Figures 6 and 7 show the Hausdorff Distance function at every point in time for many real sets of collected breathing data.

4.2.3 Base Frame for Hausdorff Distance

In the calculation of the Hausdorff distance functions, as shown in Figures 6 and 7, a base frame must be chosen from which the distance to every frame is calculated. In the cases shown here, the base frame is chosen to be the frame for which the sum of the raw laser rangefinder elements is maximized. In other words, the base frame is chosen to be the frame in which the patient's chest is farthest from the laser rangefinder.

The selection of the base frame is important because a good selection will give normal breathing motion a plot with many regular curves. With a poor selection of the base frame, the plot looks off-center and irregular, even though the underlying data is regular. By choosing as the base frame the frame in which the chest of the patient is as far away from the laser scanner as possible, it is more likely that regular data will appear regular in the plot.

4.2.4 Power Spectrum of the Hausdorff Distance Function

Breathing is a periodic motion. Therefore, the power spectrum of the Hausdorff distances between different frames should encapsulate this information. The power spectrum is calculated using the Fast Fourier Transform of the Hausdorff distance function.

4.3 Data Comparisons

Several sets of data were taken of actual people breathing using the laser rangefinder. Figure 6 shows normal breathing data for three people, with two three-minute periods each. These plots tend to show very regular breathing periodicity, even if the amplitude varies in a couple of the examples. This amplitude variation can be due to breathing more shallowly or deeply.

In contrast, Figure 7 shows normal breathing punctuated by coughing behavior by the same three people in the first figure. These periods of abnormal breathing can be seen throughout these plots.

Finally, Figure 8 shows the power spectra of the normal breathing periods shown in Figure 6. Similarly, Figure 9 shows the power spectra of the abnormality-punctuated breathing periods shown in Figure 7. From these plots, it appears the power spectrum is not a good tool to determine the overall periodicity of breathing data of this type. For both the normal and punctuated breathing, there is a lot of low-frequency activity in the power spectrum, which makes it difficult to use to separate the two classes of breathing activity.

4.4 Final Thoughts

Laser range finders can be used to record a single scan line of data about a patient's breathing over time. The examples shown in Figures 6 and 7 demonstrate that, by using the Hausdorff distance, breathing periods can be clearly seen, and even abnormalities can be delineated on inspection. Unfortunately, it is difficult to systematically pull useful information about the presence of abnormalities out of this data using the power spectrum. Overall, the Hausdorff distance functions for both

normal and abnormal breathing contain a lot of noisy, low-frequency activity, causing the signal to get lost in the overall noise.

This investigation shows that sensors can collect dense, useful information about a patient without the use of physically attached markers. The laser rangefinder used here is limited to only a single scan line of data, collected ten times per second. In the next chapter, a more data rich sensing modality is explored–structured light.

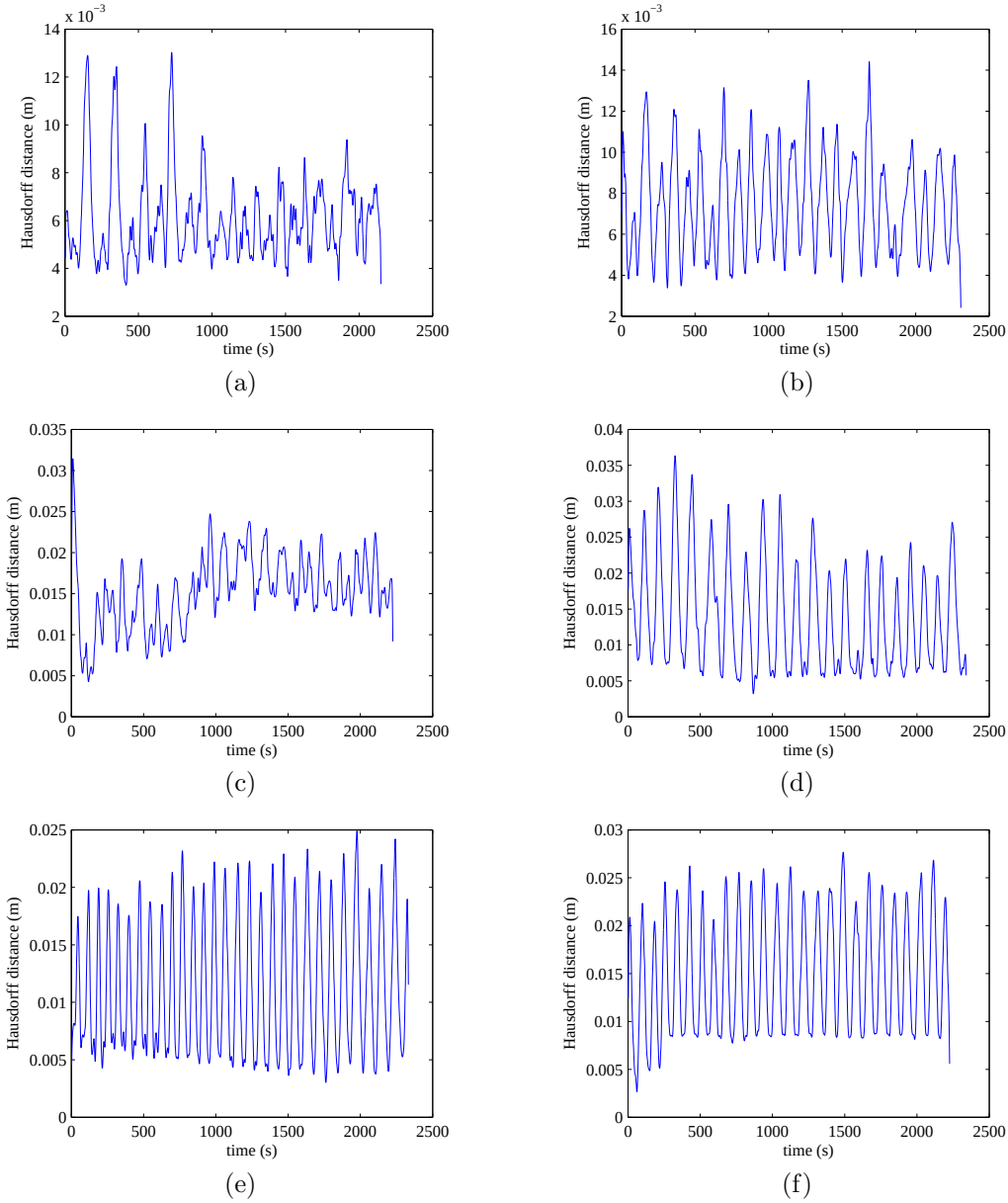


Figure 6: Hausdorff distance plots for normal breathing. (a, b) Subject 1 (c, d) Subject 2 (e, f) Subject 3.

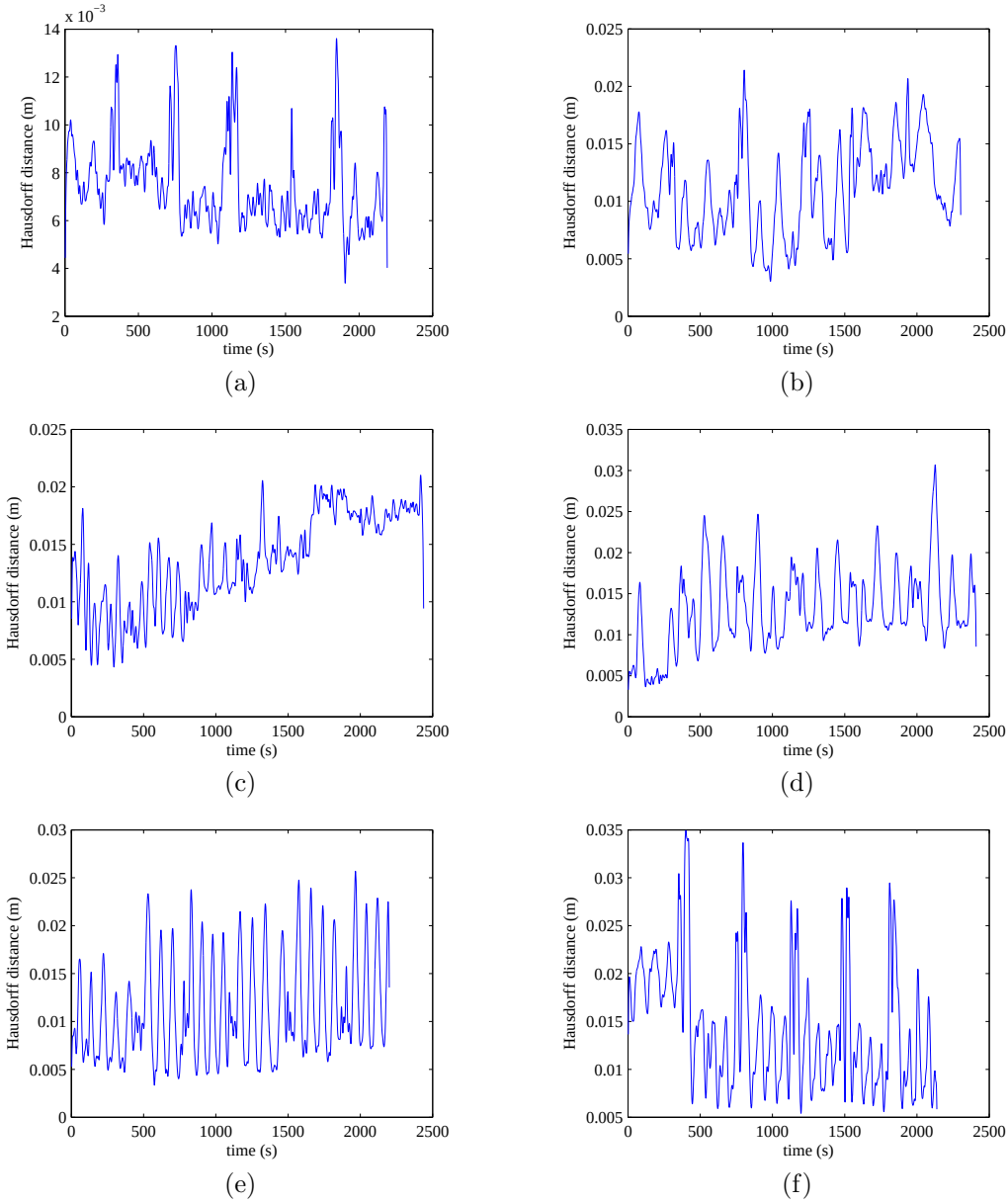


Figure 7: Hausdorff distance plots for normal breathing periodically interrupted by bouts of coughing. (a, b) Subject 1 (c, d) Subject 2 (e, f) Subject 3.

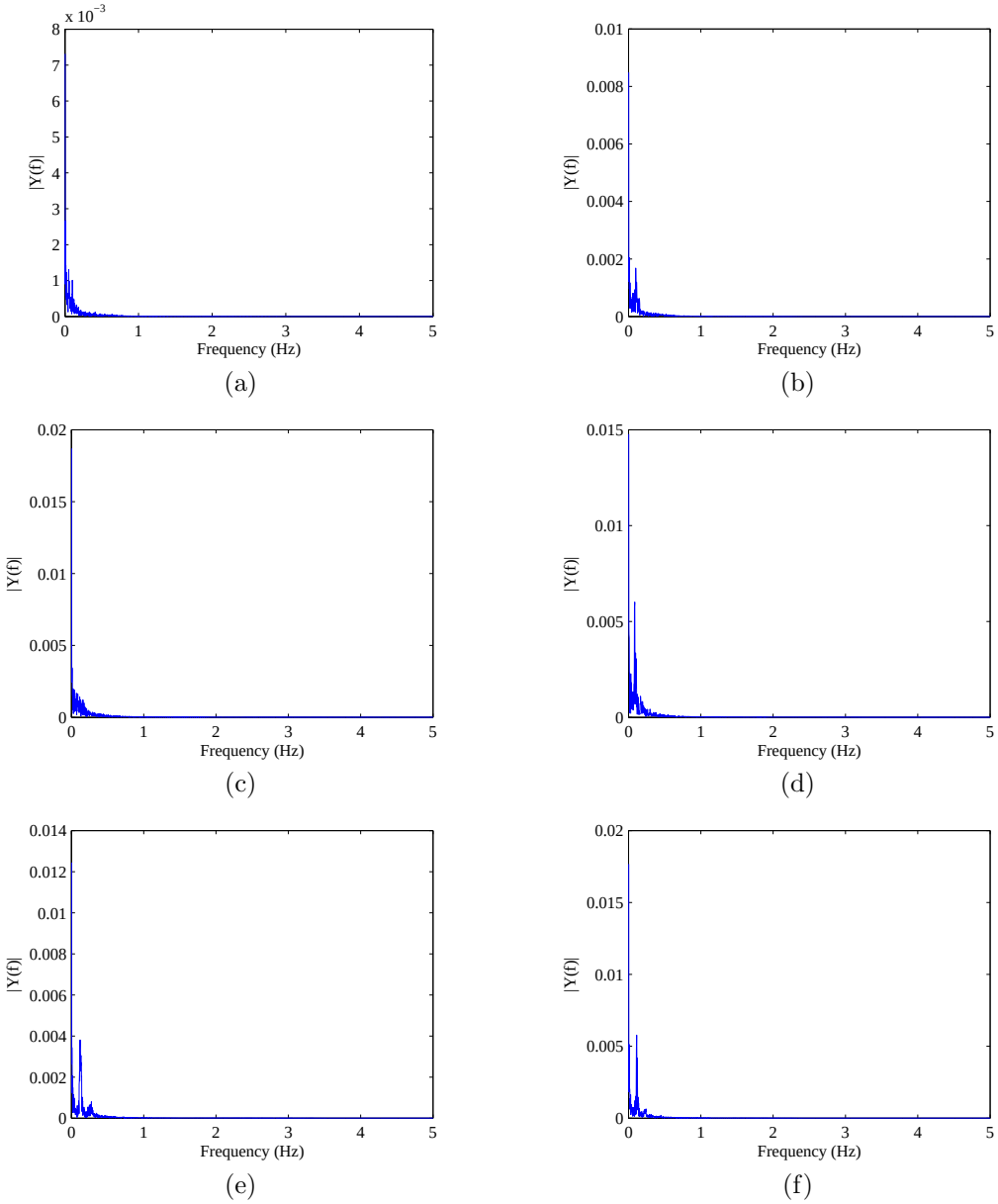


Figure 8: Power spectra of Hausdorff distance plots for normal breathing. (a, b) Subject 1 (c, d) Subject 2 (e, f) Subject 3.

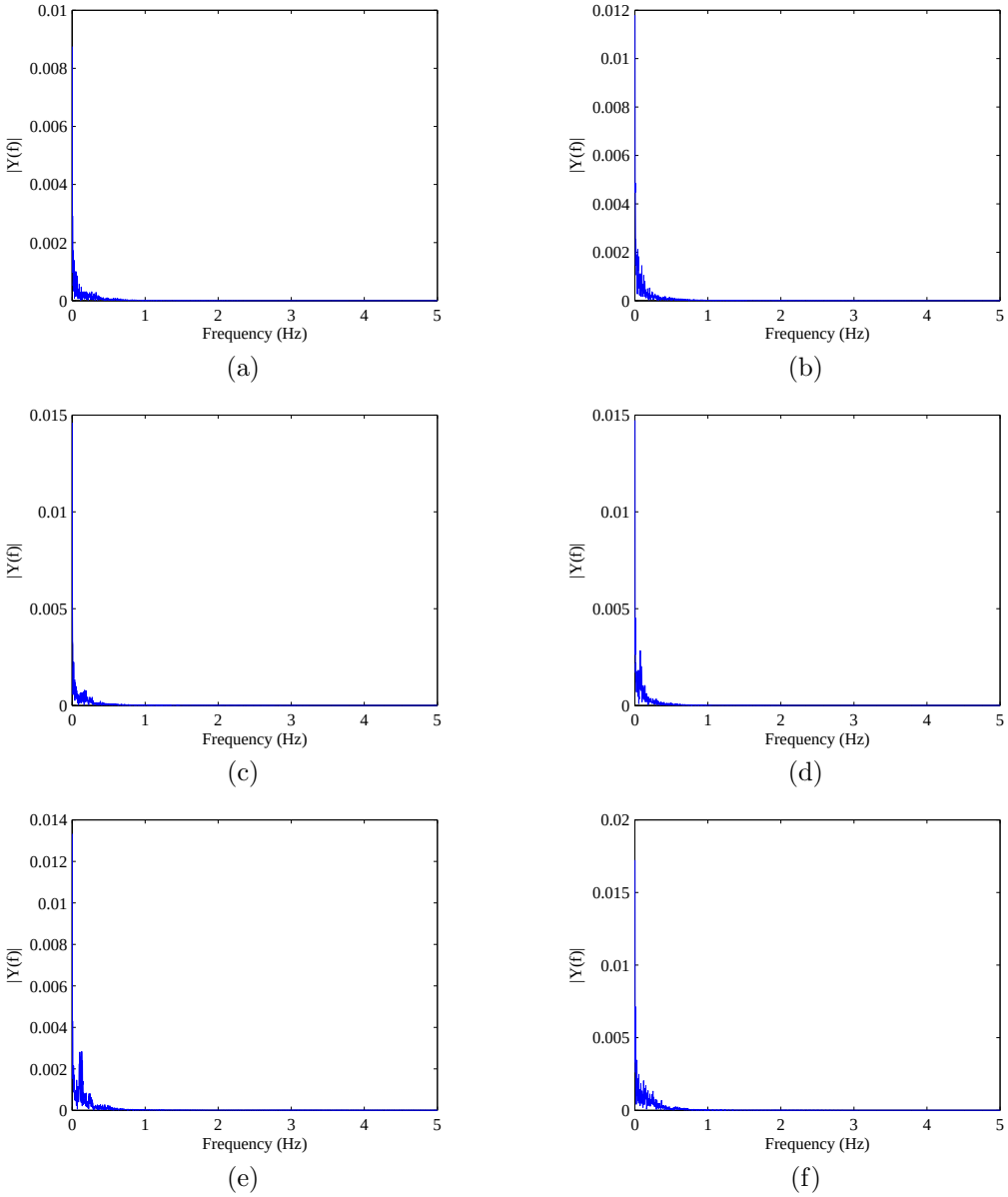


Figure 9: Power spectra of Hausdorff distance plots for normal breathing periodically interrupted by bouts of coughing. (a, b) Subject 1 (c, d) Subject 2 (e, f) Subject 3.

5 Structured Light Background

The work presented in Chapters 3 and 4 demonstrate the efficacy of applying vision techniques to monitoring patient body motion. However, there are still issues to be resolved, mainly revolving around the use of markers. They are still somewhat invasive and they need to be properly repositioned on the exact same location every time the scan is run to accurately track exact body regions. One way to improve upon these areas is to forego the use of the artificial near-infrared reflective markers and look into non-marker-based vision methods. The solution examined in detail in the remainder of this thesis is structured light.

5.1 Example Structured Light System for Patient Body Tracking

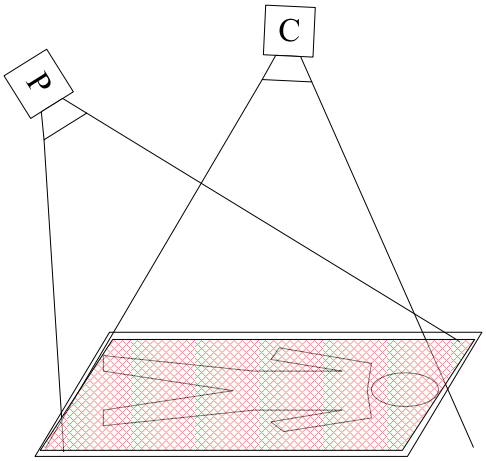


Figure 10: Example setup for a structured light-based patient body tracking system.

One possible way to create a device which ascertains the patient position and articulation in a completely non-intrusive manner that foregoes utilizing artificial markers

is to use a structured light vision system. A depiction of how such a system may be setup is shown in Figure 10. Using structured light techniques, a set of known patterns can be projected using the projector onto the patient. The patterns from the projector will intersect the patient and the rest of the scene, and a set of images will be captured by the camera. Using the system calibration, a 3D reconstruction of the patient's body surface can then be reconstructed using the algorithms described in Section 5.2. Computation of the 3D depth information is computed in a conceptually similar way to stereoscopic camera systems, only the correspondence problem is greatly simplified. Medical applications are well suited for structured light as a 3D registration modality as the environment and lighting can be strictly controlled.

A surface modeling algorithm would then be used with the detected surface data points to determine the location and articulation of the patient's body, including probabilities on locations for subsurface body regions like bones, organs, or other areas within the body that are of interest for the radiation procedure.

Because this hypothetical system is intended to provide feedback to the medical practitioners positioning their patients, the projector used for structured light should also be able to provide this feedback. Projecting the results on patient body positioning directly back onto the patient using the same projector that was used in the 3D registration procedure makes a lot of sense in this case, as this presentation method conveys the information intuitively. This is a version of augmented reality, as it adds virtual information to the real world. This feedback would be done similarly to projector-based augmented reality systems–systems where a computer controls a projector to add extra information to a real-world object. For instance, a pattern could be projected that would only satisfy some mathematical criteria if the patient was in the proper position. Alternatively, perhaps a red pattern could be used to indicate out-of-position portions of the patient, while a green pattern could indicate in-positions portions of the patient. By providing this feedback to the medical practitioners, patient posture adjustments can be performed quickly and more reliably than the current initial placement procedure. After the patient is positioned, the system can then be set to quickly monitor for changes in the patient's placement and articulation while the helical tomotherapy device is running.

Figure 11 shows a block diagram outlining the portions of the hypothetical system

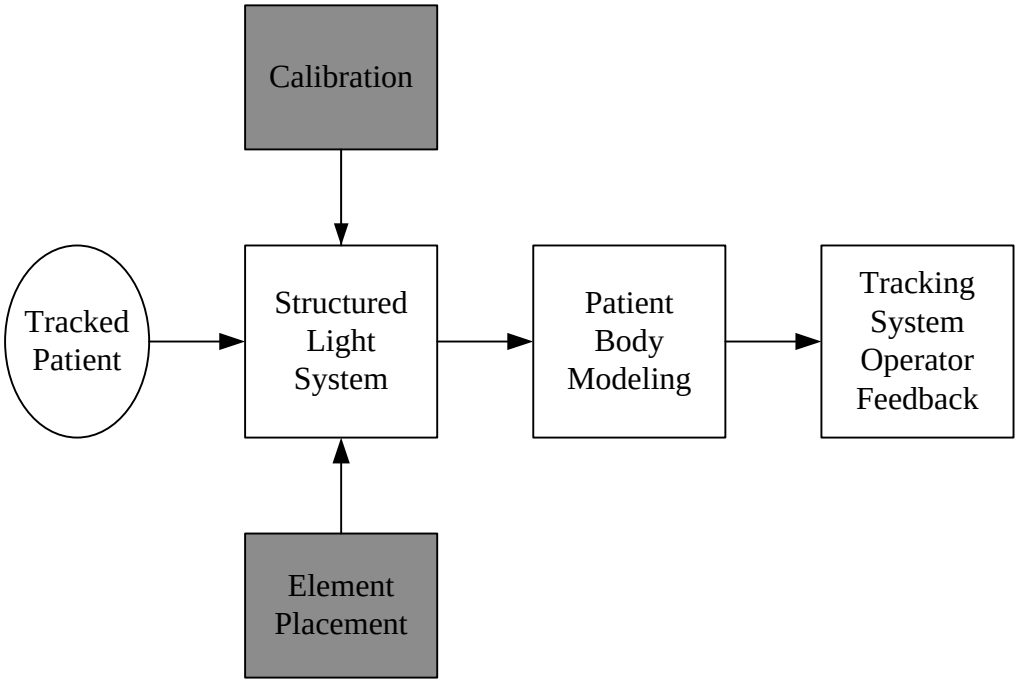


Figure 11: Block diagram of a hypothetical structured light-based patient body tracking system. The portions of the system which this thesis addresses are shaded.

outlined here. The shaded blocks are the portions addressed by the remainder of this thesis. The remainder of the this thesis addresses only issues associated with the structured light system. Surface modeling and augmented reality work are left for later work, and are discussed more in Section 8.1.

5.2 Structured Light Mathematics

In this section, we briefly walk through the mathematics behind structured light vision, highlighting similarities to stereoscopic vision.

5.2.1 Basic Multiview Geometry

Projection Matrix Elements The projection matrix $\mathbf{P}$ is standard method by which a (pinhole) camera is described mathematically. It is the main quantity about the camera which is important for extracting information from captured images.

The homogenous camera projection matrix $\mathbf{P}$ is a 3×4 matrix which maps coordinates in the world to coordinates in the camera's image plane. Another important quantity is the four-term vector $\mathbf{c}$ which represents the camera center in homogenous world coordinates. The camera center can be extracted from $\mathbf{P}$, however (more on this later). Take a projection matrix $\mathbf{P} = [\mathbf{Q}|\mathbf{t}]$, where $\mathbf{Q}$ is a 3×3 matrix, and $\mathbf{t}$ is a three vector. Assuming $\mathbf{Q}$ is invertible, the camera center is then $\mathbf{c} = ((-\mathbf{Q}^{-1}\mathbf{t})^\top, 1)^\top$.

Another form of the equations for $\mathbf{P}$ is shown in:

$$\mathbf{P} = [\mathbf{KR} | -\mathbf{KRc}]. \tag{11}$$

In Equation (11), $\mathbf{c}$ is the inhomogeneous three vector describing the camera center in the world frame, $\mathbf{R}$ is the 3×3 rotation matrix describing the rotation from the world reference frame to the camera's 3D reference frame, and $\mathbf{K}$ is the matrix which maps points in the camera's world frame to the camera's image plane. This final matrix, $\mathbf{K}$, contains the camera's intrinsic parameters and takes the form shown in:

$$\mathbf{K} = \begin{bmatrix} \alpha_x & s & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{12}$$

The parameters of $\mathbf{K}$ in Equation (12) all have specific, real-world meaning. First is the scale factors in the x- and y-coordinate direction, α_x and α_y respectively. These basically scale the location on the world image plane to the pixels in the image plane. The skew is represented by s, and should be zero for most cameras. If the skew is non-zero, the camera sensor elements are arranged such that the x- and y-axes are not perpendicular. Finally, $(x_0, y_0)^\top$ are the coordinates of the principal point of the image. Essentially, this is the offset from $(0, 0)^\top$ to the center of the image plane.

Inverse Projection Matrix The inverse mapping from the camera image to the world is sometimes needed as well. This mapping can be written as $\mathbf{P}^{-1}$, which is *not* the pseudoinverse of $\mathbf{P}$, although $\mathbf{PP}^{-1} = \mathbf{I}$. Equation (13) shows how $\mathbf{P}^{-1}$ can be computed.

$$\mathbf{P}^{-1} = \begin{bmatrix} \mathbf{R}^{-1} & \mathbf{c} \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} & \mathbf{K}^{-1} & \\ 0 & 0 & 1 \end{bmatrix}. \tag{13}$$

Projecting Points with the Projection Matrix The first common function examined is how a point in the world is mapped to a point on the image plane. This operation can be written compactly as $\mathbf{x}_i = \mathbf{P}\mathbf{x}_w$, where $\mathbf{x}_i$ is the three-element homogenous coordinate vector for the point in the image plane while $\mathbf{x}_w$ is the four-element homogenous coordinate vector for the point in the world. Given this, the world coordinate of points detected in the image plane can be determined.

Going in the opposite direction, a point in the image plane maps to a ray in the world. This is because a point in the image only has two degrees of freedom, as does the ray in the world. In 3D, a ray can be specified as the join of two points. The two points in the world used to describe this ray can thus be the camera center $\mathbf{c}$, and the point $\mathbf{x}_w = \mathbf{P}^{-1}\mathbf{x}_i$, where $\mathbf{x}_i$ is the point of interest in the camera's image plane. Thus, points on the ray in the world satisfy Equation (14) for arbitrary real values of λ.

$$\mathbf{x}_w(\lambda) = \mathbf{c} + \lambda(\mathbf{x}_w - \mathbf{c}). \tag{14}$$

Essentially, whatever feature registered as a point in the camera's image plane must exist on the ray projected from that point out into the word. With a single camera, it is not possible to determine where along this ray the actual world feature lies, but it is somewhere on the ray.

Decomposition of the Projection Matrix The projection matrix $\mathbf{P}$ can also be decomposed into its constituent components by noting that $\mathbf{KR} = \mathbf{P}_{3\times3}$, where $\mathbf{P}_{3\times3}$ is the first 3×3 elements of $\mathbf{P}$. $\mathbf{P}_{3\times3}$ is decomposed into $\mathbf{K}$ and $\mathbf{R}$ using the RQ-decomposition. RQ-decomposition works because $\mathbf{R}$ is a rotation matrix, and $\mathbf{K}$ is upper triangular.

After RQ-decomposition is performed, a regularization step is needed because the diagonal elements of $\mathbf{K}$ must be of the same sign. Essentially, a matrix $\mathbf{H}$ is used, which has the property $\mathbf{HH} = \mathbf{I}$. Essentially, there are three choices $\mathbf{H}$, which

correspond to each permutation of the 3×3 identity matrix in which the sign is reversed on two elements on the diagonal. The corrected matrices are then found as $\mathbf{K}' = \mathbf{KH}$ and $\mathbf{R}' = \mathbf{HR}$.

The camera center $\mathbf{c}$ can be found as the solution to the set of linear equations $\mathbf{Pc} = \mathbf{0}$.

5.2.2 Light Striping

To examine structured lighting, we introduce the concept of light striping. Light striping is an established vision methodology in which a single stripe of light is projected onto a scene. Suppose we have a system which consists of a projector and a camera pointed at a scene such that the image projected by the projector is visible by the camera. The projector is then used to project a light stripe onto the scene. This stripe illuminates a single 2D plane in the world. Thus, all points in the scene illuminated by the projected stripe lie on the same 2D plane. If a camera can see these illuminated points, their world locations can be determined assuming the projection matrices for both the camera and projector are known. Light striping corresponds to the case in Fig. 12(b). Projecting any pattern more complicated than a line, as in Fig. 12(c), results in non-uniquely solvable systems.

Determining the Light Stripe Plane The equation for a plane is $\pi_1 x_1 + \pi_2 x_2 + \pi_3 x_3 + \pi_4 = 0$ where x_1, x_2, and x_3 are the spatial dimensions and π_1, π_2, π_3, and π_4 are known coefficients. In homogenous coordinates, this can be written in dot product notation as $\pi \cdot \mathbf{x}_w = \mathbf{0}$ where $\mathbf{\Pi} = (\pi_1, \pi_2, \pi_3, \pi_4)^\top$ is the homogenous representation of the plane and $\mathbf{x}_w = (x_1, x_2, x_3, 1)^\top$ is the homogenous representation of a point in the world which lies on the illuminated plane.

The coefficients π_1, π_2, π_3, and π_4 can be found using the projector's projection matrix, $\mathbf{P}_p$, the projector's center $\mathbf{c}_p$, and the line in the projector's image plane being illuminated, $\mathbf{l}_p = (l_1, l_2, l_3)^\top$.

To define a plane, three points are needed. One of these points can be taken to be the projector's center in world coordinates, $\mathbf{c}_p$. For the other two points, choose two points $\mathbf{x}_{p1}$ and $\mathbf{x}_{p2}$, which lie on the line $\mathbf{l}$ in the projector's image plane. The points $\mathbf{x}_{p1}$ and $\mathbf{x}_{p2}$ can be selected arbitrarily on $\mathbf{l}$, but edge cases must be noted as

either l_1 or l_2, but not both, can be zero. Thus, the three world points from which the coefficients for $\mathbf{\Pi}$ can be determined using the three world points specified in:

$$\begin{aligned}
\mathbf{x}_{w1} &= \mathbf{P}_p^{-1}\mathbf{x}_{p1} \\
\mathbf{x}_{w2} &= \mathbf{P}_p^{-1}\mathbf{x}_{p2} \\
\mathbf{x}_{w3} &= \mathbf{c}_p.
\end{aligned} \tag{15}$$

The points found in Equation (15) must then be normalized such that $\mathbf{x}'_{wi} = \mathbf{x}_{wi}/x^4_{wi}$. Given these three, non-collinear points, the coefficients of $\mathbf{\Pi}$ can be determined as shown in Equation (16). Again, $\mathbf{\Pi} = (\pi_1, \pi_2, \pi_3, \pi_4)^\top$.

$$\begin{aligned}
D &= det\left(\begin{bmatrix}\mathbf{x}_{w1}^\top \\ \mathbf{x}_{w2}^\top \\ \mathbf{x}_{w3}^\top\end{bmatrix}\right) = det\left(\begin{bmatrix} x^1_{w1} & x^2_{w1} & x^3_{w1} \\ x^1_{w2} & x^2_{w2} & x^3_{w2} \\ x^1_{w3} & x^2_{w3} & x^3_{w3}\end{bmatrix}\right) \\
\pi_1 &= -\frac{1}{D}det\left(\begin{bmatrix} 1 & x^2_{w1} & x^3_{w1} \\ 1 & x^2_{w2} & x^3_{w2} \\ 1 & x^2_{w3} & x^3_{w3}\end{bmatrix}\right) \\
\pi_2 &= -\frac{1}{D}det\left(\begin{bmatrix} x^1_{w1} & 1 & x^3_{w1} \\ x^1_{w2} & 1 & x^3_{w2} \\ x^1_{w3} & 1 & x^3_{w3}\end{bmatrix}\right) \\
\pi_3 &= -\frac{1}{D}det\left(\begin{bmatrix} x^1_{w1} & x^2_{w1} & 1 \\ x^1_{w2} & x^2_{w2} & 1 \\ x^1_{w3} & x^2_{w3} & 1\end{bmatrix}\right) \\
\pi_4 &= 1.
\end{aligned} \tag{16}$$

Solving for Detected Points on the Light Stripe Plane Suppose a set of points, $\{\mathbf{x}_{c1}, \mathbf{x}_{c2}, ..., \mathbf{x}_{cn}\}$, has been found in the camera's image plane which corresponds to a set of points in the world illuminated by the light stripe. The camera's projection matrix is denoted $\mathbf{P}_c$. Each of these points has a corresponding a ray, $\mathbf{L}_i(\lambda_i)$, emanating from the camera center, $\mathbf{c}_c$, such that $\mathbf{L}_i(\lambda_i) = \mathbf{c}_c + \lambda_i(\mathbf{x}_{wi} - \mathbf{c}_c)$.

In this case, $\mathbf{x}_{wi} = \mathbf{P}_c^{-1}\mathbf{x}_{ci} = (x_1, x_2, x_3, x_4)^\top$, for the detected point $\mathbf{x}_{ci}$.

The point at which this ray intersects the light stripe plane is the point which satisfies:

$$\mathbf{\Pi} \cdot \mathbf{L}_i(\lambda_i) = 0. \tag{17}$$

Equation (17) is satisfied for the value of λ_i shown in Equation (18). Note that in Equation (18), $\mathbf{c}_c = (c_{c1}, c_{c2}, c_{c3})^\top$.

$$\lambda_i = \frac{-\pi_1 c_{c1} - \pi_2 c_{c2} - \pi_3 c_{c3} - \pi_4}{\pi_1 x_1 + \pi_2 x_2 + \pi_3 x_3 + \pi_4 x_4 - \pi_1 c_{c1} - \pi_2 c_{c2} - \pi_3 c_{c3} - \pi_4}. \tag{18}$$

Thus, the point in the world corresponding to camera image point $\mathbf{x}_{ci}$ is $\mathbf{L}_i(\lambda_i)$.

5.2.3 Light Striping Practicality

Light striping is an established way to estimate the reconstruction of a given scene. However, projecting only a single plane of light per frame makes capturing the entire scene a slow process requiring many frames. When information must be collected swiftly, the amount of information gathered per frame must be increased.

One way is to used gray codes (see Batlle, *et al.* [5]), which is a fascinating method to encode all the stripes to be scanned into fewer individual scans. An issue with this is that, much like regular stereoscopic vision, detection becomes harder. The presence of noise in camera images also worsens results for gray coded detection methods. The algorithm used to detect the changes due to changed lighting thus need more tuning to operate properly.

Another possible solution then is to project and capture multiple light planes simultaneously. Unfortunately, projecting multiple light planes per frame leads to ambiguities in reconstruction. With a single light plane, any illuminated points must be on the plane. However, if multiple stripes are projected simultaneously, illuminated points can belong to either plane, as shown in Figure 12(c). Figure 13 gives a more concrete demonstration of this problem. In this figure, the projector P projects two planes, which intersect the scene at two different places, a floating object and the back wall. However, because of the way the objects in this scene are organized, the top plane projected by P is seen on the bottom by C. Similarly, the bottom plane

projected by P is seen on the top by C. If the reconstruction algorithm expects the top plane in P to also be the top plane in C, an incorrect solution is sure to result. This example demonstrates that projecting patterns more complex than a single plane of light onto the scene introduces uncertainties into the reconstruction of the scene.

One possible method to combat uncertainty in projected patterns is to project light planes or other patterns of differing colors. Then, the color of lighted points in the camera can be examined to determine which plane from the projector illuminated them. There could be uncertainties here as well because the color recorded by the camera is the color of the projected light plane only if all objects in the scene are white. If the objects in the scene can be any color, ambiguities can be present in determining what color the light striking the object is.

5.3 Current Holes in Structured Light Understanding

There are certainly plenty of areas worthy of examination in the area of structured light. One such area is how projector and camera placement affects 3D scene reconstruction using such systems. As outlined in Section 2.5, there has been work in the area of camera placement in multi-camera systems for various tasks. But, to our understanding, placement of cameras and projectors has not been examined for use in structured light systems. This thesis presents work in this area, shown in Chapter 6.

Another hole in the understanding of structured light systems is proper calibration of such systems. Section 2.6 outlines existing calibration methods in the literature. The one thread that unites them all is that they all *must* have a priori knowledge of the physical scene structure for calibration. This is not necessary, as work presented in Chapter 7 shows.

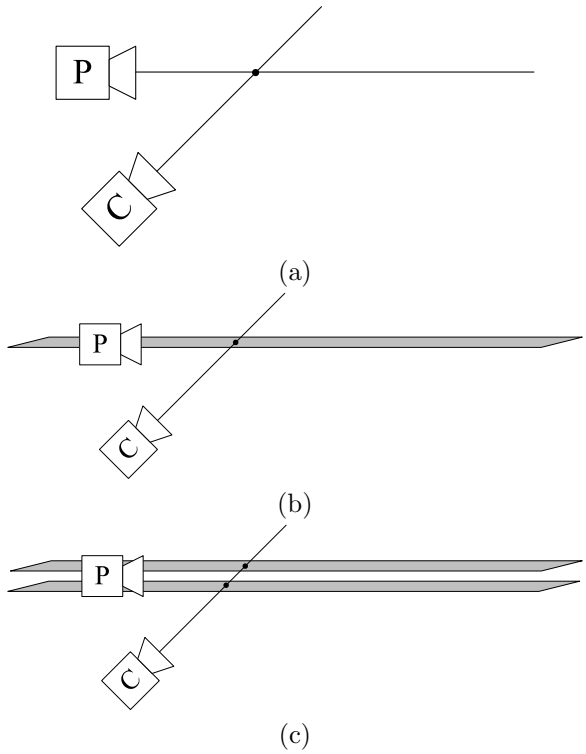


Figure 12: Different types of geometric intersections are possible using a structured light system. Uniquely solvable cases are shown in (a) and (b), while (c) is not uniquely solvable. In (a), a single point in the projector and a single point in the camera each project to a ray in the world, which intersect at most one point. In (b), a single line in the projector projects to a plane in the world and a single point in the camera projects to a ray in the world, which intersect at most one point. However, in (c), multiple lines in the projector project to multiple planes in the world and a single point in the camera projects to a ray in the world, with multiple possible solutions.

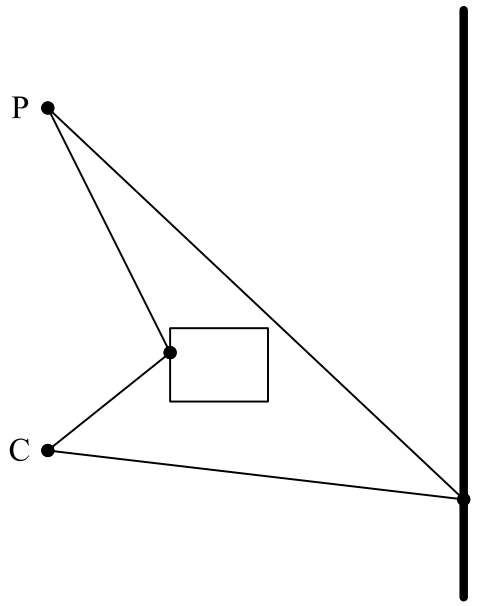


Figure 13: Demonstration of an ambiguity that can arise when multiple light planes are intersected a scene. The projector P and camera C frames are perfectly aligned, except for a translation along a single axis. In this situation, the top plane projected by the projector is the bottom plane detected by the camera, and vice-versa. This shows that without specific knowledge of the scene, there is no way of knowing which plane each point detected by C corresponds to.

6 Element Placement in Structured Light Systems

This chapter presents a mathematical basis for judging the quality of camera and projector placement in 3D for structured light systems. Here, two important quality metrics are considered: visibility, which measures how much of the target object is visible; and scale, which measures the error in detecting the visible portions. A novel method for computing each of these metrics is presented. An example is discussed which demonstrates use of these two metrics. The proposed techniques have direct applicability to the task of monitoring patient safety for radiation therapy applications.

6.1 Placement Problem Description

In this chapter, the problem of measuring the quality of camera and projector placement for structured light systems is examined. Placement of sensors is important for all detection tasks as clever algorithms can be defeated by poor sensor placement, but good placement can lead to acceptable results even from subpar algorithms. Camera placement for observation tasks in multiple camera systems is an area that has been explored. However, the interaction between projectors and cameras in structured light systems is significantly different than the interactions in multiple camera systems, leading to different placement criteria.

The quality of differing camera and projector placements is determined here by examining the physics of the problem. The spread of light from the projector into the scene and the uncertainty in the camera's detection of this light is modeled and accounted for. Taking this into consideration, judging the placement of cameras and projectors in such systems can be performed mathematically. One area where improved placement is applicable is tracking a patient's body position during radiation therapy, in which small body movements direct radiation where it is not intended. To detect such movements, the cameras and projectors in a structured light system

must be placed such that the body can be reconstructed with the desired precision.

This chapter is organized as follows. Section 6.2 then gives an intuitive description of the quality metrics used. The mathematics of these metrics are presented in Section 6.3. An example problem is analyzed in Section 6.4. Concluding comments are presented in Section 6.5. Looking back, related literature is surveyed in Section 2.5.

6.2 Placement Problem Formulation

The problem addressed in this chapter is how to classify different camera and projector placements around an example target object by how good they are for structured light-based reconstruction. There are two competing definitions for what constitutes good detection in a structured light system. The first metric, visibility, is concerned with how much of the target object is visible at any given point in time. The second metric, scale, is concerned with how precisely points on the target object can be detected. To examine the tradeoff between these metrics, consider the following limit cases. When the camera and projector are very far away from the object, the entire object is contained within the field-of-view, but points on the surface are not distinguishable from each other. Alternatively, when the camera and projector are as close as possible to the surface of the object, points on the surface are distinguishable, but not many of them are within the images.

The *visibility metric* quantifies how much of the target object is visible. When the target object is represented as a set of surface points, the visibility metric for a given camera and projector setup is then the percentage of these points visible to at least one camera and at least one projector. A point can be considered visible when its projection is within the width and height of the camera or projector's image plane.

The *scale metric* quantifies the accuracy of detection for each of the visible points. The intuition behind this metric is shown in Figure 14. Consider a single pixel being illuminated on a projector's image plane. This point illuminates a cone which spreads out into the world from the projector. If this cone intersects an object that is locally planar about the area of intersection, an ellipse is illuminated on the surface of the object. This process is shown in green in Figure 14. To perform reconstruction, a camera must then observe this illuminated ellipse. However, the camera cannot detect

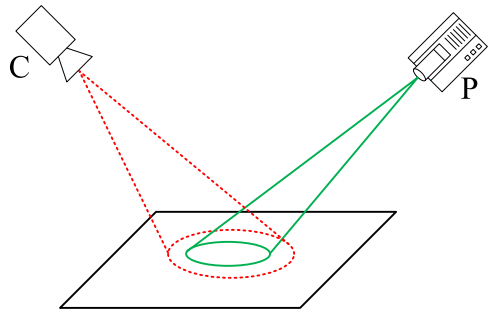


Figure 14: The intuition behind the scale quality metric presented in this chapter. Illuminating one point on the projector's image plane projects a cone of light into the scene. This cone of light intersects the scene in an ellipse shape (shown in green). The camera detects this illuminated ellipse imperfectly, making the real-world error bound the dotted red ellipse shown.

this perfectly–there is noise associated with the camera detecting the illuminated point, essentially enlarging the area in which the ideal point could be in the world, shown in red in Figure 14. The error is additive due to there being only one sensor, a major difference between camera and projector systems and stereo camera systems.

Note that the model assumes the target object to be locally planar about the area where it intersects with the illumination cone. This is reasonable as this area is on the scale of a single pixel being illuminated.

For clarity, we will now walk through how the scale quality metric is computed for a given camera and projector position. A graphical depiction of this process is shown in Figure 15. In this case, the camera and projector are arranged relative to a single point of interest, as shown in Figure 15(a). The point of interest is illuminated by a single point on the projector image plane, shown in Figure 15(b). This point of interest is also visible as a single point on the camera image plane, shown in Figure 15(c).

The points in each of the projector and camera image planes have error circles around them with the diameter of a single pixel in their respective planes. These are shown in Figure 15(d) and Figure 15(e).

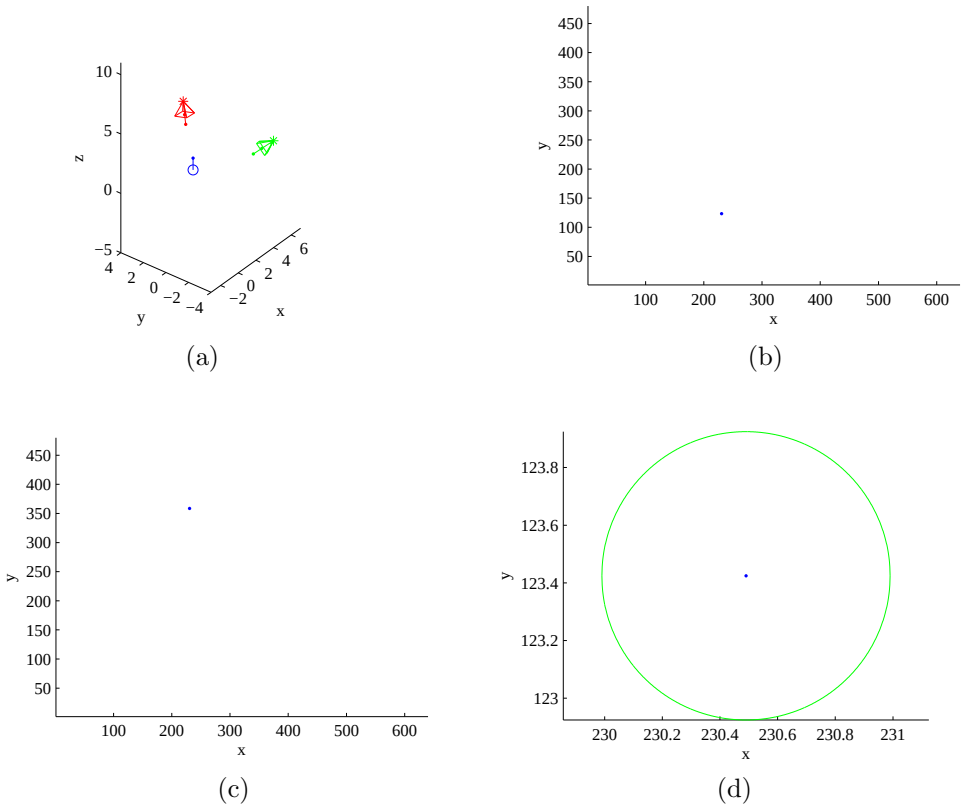


Figure 15: (a) Setup of the camera and projector system. The projector is shown in green, the camera is shown in red, and the only target point is shown in blue with an arrow representing its normal vector. (b) Target point location on the projector image plane. (c) Target point location on the camera image plane. (d) Half-pixel width error ellipse around the target point location on the projector image plane.

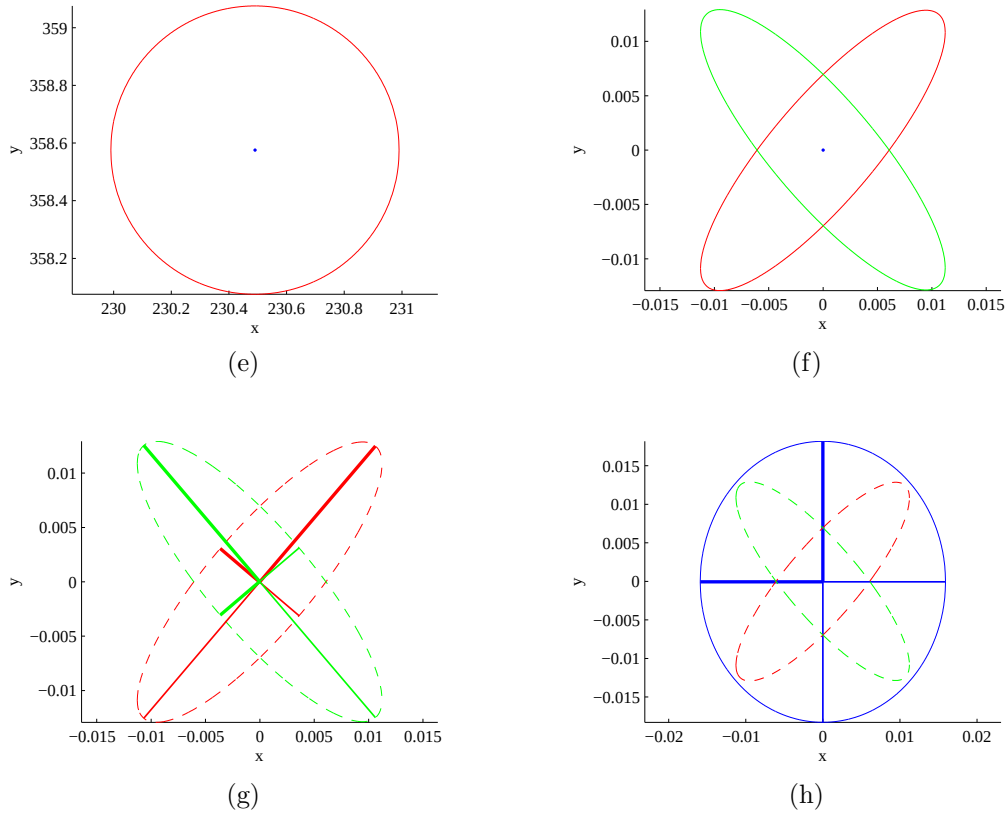


Figure 15: (e) Half-pixel width error ellipse around the target point location on the camera image plane. (f) Error ellipses from the projector (green) and camera (red) projected onto the target point tangent plane. (g) Gaussian distributions associated with the error ellipses on the target point tangent plane. (h) Final convolved Gaussian distribution on the target point tangent plane.

The target object is assumed to be planar in the area immediately surrounding the target point. Thus, the error circles on the camera and project image planes can then both be projected to ellipses on the target point tangent plane as shown in Figure 15(f).

To find how well the camera can detect the illuminated ellipse from the projector, the underlying Gaussian distribution for each of the projected ellipses must be found, as shown in Figure 15(g). These distributions are then convolved, giving the overall distribution shown in Figure 15(h). The semimajor and semiminor axis lengths are then extracted from this combined distribution, providing real-world bounds on the scale of the error for detecting the target point from the camera and projector setup given.

Note that the propagation of the camera and projector error ellipses is fundamentally different than how such error propagates in multiple camera systems. In multiple camera systems, an intersection operation would be used on the error ellipses on the target point tangent plane (Figure 15(f)). This is because multiple cameras are multiple sensors, and uncertainty decreases with multiple measurements. However, as only the camera in a camera and projector system is a sensor, only one actual measurement is taking place. Thus, the error from the projector adds to the error in sensing, so the total error increases instead of decreasing. Thus, the error ellipses must be convolved.

6.3 Placement Problem Mechanics

6.3.1 Camera Parameters

A camera can be defined by its intrinsic and extrinsic parameters. A camera's intrinsic parameters are described by the matrix $\mathbf{K}$, which represents the transformation from the camera's 3D frame to the image plane of the camera. The form of this matrix is shown in Equation (19). The camera's 3D frame has the z-axis perpendicular to the image plane, pointing out from the camera. The parameters of $\mathbf{K}$ are as follows. The scale factors along the x- and y-axis are α_x and α_y. The camera skew is s, which should be zero for most cameras. If s is not zero, it means that the x-axis and y-axis of the elements on the camera sensor are not perpendicular. Finally, the principal

point of the image is $(x_0, y_0)^\top$, which is essentially the offset from $(0, 0)^\top$ to the center of the image.

$$\mathbf{K} = \begin{bmatrix} \alpha_x & s & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{19}$$

The camera's extrinsic parameters describe the camera's location within the greater world. This is represented as a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$. Together, $\mathbf{R}$ and $\mathbf{t}$ transform from the frame of reference for the entire world to the frame of reference for the camera.

A projection matrix $\mathbf{P}$ can be defined which will map homogenous world points to homogenous points in the image, as shown in Equation (20), where the vector $\mathbf{0} = (0, 0, 0)^\top$.

$$\mathbf{P} = \begin{bmatrix} \mathbf{K} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{20}$$

6.3.2 Projector Parameters

A projector can be described by the same parameters used to describe a camera. A projector has an intrinsic parameter matrix $\mathbf{K}$, as well as extrinsic parameters described by a rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$. Keep in mind, however, that while the equations are the same, the physical process works in reverse for the projector. Instead of the scene being projected upon the image plane, as is the case with a camera, the image plane is being projected onto the scene. This distinction is important for determining the measurement error for a given structured light system.

6.3.3 Target Point Parameters

It is assumed that a representative target is given, which can be used to determine the quality of the placement of the camera(s) and projector(s). In this case, the target is recorded as a set of 3D points in the world coordinate system, $\{p_i\}$, and a set of surface normal vectors associated with those points, $\{n_i\}$. All points must lie on a convex surface.

6.3.4 Determining Visibility of Target Points

For any of the points p_i, there are two steps to finding out if it is visible to a single camera or projector (the process is identical). The first is to ensure that the point in the world projects to a point within the bounds of the camera/projector image plane. The projection of p_i onto the image plane of the camera/projector with projection matrix $\mathbf{P}$ is the point p_{Ii}. This point can be found as shown in Equation (21). It is within the bounds of the image plane when $1 \leq \frac{p_{Ii}(1)}{p_{Ii}(3)} \leq \alpha_x$ and $1 \leq \frac{p_{Ii}(2)}{p_{Ii}(3)} \leq \alpha_y$.

$$p_{Ii} = \mathbf{P} \begin{pmatrix} p_i \\ 1 \end{pmatrix}. \tag{21}$$

The angle of the normal vector of the target point and the camera must be checked to ensure visibility. If l_i is the location of the camera/projector in the world, we find $v_i = l_i - p_i$. Then, the angle β can be found as shown in Equation (22). The point is visible if $\beta < \frac{\pi}{2}$. This ensures that the normal of the target point is not directed away from the camera, which would mean the surface at that point is not visible.

$$\beta = \cos^{-1}\left(\frac{n_i \cdot v_i}{\|n_i\|_2 \|v_i\|_2}\right). \tag{22}$$

Note that for a point p_i to be considered visible by a coded structured light system, it must be visible by at least one camera *and* at least one projector. Also note that self-occlusions are not considered by this visibility model. If the target points lie on a non-convex surface, a more robust method must be used to determine point visibility.

6.3.5 Visibility Quality Metric

The visibility quality metric is the ratio of visible points ($n_{visible}$) on the target object to the total number of points on the object (n_{total}). This ratio is shown in Equation (23). In this case, a higher ratio is better, but it will always be less than one.

$$q_{visible} = \frac{n_{visible}}{n_{total}}. \tag{23}$$

The quality measure $q_{visible}$ records the percentage of the points on the object that are visible. However, it does not encode how finely those points are detected.

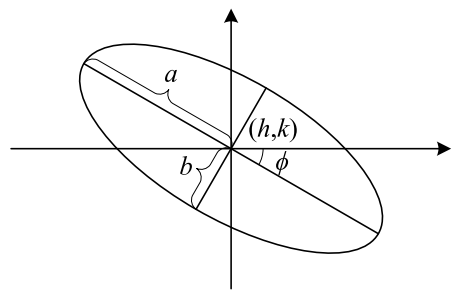


Figure 16: Depiction of an ellipse with its center (h, k), semi-major axis length a, semi-minor axis length b, and angle ϕ labeled.

6.3.6 Homography Matrix

A homography matrix between the plane around a target point and a camera can be found as shown in Equation (24). The target point is p_i and its surface normal is n_i. Also, $\mathbf{R_T}$ is the rotation matrix between the target plane's frame of reference and that of the world frame. The matrices $\mathbf{K}$, $\mathbf{R}$, and the vector $\mathbf{t}$ are the camera's parameters.

$$\mathbf{H} = \begin{bmatrix} \mathbf{K} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{R_T} & p_i \\ \mathbf{0}^\top & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{24}$$

The homography matrix $\mathbf{H}$ maps from the plane around p_i to the image plane of the camera, as shown in Equation (25). Here, p_T is a point in the target plane and p_I is the corresponding point in the camera's image plane.

$$p_I = \mathbf{H} p_T. \tag{25}$$

6.3.7 Ellipses

Ellipses can be represented two ways. The first is the parametric form, in which five parameters are used to describe the ellipse: the ellipse center (h, k), the semi-

major axis length a, the semi-minor axis length b, and the angle with the x-axis ϕ (Figure 16).

The other representation of an ellipse is the general form, as shown in Equation (26). Note this is the general form for a conic section.

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0. \tag{26}$$

Parametric Form to General Form One possible set of equations for the transformation from the set of ellipse parameters to the general form is shown below. Space is limited for these equations, so $c\phi = \cos\phi$, and $s\phi = \sin\phi$.

$$A = \frac{c^2\phi}{a^2} + \frac{s^2\phi}{b^2} \tag{27}$$

$$B = 2\left(\frac{1}{b^2} - \frac{1}{a^2}\right) c\phi s\phi \tag{28}$$

$$C = \frac{s^2\phi}{a^2} + \frac{c^2\phi}{b^2} \tag{29}$$

$$D = \left(\frac{2k}{a^2} - \frac{2k}{b^2}\right) c\phi s\phi - 2h\left(\frac{c^2\phi}{a^2} + \frac{s^2\phi}{b^2}\right) \tag{30}$$

$$E = 2h\left(\frac{1}{a^2} - \frac{1}{b^2}\right) - 2k\left(\frac{s^2\phi}{a^2} + \frac{c^2\phi}{b^2}\right) \tag{31}$$

$$F = \left(\frac{h^2}{a^2} + \frac{k^2}{b^2}\right) c^2\phi + 2hk\left(\frac{1}{b^2} - \frac{1}{a^2}\right) c\phi s\phi + \left(\frac{k^2}{a^2} + \frac{h^2}{b^2}\right) s^2\phi - 1. \tag{32}$$

General Form to Parametric Form The set of equations for the transformation from the general form of an ellipse to the parametric form is shown below. This is for the case in which $B \neq 0$.

$$\phi = \frac{1}{2}\tan^{-1}\left(\frac{B}{C-A}\right) \tag{33}$$

$$a = \sqrt{\frac{2|F|\sin(2\phi)}{(A+C)\sin(2\phi) - B}} \tag{34}$$

$$b = \sqrt{\frac{2|F|\sin(2\phi)}{(A+C)\sin(2\phi) + B}} \tag{35}$$

$$h = \frac{BE - 2CD}{4AC - B^2} \tag{36}$$

$$k = \frac{2AE - DB}{B^2 - 4AC}. \tag{37}$$

If $B = 0$, then the following equations for a and b must be used instead.

$$a = \frac{1}{\sqrt{A}} \tag{38}$$

$$b = \frac{1}{\sqrt{C}}. \tag{39}$$

6.3.8 Discussion of Gaussian Distributions

An ellipse can be considered a level set of a Gaussian distribution. Thus, we can find the Gaussian distribution associated with a given ellipse, and vice-versa.

Ellipse Parameters to Gaussian Distribution A fixed Gaussian distribution is represented by a vector mean μ and a covariance matrix $\mathbf{\Sigma}$. These can be found from the parameters as follows. Note here that $c\phi = \cos\phi$, and $s\phi = \sin\phi$.

$$\mu = \begin{pmatrix} h \\ k \end{pmatrix} \tag{40}$$

$$\mathbf{\Sigma} = \begin{bmatrix} c\phi & s\phi \\ -s\phi & s\phi \end{bmatrix} \begin{bmatrix} a^2 & 0 \\ 0 & b^2 \end{bmatrix} \begin{bmatrix} c\phi & -s\phi \\ s\phi & c\phi \end{bmatrix}. \tag{41}$$

Gaussian Distribution to Ellipse Parameters The parameters ϕ, a, and b require us to find the eigenvalues and eigenvectors of $\mathbf{\Sigma}$. The eigenvalues are λ_1 and λ_2, where $\lambda_1 \geq \lambda_2$. The associated eigenvectors are $\mathbf{e}_1$ and $\mathbf{e}_2$.

$$\phi = -\tan^{-1}\left(\frac{\mathbf{e}_1(1)}{\mathbf{e}_1(2)}\right) \tag{42}$$

$$a = \sqrt{\lambda_1} \tag{43}$$

$$b = \sqrt{\lambda_2} \tag{44}$$

$$h = \mu(1) \tag{45}$$

$$k = \mu(2). \tag{46}$$

Convolution of Gaussian Distributions Convolving Gaussian distributions is very useful, and can be accomplished with the following equations. The convolved Gaussian distribution is represented by mean μ_c and covariance $\mathbf{\Sigma}_c$.

$$\mu_c = \mu_1 - \mu_2 \tag{47}$$

$$\mathbf{\Sigma}_c = \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2. \tag{48}$$

6.3.9 Projection of Ellipses

Taking an ellipse in the form of Equation (26), a matrix representation of the ellipse, $\mathbf{C}$, can be created as shown in Equation (49).

$$\mathbf{C} = \begin{bmatrix} A & B/2 & D/2 \\ B/2 & C & E/2 \\ D/2 & E/2 & F \end{bmatrix}. \tag{49}$$

Once the ellipse is in this matrix form, the projection of the ellipse onto another plane can be found using the homography matrix calculated in Equation (24). For instance, Equation (50) shows the projection of an ellipse in the camera image plane ($\mathbf{C}_I$) onto the target point tangent plane ($\mathbf{C}_T$).

$$\mathbf{C}_T = \mathbf{H}^\top \mathbf{C}_I \mathbf{H}. \tag{50}$$

6.3.10 Scale Quality Metric

The scale quality metric encodes the scale at which the points that are visible is detected. Here, i iterates over all visible points. The area of an ellipse is $ab\pi$, where a and b are the semi-major and semi-minor axis lengths. Thus, if a_c and b_c are the semi-major and semi-minor axis of the convolved Gaussian distribution, we get the equation for q_{scale} shown in Equation (51).

$$q_{scale} = \frac{\pi}{n_{visible}} \sum_{i=1}^{n_{visible}} a_{ci} b_{ci}. \tag{51}$$

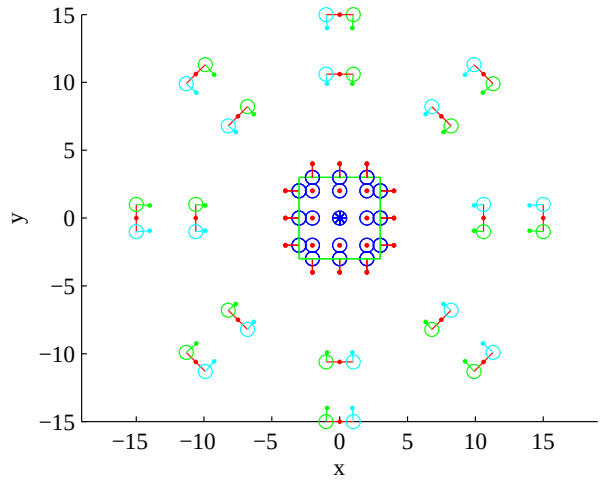
The quality metric q_{scale} is in world units. Thus, it can be used if resolution at a desired scale can be observed and tracked by the system. The best camera and projector placement is the one in which q_{scale} is minimized.

6.3.11 Multiple Cameras and/or Projectors

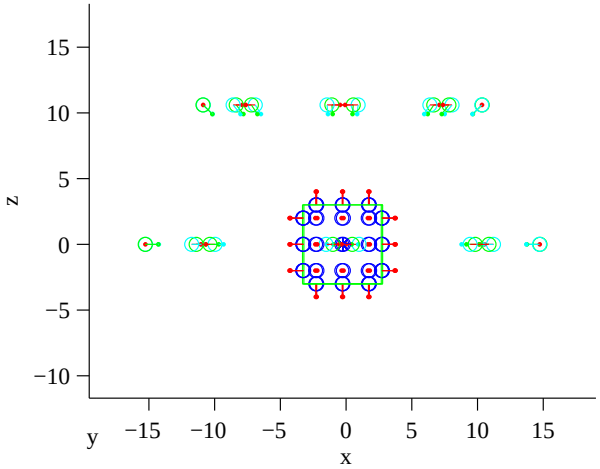
For situations where multiple cameras and/or projectors are utilized, it should be generally assumed that the worst-case scale and visibility metrics calculated be used for each point on the target object. This is because it is not possible to tell just from the projected patterns and corresponding camera images in the real system which image or pattern is better than another, especially if all projectors and cameras operate simultaneously. This being the case, the error has to be assumed to be the worst possible when scoring a particular placement setup.

6.4 Placement Example

The quality of a set of camera and projector pair placements around a cubic target object are examined in this example. The target points chosen are regularly distributed across the faces of the cube. For all camera and projector placements, the center point between the camera and projector is a constant distance from the center of the cube and the camera and projector are a constant distance from each other.



(a)



(b)

Figure 17: Two views of the set of camera and projector placements tested, seen around the cubic target object.

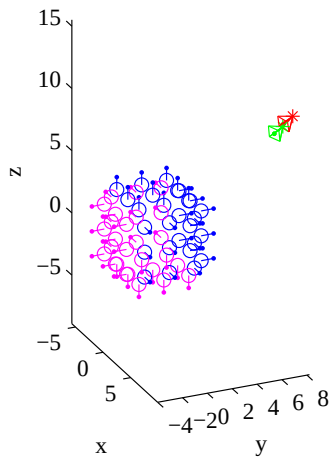


Figure 18: The camera/ projector placement tested with the best $q_{visible}$ score. Visible target points are shown in blue, nonvisible points in magenta. Note that three faces (50%) of the cube are visible.

The camera and projector point towards the center of the cube. The purpose of this example is to show which of the camera and projector pair positions has the best $q_{visible}$ score and which has the best q_{scale} score.

Two views of the set of examined camera/projector pairs around the target object are shown in Figure 17. Each of these views was examined to determine which view was best in terms of each of the quality metrics.

The camera/projector pair location with the best $q_{visible}$ score is shown in Figure 18. Note that three faces of the cube are visible to both the camera and projector in this setup. This placement makes 50% of the cube visible, the maximum amount possible. Thus, this camera and projector placement maximizes the visibility metric.

Finally, the camera/projector pair location with the best q_{scale} score is shown in Figure 19. Here, only one face is visible to the camera and projector. This is reasonable because if multiple faces are visible, the error ellipses on the faces are longer due to the grazing angles between the cube faces and the camera or projector, leading to larger error bounds. Thus, a view of a single face reduces the average error per point.

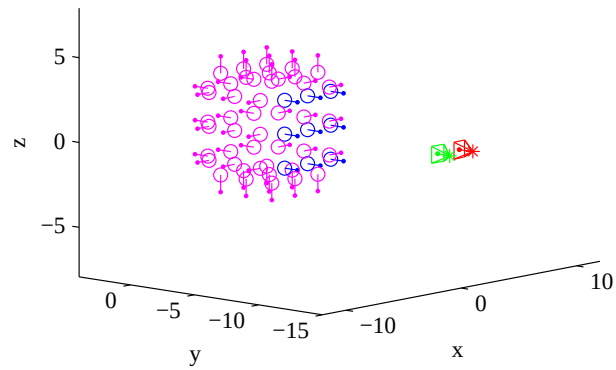


Figure 19: The camera/ projector placement tested with the best q_{scale} score. Visible target points are shown in blue, nonvisible points in magenta. Here, only one face of the cube is visible, as this minimizes the size of the error projections.

6.5 Placement Final Thoughts

In this chapter, a mathematical basis for judging the quality of camera and projector placement for structured light systems was presented. Two metrics of quality must be considered: visibility, which measures how much of the target object is visible, and scale, which measures how well the parts that are visible can be seen. Methods for computing each of these metrics were presented. An example was shown which demonstrates these quality metrics.

The method presented here is useable in domains containing multiple cameras and multiple projectors. In addition, differing resolutions between the cameras and projectors are taken into account. An application where placement quality is important is patient body tracking, in which cameras and projectors must be placed such that a patient's body can be reconstructed to a desired precision.

Future directions include addressing self-occlusion in the target object, which requires a more robust visibility model. This may be possible using something similar to the visibility regions discussed by Tarabanis, *et al.* [71]. In addition, while it is expected that the relative importance of the two quality metrics is dependent on

application, the optimal trade-off between them must still be examined in depth.

Using the method presented here, the quality of competing camera and project placements in structured light systems can now be mathematically compared. No longer must guesswork be used in placing cameras and projectors for structured light systems.

7 Optimal Calibration of Camera and Projector Systems

This chapter presents a method to perform Euclidean calibration on camera and projector-based structured light systems, without assuming specific scene structure. The vast majority of the methods in the pertinent literature rely on some a priori knowledge of the 3D scene geometry. Examples of a priori information used include manual manipulation of occlusions so the pattern shines on them in a predetermined manner, manually measuring the world location of projected points or lines, scanning specific calibration objects with the light pattern, or ensuring the entire scene obeys a specific setup. The main contribution of this work is to eliminate the requirement for a priori knowledge of the scene. By using multiple cameras, the method presented here is able to calibrate camera and projector systems without requiring specific geometric constraints on the scene, objects in it, tedious manual manipulation of occlusions, or manual world point measurements. The method presented here is optimal in terms of image-based reprojection error. Simulations are shown which characterize the effect noise has on the system, and experimental verification is performed on complex and cluttered scenes.

7.1 Calibration Problem Description

Structured light systems operate on similar principles to stereoscopic camera systems, except that at least one camera is replaced by a projector. The 3D scene reconstructions generated can be considered as good as that constructed by a stereoscopic camera system. One central difficulty in using a stereoscopic camera set to perform scene reconstruction is that features in one camera's image have to be precisely matched to features in the other cameras' images. This can be difficult, and it is known as the correspondence problem. Structured light systems are often used

instead of stereoscopic camera systems because by illuminating known patterns with a projector, the correspondence problem is greatly simplified. Batlee, *et al.* [5] and Salvi, *et al.* [64] present a good pair of survey papers on the topic.

Structured light system work best in indoor environments where the light can be strictly controlled. This makes them ideal for medical applications such as conformal radiation treatments where the surface of a patient's body must be precisely tracked to ensure that high radiation dosages are targeted at the correct locations. Other medical uses for visual body-surface scans are described by Treleaven and Wells [75] as well. Unfortunately, to reconstruct the scene as accurately as possible, the most accurate calibration must first be found. This chapter addresses the problem of finding the optimal Euclidean calibration of structured light systems in general scenes.

The calibration methods that these authors have identified in the literature all require a priori knowledge of the environment in which the structured light system to be calibrated exists. Often, these constraints can be very inconvenient to fulfill. This knowledge is necessary because a projector does not act as a sensor, so it is impossible to tell which point in the projector image corresponds to a specific world point by examining only the projector image. This is different from a camera where this is possible, a property which makes calibration easier on stereoscopic camera systems. The method presented here does away with this requirement by using multiple cameras for the calibration.

7.2 Calibration Approach Outline

Related work in the area does not find a truly optimal solution due to the way in which projectors function (differently than cameras). The common thread is that they all develop methods in an attempt to make up for not being able to find the world coordinates of scene points illuminated by the projector using a single camera.

The structured light calibration method presented here differs from existing methods by functioning on general scenes. That is, this method does not depend on the projected pattern in the scene meeting specific geometric constraints in the real world, merely that points are visible to the cameras involved. Also, this method does *not* require a human operator to move objects around the scene so projector patterns

shine on them at the exact right angle or otherwise painstakingly measure the exact world location of projected points.

In existing methods, the reason that the world coordinates of projected points are not readily available is that the projector is not a sensor. It is an illumination device that can illuminate points in the scene that can be detected by a camera. This is very useful once the system is calibrated because it greatly simplifies 3D reconstruction of the scene by greatly simplifying the correspondence problem. However, it is a problem when calibrating because the projector is not able to detect specific points on a physical calibration pattern in the scene. The projector merely illuminates indiscriminate points in the scene. That is, we cannot know ahead of time what the world coordinates of projector points will be illuminated by any given projection pattern. Such information is required when calibrating the system based upon their images.

The existing literature is full of tricks which attempt to get around this limitation, usually by involving a human in the loop either moving a physical object around the scene, trying to get the projected pattern to shine on it at the exact right angle, or by painstakingly measuring the world coordinates of points illuminated by the projector after the projector projects them. Or some combination of these.

To find the optimal Euclidean calibration of a camera and projector system with minimal human interaction, the strengths of the stereoscopic camera approach need to be integrated with the strengths of the structured light approach. To do this at least two cameras are required to calibrate one projector, and all camera and projector views must overlap with each other.

By using two cameras instead of one, the problem becomes theoretically sound. With a calibrated camera pair, it does not matter if a projected pattern conforms to expected geometric constraints in the real world or not. The world coordinates of any arbitrary points illuminated by the projector can be estimated using their measured location in the camera images. These estimated world coordinates essentially make an ad hoc calibration pattern which can be used to calibrate the projector.

7.2.1 Two-Camera Requirement for Calibration

Optimal Euclidean calibration of a structured light system requires at least two cameras whose fields-of-view overlap the field-of-projection of each projector. What follows is a discussion of why a single camera is not sufficient.

Consider a single camera, denoted c, whose field-of-view overlaps with a single projector's, denoted p, field-of-projection. The camera's projection matrix $\mathbf{P}_c$ can be found by calibrating the camera against a fixed calibration target. Meanwhile, point correspondences between the projector and camera images can be found by having the projector project single points onto the scene which the camera then detects. Using these point correspondences, the fundamental matrix $\mathbf{F}$ can be found which links the camera and projector projectively such that $\mathbf{x}_p^\top \mathbf{F} \mathbf{x}_c = 0$, where $\mathbf{x}_p$ is a homogenous-coordinate representation of a point in the projector image plane and $\mathbf{x}_c$ is a homogenous-coordinate representation of the corresponding point in the camera image plane.

For general projection matrices $\mathbf{P}_c$ and $\mathbf{P}_p$, the fundamental matrix can computed as shown in Equation (52). Here, $\mathbf{e}_p$ is the epipole of the camera in the projector's image plane.

$$\mathbf{F} = [\mathbf{e}_p]_\times \mathbf{P}_p \mathbf{P}_c^+. \tag{52}$$

Note that $\mathbf{e}_p$ can be computed directly from the Singular Value Decomposition of $\mathbf{F}$, and can thus be considered known.

Since $\mathbf{F}$ is a 3×3 matrix, we have nine equations for 11 unknowns if we naively try to solve for $\mathbf{P}_p$. There are 11 unknowns instead of 12 because a projection matrix is only accurate up to a scale factor. Note that $\mathbf{F}$, $\mathbf{P}_c$, and $[\mathbf{e}_p]_\times$ are known.[2]

Thus, it is not possible to find a single solution for $\mathbf{P}_p$ given the set of equations arising from a single camera and a single projector. Any number of solutions can be found for $\mathbf{P}_p$, but they would only yield projective reconstructions of the scene, not Euclidean. However, by adding a second camera, estimates of the 3D locations of points illuminated by the projector can be found, creating a transient calibration pattern against which the projector can be calibrated.

Eliminating Variables The astute reader at this point will note that since there are too many variables to solve for, it would be prudent to use the structure of these equations to our advantage to get rid of variables we can estimate from other sources. Indeed, we know that a general projection matrix $\mathbf{P}$ has the form shown in Equation (54), where $\mathbf{K}$ is the camera intrinsic parameter matrix, $\mathbf{R}$ is the rotation matrix of the camera with respect to the world, and $\mathbf{c}$ is the location of the camera in the world.

Now, $\mathbf{R}$ is a rotation matrix, which means that it can be represented using three unknowns. Also, the camera center, c, is a three vector, having three unknowns. Finally, $\mathbf{K}$ has the form shown in Equation (55), where $a_x = fm_x$ and $ay = fm_y$. Here, f is the focal length of the camera, m_x is the number of pixels along the x-axis in the image, m_y is the number of pixels along the y-axis in the image, s is the skew factor, x_0 is the image center along the x-axis, and y_0 is the image center along the y-axis. Thus, naively, there are six unknowns in $\mathbf{K}$.

Since we have the actual images that we are projecting though, m_x, m_y, x_0, y_0, and s can be considered known. This then leaves seven unknowns and nine equations from Equation (52).

At first glance, this is a good set of equations, but it unfortunately is not actionable due to the nonlinearities and numerical instabilities involved.

7.3 Algorithm

Our algorithm for solving for the overall calibration of a structured light system is described as follows.

7.3.1 Initial Camera Projection Matrix Estimation

In the first step of our method, initial estimates for the projection matrices for the two cameras (denoted $\mathbf{P}_{c_1}$ and $\mathbf{P}_{c_2}$) are found based upon a physical calibration pattern placed in the scene. This calibration pattern consists of a set of points with known world coordinates that is imaged by both cameras. The calibration method uses these world coordinates to estimate the camera parameters. The algorithm we use is the "Gold Standard" algorithm for estimating $\mathbf{P}$ using corresponding world and image

points as described by Hartley and Zisserman [31]. This is not an endorsement of point-correspondence camera calibration algorithms over algorithms that use other types of known world calibration pattern geometry, *e.g.*, methods based on parallel and orthogonal lines with known lengths. Any Euclidean calibration algorithm can be used.

Note that a physical calibration pattern is needed at some point. This cannot be avoided. Without some known geometric dimensions in the scene and their corresponding images, a Euclidean reconstruction can not be found. At best, a projective reconstruction can be found, which is insufficient for most reconstruction tasks.

7.3.2 Projector Pattern World Point Coordinate Estimation

In the second step of our method, a set of points are sequentially projected onto the scene from the projector. The subset of these points which are visible to both cameras is found, and the optimal world coordinates of the points are estimated. Optimal in this sense means finding an estimate for the world coordinates that minimizes the reprojection error, *i.e.*, find the estimated value of the world point $\hat{\mathbf{x}}_w$ which minimizes the value of err in Equation (53), where $\mathbf{x}_{c_1}$ is the detected point in the first camera image and $\mathbf{x}_{c_2}$ is the detected point in the second camera.

$$err = \|\mathbf{x}_{c_1} - \mathbf{P}_{c_1}\hat{\mathbf{x}}_w\|_2 + \|\mathbf{x}_{c_2} - \mathbf{P}_{c_2}\hat{\mathbf{x}}_w\|_2. \tag{53}$$

7.3.3 Initial Projector Projection Matrix Estimation

In the third step, we use the estimated value of the world points to estimate the projection matrix of the projector. Since we have the projector image points $\mathbf{x}_p$ corresponding to the estimated world point locations $\hat{\mathbf{x}}_w$, we can use the same calibration algorithm used in step one to generate an initial estimate of the projector's projection matrix $\mathbf{P}_p$. In this case, however, the calibration algorithm used must be a point correspondence algorithm, as other types of geometric constraints (such as lines) do not hold when projected onto a general scene, which is the situation we address.

7.3.4 Iterative Nonlinear Solution Refinement

The fourth and final step uses the Levenberg-Marquardt iterative nonlinear optimization algorithm to refine the calibration estimate. In this case, we optimize over the parameters of the three projection matrices and the world coordinate estimates for the points illuminated by the projector.

Consider a general projection matrix $\mathbf{P}$. The projection matrix is composed by its constituent components as shown in Equation (54). These components correspond to the intrinsic parameter matrix $\mathbf{K}$, the camera rotation matrix with respect to the world $\mathbf{R}$, and the camera center $\mathbf{c}$. As $\mathbf{K}$ is upper triangular, and $\mathbf{R}$ is a rotation matrix, $\mathbf{K}$ and $\mathbf{R}$ can be found by performing RQ-decomposition on the first three columns of $\mathbf{P}$. The camera center $\mathbf{c}$ can then be solved for.

$$\mathbf{P} = [\mathbf{KR} | -\mathbf{KRc}] \,. \tag{54}$$

The intrinsic parameter matrix $\mathbf{K}$ has the form as shown in Equation (55). Here, a_x is the scale factor in the x-axis, a_y is the scale factor in the y-axis, s is the skew factor, and (x_0, y_0) is the principal point of the image.

$$\mathbf{K} = \begin{bmatrix} a_x & s & x_0 \\ 0 & a_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} . \tag{55}$$

The rotation matrix $\mathbf{R}$ can be represented by three values, for instance by Rodrigues' rotation, which yields a 3-vector $\mathbf{w}$. The camera center $\mathbf{c}$ is a 3-vector, and as we have just seen, $\mathbf{K}$ contains five components. Thus, $\mathbf{P}$ can be decomposed into a total of 11 components over which we perform the nonlinear optimization–a_x, a_y, s, x_0, y_0, the three components of $\mathbf{w}$, and the three components of $\mathbf{c}$.

We further optimize over the set of 3D world coordinate estimates of the points illuminated by the projector $\{\hat{\mathbf{x}}_w\}$ as found in the second step. Thus, if n_{proj} is the number of points in the set $\{\hat{\mathbf{x}}_w\}$, then a total of $33 + 3n_{proj}$ variables are optimized over using the Levenberg-Marquardt method.

The cost function utilized by the Levenberg-Marquardt method for this optimization is the sum of the reprojection error for each point in the three images (two camera

images, and one projector image). The cost function is shown in Equation (56). Note that we index the points by the superscript i to make it clear that this cost is calculated for just one point at a time. The Jacobian of c^i_{proj} is solvable by symbolic mathematics packages.

$$c^i_{proj} = \|\mathbf{x}^i_{c_1} - \mathbf{P}_{c_1}\hat{\mathbf{x}}^i_w\|_2 + \|\mathbf{x}^i_{c_2} - \mathbf{P}_{c_2}\hat{\mathbf{x}}^i_w\|_2 + \|\mathbf{x}^i_p - \mathbf{P}_p\hat{\mathbf{x}}^i_w\|_2. \tag{56}$$

In addition to this optimization, we optimize over the reprojection error in the two camera images for the calibration pattern points. This cost function is shown in Equation (57). Here, we index the points by the superscript j again to make it clear that this cost is calculated for a single point, but also to make the distinction between the image points used. The points indexed by i are the projector illuminated points, and the points indexed by j are the calibration pattern points. The Jacobian of c^j_{calib} is also solvable by using symbolic mathematical packages.

$$c^j_{calib} = \|\mathbf{x}^j_{c_1} - \mathbf{P}_{c_1}\mathbf{x}^j_w\|_2 + \|\mathbf{x}^j_{c_2} - \mathbf{P}_{c_2}\mathbf{x}^j_w\|_2. \tag{57}$$

7.4 Simulation

One positive aspect of our calibration method is that it can be readily simulated. Unlike methods that require a human manually place occlusions precisely in the environment, the only error that we need to model is the error in measuring the image location of the points of interest, whether they are illuminated by the projector or situated on the calibration pattern. No model needs to be created for error induced by manually placing obstacles, for instance.

7.4.1 Single-Run Simulation

For one run of the simulation, two virtual cameras and a virtual projector are placed (pseudo) randomly on a sphere pointed at the sphere's center. Two sets of (pseudo) random points are then created within the confines of the sphere. The structured light calibration method described in Section 7.3 is then run utilizing both sets of points. One set of points represents the calibration pattern, and the world coordinates of the points are considered known. These points are used to perform the initial camera

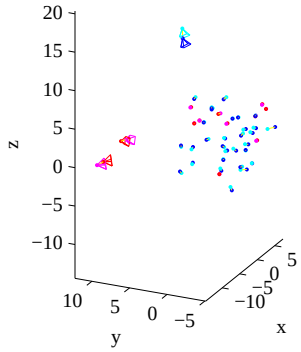


Figure 20: 3D reconstruction of a simulated scene as described in Section 7.4.1. The real cameras are shown in magenta and the estimated cameras are shown in red. The real projector is shown in cyan and the estimated projector is shown in blue. For the projected world points, the real points are shown in cyan and the reconstructed points in blue. For the calibration pattern points, the real points are shown in magenta and the reconstructed points in red. Corresponding points are connected with yellow lines.

calibration as described in Section 7.3.1 above. The other set of points represents the points illuminated by the projector, so the world coordinates of that set are considered unknown. An initial estimate of the world coordinates of these points is estimated as described in Section 7.3.2. This estimate is used to perform the initial projector calibration as described in Section 7.3.3. Finally, the estimates of all calibration matrices are iteratively refined as described in Section 7.3.4.

To make the simulation more realistic, the image-plane projections of both sets of points are calculated for both cameras, and the point locations are corrupted by Gaussian noise. The image locations of the illuminated points are calculated for the projector, but not corrupted since the projector is not a sensor in the physical system we are modeling. The output for one example calibration is shown in Figures 20, 21, 22, and 23.

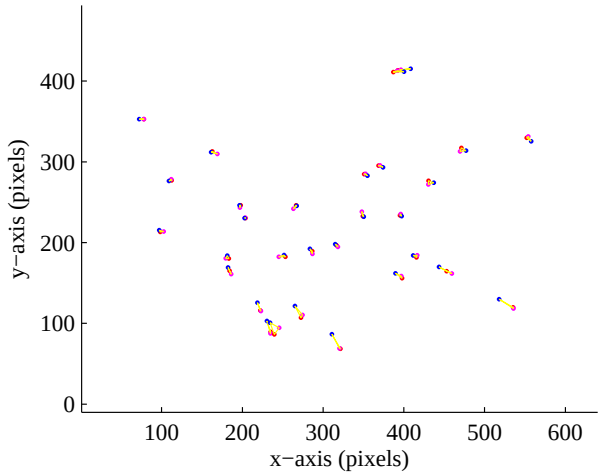


Figure 21: Image from camera one of the simulated scene depicted in Figure 20. Here, the uncorrupted point location is shown in red, the corrupted location is shown in magenta, and the projection of the real world point using the estimated projection matrix for the camera is shown in blue. Corresponding points are connected with yellow lines.

7.4.2 Multiple-Run Simulations

A set of simulations (described above in Section 7.4.1) was run in which the degree of noise corrupting the imaged points in the camera images was varied. This set of simulations was then used to characterize the effect that different levels of noise in the camera images has on the overall system performance.

The standard measure of error used in the camera calibration community to judge quality of camera calibration methods is the reprojection error. There are good reasons for this. The primary reason is that aspects of the projection matrix are hard to classify individually with traditional error metrics, *e.g.*, there is no good way to quantify how different two rotation matrices are. There is also no good method to determine how much weight should be given to the intrinsic versus extrinsic parameters of the camera.

A set of simulations was run in which the standard deviation of the Gaussian noise applied to the image points ranged from one to five pixels. Five hundred simulations

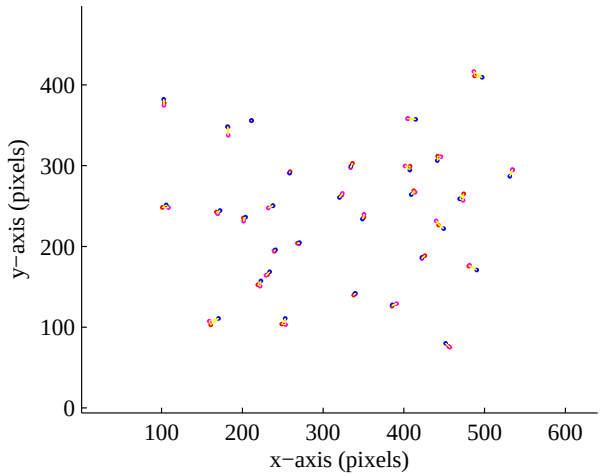


Figure 22: Image from camera two of the simulated scene depicted in Figure 20. Here, the uncorrupted point location is shown in red, the corrupted location is shown in magenta, and the projection of the real world point using the estimated projection matrix for the camera is shown in blue. Corresponding points are connected with yellow lines.

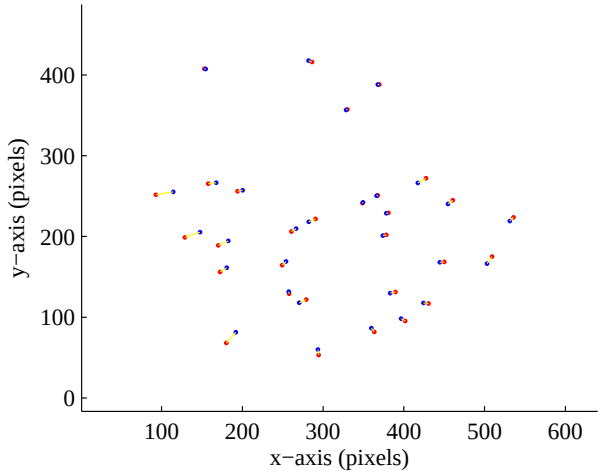


Figure 23: Image from the projector of the simulated scene depicted in Figure 20. Here, the uncorrupted point location is shown in red and the projection of the real world point using the estimated projection matrix for the camera is shown in blue. Corresponding points are connected with yellow lines.

were run for each level of Gaussian noise, for a total of 2500 simulations. The average error recorded is shown in Figure 24 in solid lines, with the one-standard deviation mark denoted in dotted lines. The blue solid line shows the mean of all $\frac{1}{2}c^j_{calib}$ for each image corruption level, and the blue dotted line shows the one standard deviation mark. The green solid line shows the mean of all $\frac{1}{3}c^i_{proj}$ for each image corruption level, and the green dotted line shows the one standard deviation mark. Note that c^i_{proj} is decimated by 3 for this comparison because it is an error measure across three images (both cameras and the projector), while c^j_{calib} is decimated by 2 because it is an error measure across two images (only the two cameras). Thus, the relative scale can be readily compared.

Examining Figure 24, it can be seen that the per-point-per-image reprojection error for the projection-pattern-based calibration ($\frac{1}{3}c^i_{proj}$) is always higher than that of the per-point-per-image reprojection error for the calibration-pattern-based calibration ($\frac{1}{2}c^j_{calib}$). This follows because the result of the calibration-pattern-based calibration of just the two cameras is utilized as an initialization point for the optimization of the overall calibration of the entire system across both cameras and the projector.

However, it can also be seen that the variance in $\frac{1}{3}c^i_{proj}$ increases while the variance in $\frac{1}{2}c^j_{calib}$ remains fairly steady. This is due mostly to the fact that our simulation selects random camera and projector placements. When the cameras are poorly placed with respect to each other, *e.g.*, either very close to one another or facing each other across the generated point clouds, both scene reconstruction and calibration become much more sensitive to noise. Increasing image point corruption expands the set of overly sensitive relative placement positions, causing the increases in the variance of $\frac{1}{3}c^i_{proj}$. The simulation yielding the worst $\frac{1}{3}c^i_{proj}$ value is shown in Figure 25, which was incidently corrupted with the maximum noise level. Notice how the cameras are positioned directly across the point clouds from each other, yielding large errors in the projection point positions. While this shows that high image noise does produce worse results, it is hopeful in that it also shows that even in the face of high image noise in the cameras, error can be minimized by choosing good camera and projector placement.

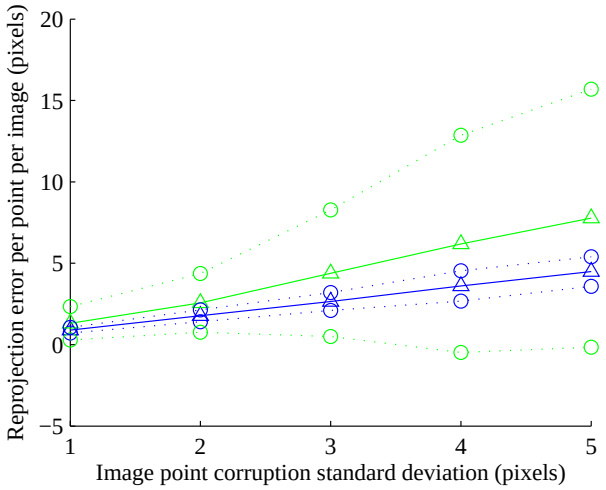


Figure 24: Results of a set of simulations showing the effect on the reprojection errors induced by different levels of corruption in the camera images. Shown is the mean of all $\frac{1}{2}c^j_{calib}$ for each image corruption level (blue solid line), and its one standard deviation mark (blue dotted line). Also shown is the mean of all $\frac{1}{3}c^i_{proj}$ for each image corruption level (green solid line), and its one standard deviation mark (green dotted line). See the text for further explanation.

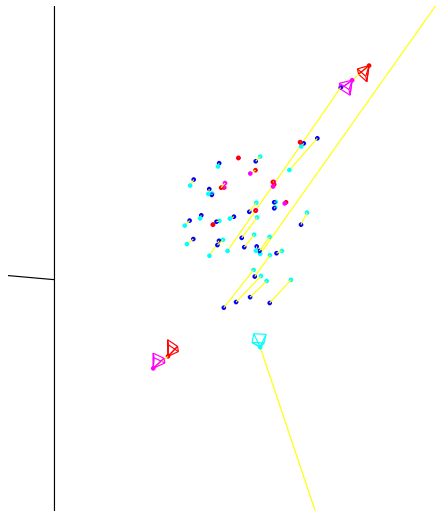


Figure 25: 3D reconstruction of the simulated scene with the worst $\frac{1}{3}c^i_{proj}$ value as described in Section 7.4.2. The real cameras are shown in magenta and the estimated cameras are shown in red. The real projector is shown in cyan and the estimated projector is out of range. For the projected world points, the real points are shown in cyan and the reconstructed points in blue. For the calibration pattern points, the real points are shown in magenta and the reconstructed points in red. Corresponding points are connected with yellow lines. Note how the cameras are pointed almost directly at each other, yielding high errors in the calibration pattern point reconstructions.

7.5 Real-World Verification

A set of real-world tests was performed with cameras positioned around cluttered scenes, demonstrating one of the main contributions of this method: a structured light calibration method that does not assume specific scene structure. The following subsections show the results of two of these tests.

7.5.1 Real-World Test 1

The overall system setup for this test is shown in Figure 26(a). The calibration pattern, with measured and reprojected points is shown in Figure 27(a) for camera 1 and Figure 27(b) for camera 2. The projected pattern, with detected and reprojected points is shown in Figure 27(c) for camera 1, Figure 27(d) for camera 2, and Figure 27(e) for the projector. Finally, the overall 3D scene reconstruction for the cameras, projectors, calibration pattern points, and projection pattern points is shown in Figure 26(b).

7.5.2 Real-World Test 2

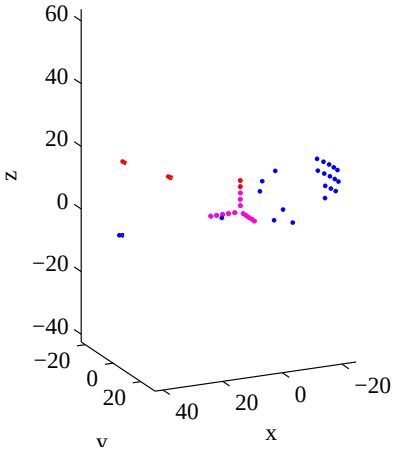
The overall system setup for this test is shown in Figure 28(a). The calibration pattern, with measured and reprojected points is shown in Figure 29(a) for camera 1 and Figure 29(b) for camera 2. The projected pattern, with detected and reprojected points is shown in Figure 29(c) for camera 1, Figure 29(d) for camera 2, and Figure 29(e) for the projector. Finally, the overall 3D scene reconstruction for the cameras, projectors, calibration pattern points, and projection pattern points is shown in Figure 28(b).

7.5.3 Real-World Test 3

This test is very similar to Real-World Test 1 in Section 7.5.1. A different calibration pattern was used in this one, and the objects are arranged differently. This shows that Test 1 was not just a one-off.

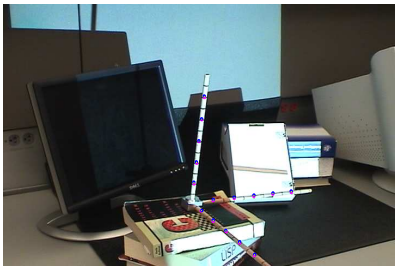
The overall system setup for this test is shown in Figure 30(a). The calibration pattern, with measured and reprojected points is shown in Figure 31(a) for cam-

(a)

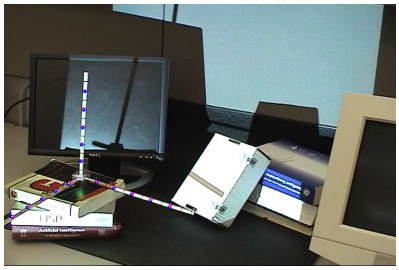


(b)

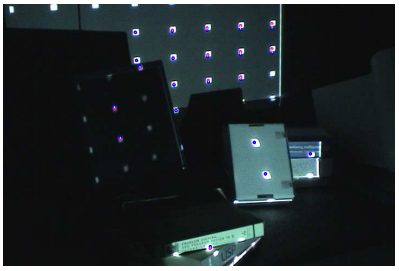
Figure 26: (a) Image of the overall system setup for the real-world verification described in Section 7.5.1. (b) 3D Reconstruction of the cameras (red pyramids), the projector (blue pyramid), the calibration-pattern points (red points), and the projected points (blue points) based upon our calibration method.



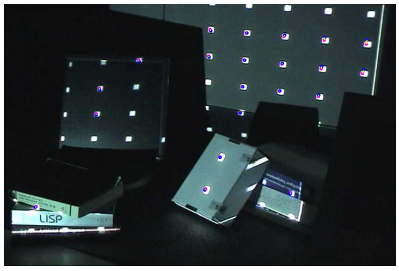
(a)



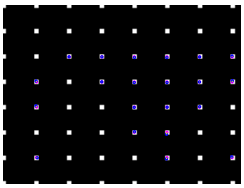
(b)



(c)



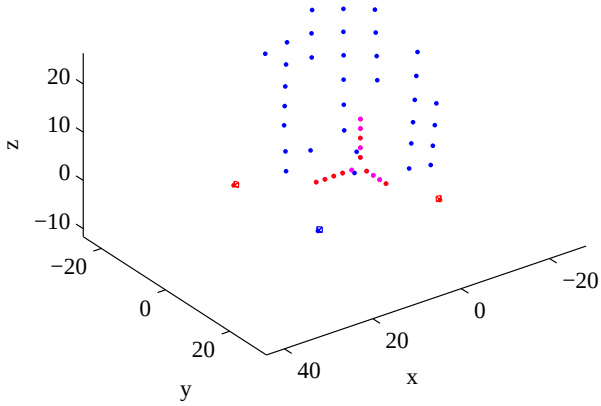
(d)



(e)

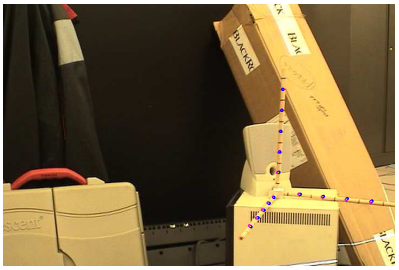
Figure 27: Images from the real-world verification described in Section 7.5.1. Measured point locations are in magenta, the reprojection of the best estimated point locations are in blue, and they are joined by a yellow line. (a) Image from camera 1 of the calibration pattern points. (b) Image from camera 2 of the calibration pattern points. (c) Image from camera 1 of the projected pattern points. (d) Image from camera 2 of the projected pattern points. And (e) image from the projector of the projected pattern points.

(a)



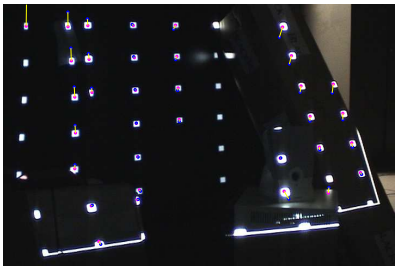
(b)

Figure 28: (a) Image of the overall system setup for the real-world verification described in Section 7.5.2. (b) 3D Reconstruction of the cameras (red pyramids), the projector (blue pyramid), the calibration-pattern points (red points), and the projected points (blue points) based upon our calibration method.

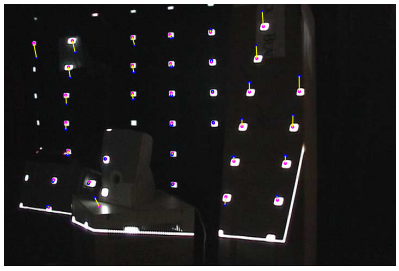


(a)

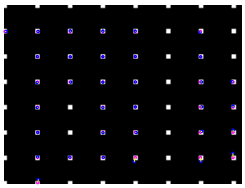
(b)



(c)

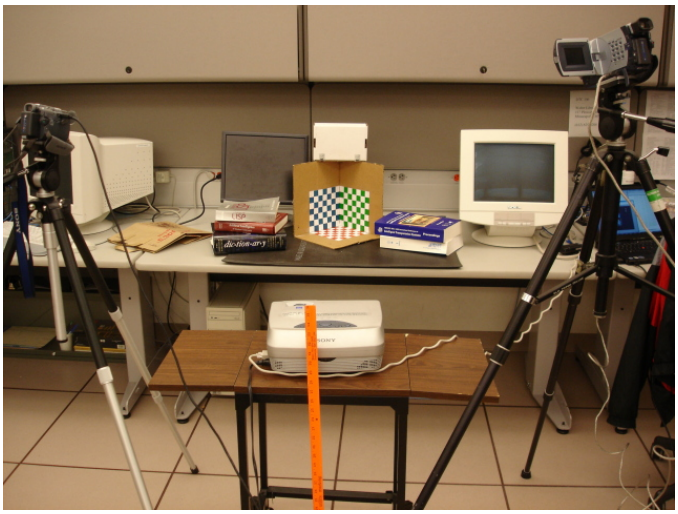


(d)

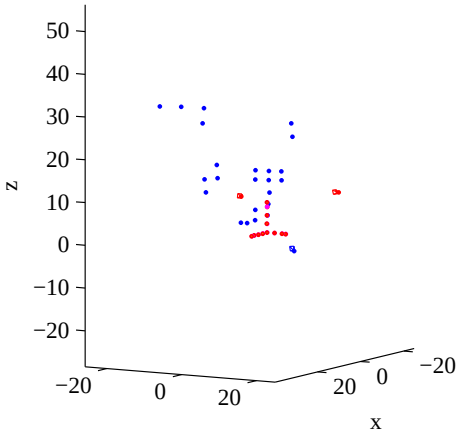


(e)

Figure 29: Images from the real-world verification described in Section 7.5.2. Measured point locations are in magenta, the reprojection of the best estimated point locations are in blue, and they are joined by a yellow line. (a) Image from camera 1 of the calibration pattern points. (b) Image from camera 2 of the calibration pattern points. (c) Image from camera 1 of the projected pattern points. (d) Image from camera 2 of the projected pattern points. And (e) image from the projector of the projected pattern points.



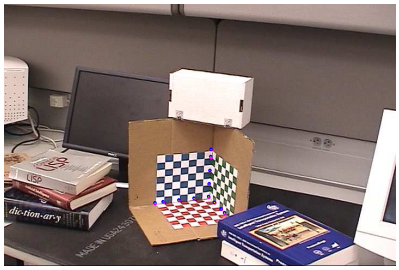
(a)



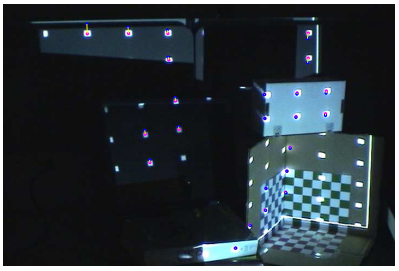
(b)

Figure 30: (a) Image of the overall system setup for the real-world verification described in Section 7.5.3. (b) 3D Reconstruction of the cameras (red pyramids), the projector (blue pyramid), the calibration-pattern points (red points), and the projected points (blue points) based upon our calibration method.

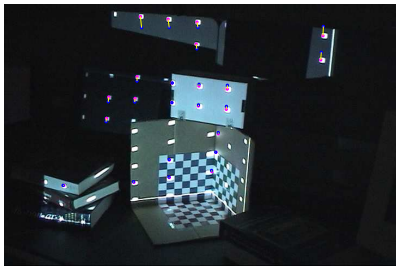
(a)



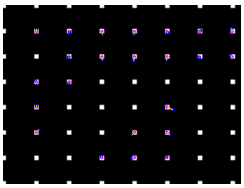
(b)



(c)



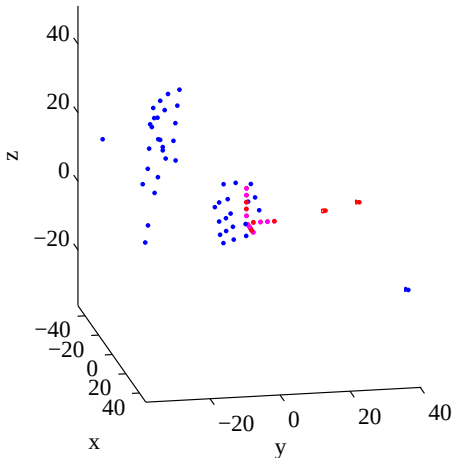
(d)



(e)

Figure 31: Images from the real-world verification described in Section 7.5.3. Measured point locations are in magenta, the reprojection of the best estimated point locations are in blue, and they are joined by a yellow line. (a) Image from camera 1 of the calibration pattern points. (b) Image from camera 2 of the calibration pattern points. (c) Image from camera 1 of the projected pattern points. (d) Image from camera 2 of the projected pattern points. And (e) image from the projector of the projected pattern points.

(a)

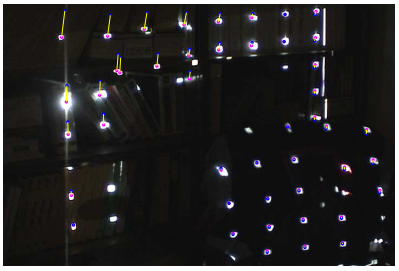


(b)

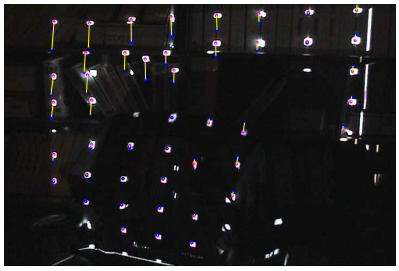
Figure 32: (a) Image of the overall system setup for the real-world verification described in Section 7.5.4. (b) 3D Reconstruction of the cameras (red pyramids), the projector (blue pyramid), the calibration-pattern points (red points), and the projected points (blue points) based upon our calibration method.

(a)

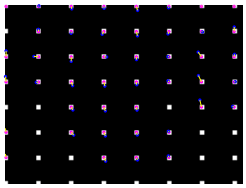
(b)



(c)



(d)



(e)

Figure 33: Images from the real-world verification described in Section 7.5.4. Measured point locations are in magenta, the reprojection of the best estimated point locations are in blue, and they are joined by a yellow line. (a) Image from camera 1 of the calibration pattern points. (b) Image from camera 2 of the calibration pattern points. (c) Image from camera 1 of the projected pattern points. (d) Image from camera 2 of the projected pattern points. And (e) image from the projector of the projected pattern points.

era 1 and Figure 31(b) for camera 2. The projected pattern, with detected and reprojected points is shown in Figure 31(c) for camera 1, Figure 31(d) for camera 2, and Figure 31(e) for the projector. Finally, the overall 3D scene reconstruction for the cameras, projectors, calibration pattern points, and projection pattern points is shown in Figure 30(b).

7.5.4 Real-World Test 4

The overall system setup for this test is shown in Figure 32(a). The calibration pattern, with measured and reprojected points is shown in Figure 33(a) for camera 1 and Figure 33(b) for camera 2. The projected pattern, with detected and reprojected points is shown in Figure 33(c) for camera 1, Figure 33(d) for camera 2, and Figure 33(e) for the projector. Finally, the overall 3D scene reconstruction for the cameras, projectors, calibration pattern points, and projection pattern points is shown in Figure 32(b).

7.5.5 Discussion

The results shown from these real-world tests on these two cluttered scenes show that the relative positions of the cameras and projector are found with respect to the scene. This can be seen by comparing the system setup with the reconstruction based on the results of the calibration method, as shown in Figures 26 and 28.

We note that there are some points for which the reprojection error is higher than for others. This is likely due to effects that were not analyzed in simulation. One such effect is foreshortening of the illuminated area around the projected point. This can be seen most clearly in the points on the slanted box face shown in Figures 29(c) and 29(d). Because of the different foreshortening of the illuminated areas on the surface in each view, there is higher error in the reconstruction of the affected points' locations. Glare occurs in the real-world as well, which also makes it more difficult to accurately locate a point by illuminating more of the scene than is intended.

7.6 Calibration Final Thoughts

In this chapter, we have presented a Euclidean calibration method for camera and projector systems in general scenes. This method is optimal in terms of the reprojection error of scene points. In addition, this method does not require manual human intervention to move specific occlusions so they intersect calibration patterns just right, or painstaking hand measurement of illuminated points.

A set of simulations was run which characterizes the reprojection error of the method with respect to the level of corruption in the camera images. It was shown that the error in the camera and projector method is, on average, a little higher than the error for calibration-pattern based stereo camera calibration, but that its variance increases with increased image noise. Poor placement of the cameras and projector was shown to be an aggravating factor in these situations.

Finally, the real-world efficacy of the method was demonstrated by calibrating a camera and projector system on complex and cluttered scenes.

8 Conclusions

This thesis starts with a motivating problem related to precisely tracking a patient's body surface. This is best done in as minimally invasive manner as possible. In other words, methods which do not need artificial markers attached to the patient are preferred to those that do. Structured light-based vision is well suited for such tasks as it does not need physical markers. The main drawback of structured light, that it needs a controlled lighting environment, is not a problem with medical procedures as lighting can be strictly controlled in medical centers. This thesis then examined, and ultimately expanded upon knowledge of structured light systems.

The following are the central contributions of this thesis:

1. A pair of physics-based, mathematical quality metrics for the placement of elements of structured light camera and projector systems. These quality metrics quantitatively determine how good a 3D reconstruction can be made of a known target for a given system setup. This is different from the placement work in the literature which focuses on stereoscopic camera systems, as camera and projector systems propagate error in a fundamentally different fashion. This work was presented in Chapter 6.

2. A calibration method for structured light camera and projector systems which is optimal in terms of reprojection error. This is different than existing calibration methods in the literature in that no a priori information about the scene structure is required. By incorporating multiple cameras for calibration, constrained scene structured is no longer required for calibration of structured light systems. This work was presented in Chapter 7.

As the central contributions of this thesis focus on structured light systems in general, it is important to note that they are applicable to a much wider range of

applications than merely patient body tracking, or even medical devices. The work presented here is applicable to any task for which structured light is utilized.

8.1 Future Directions

There are avenues of future development which build upon the work presented in this thesis.

The work on camera and projector placement for structured light systems has several possible extensions. Currently, the target model must be convex because a point cloud is used to model the target. This model cannot properly self occlusions that occur when the target is allowed to have concavities. Extending the target object model to utilize a form of surface modeling would be a very useful extension because non-convex objects could then be targeted. In addition, by incorporating surface information into the target object model, the quality the placement could be found with respect to multiple objects, instead of just one.

Developing a method to apply the placement quality metrics to moving targets would also be a meaningful extension, as the patient lies on a moving platform while the helical tomotherapy device is being used. A probability measure modeling could be incorporated in order to quantify how often a given point should be viewable by the cameras and projectors.

Another useful extension to the camera and projector placement is the development of a method which utilizes the quality metrics developed to maximize the quality of a placement setup. The quality metrics are currently only useful for comparing specific, discreet component setups. This optimization problem is difficult because both quality metrics are non-convex in high-dimensional domains, even for simple targets. It may be possible to develop a formulation of the exterior visibility problem that could be used to seed good overall component placements.

Beyond component placement, the development of a body surface model is important for tracking the patient while the helical tomotherapy device is in use. The body model would be created from the point cloud collected by the structured light system and used to determine whether the patient's body has moved out of alignment during treatment. Ideally, the model would be able to take into account the normal periodic

motion of breathing, so false alarms are not thrown based on breathing motion.

Finally, it would be useful to investigate dual purposing the projected patterns from the structured light system to not only aid in 3D reconstruction tasks, but to simultaneously provide augmented reality feedback to the medical personnel using the system. These patterns could be used to monitor the patient as well as communicate information to the system user at the same time. This would be very useful for the initial placement of the patient on the platform before treatment, as immediate feedback can be given directly on the patient, allowing swift adjustments of the patient's pose.

References

[1] A. Agovic, A. Banerjee, A. Ganguly, and V. Protopopescu. *Knowledge Discovery from Sensor Data*, chapter Anomaly Detection in Transportation Corridors Using Manifold Embedding. CRC Press, 2008.

[2] S. Audet and M. Okutomi. A user-friendly method to geometrically calibrate projector-camera systems. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 47–54, June 2009.

[3] R. Azuma. A survey of augmented reality. In *Presence: Teleoperations and Virtual Environments 6*, pages 355–385, August 1997.

[4] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre. Recent advances in augmented reality. *Computer Graphics and Applications, IEEE*, 21(6):34–47, November/ December 2001.

[5] J. Batlle, E. M. Mouaddib, and J. Salvi. Recent progress in coded structured light as a technique to solve the correspondence problem. *Pattern Recognition*, 31(7):963–982, 1998.

[6] H. Bay, T. Tuytelaars, and L. V. Gool. SURF: Speeded up robust features. In *Proceedings of the Ninth European Conference on Computer Vision*, pages 404–417, May 2006.

[7] F. Betting, J. Feldmar, N. Ayache, and F. Devernay. A new framework for fusing stereo images with volumetric medical images. In *Proceedings of Computer Vision, Virtual Reality, and Robotics in Medicine*, pages 30–39, 1995.

[8] N. Bird, S. Atev, N. Caramelli, R. Martin, O. Masoud, and N. Papanikolopoulos. Real time, online detection of abandoned objects in public areas. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3775–3780, May 2006.

[9] N. Bird, O. Masoud, N. Pananikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):167–177, June 2005.

[10] A. Bissacco, M. H. Yang, and S. Soatto. Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2007.

[11] R. Bodor, A. Drenner, P. Schrater, and N. Papanikolopoulos. Optimal camera placement for automated surveillance tasks. *Journal of Intelligent and Robotic Systems*, 50:257–295, 2007.

[12] R. Bodor, P. Schrater, and N. Papanikolopoulos. Multi-camera positioning to optimize task observability. In *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2005.

[13] J.-Y. Bouguet and P. Perona. 3d photography on your desk. In *International Conference on Computer Vision*, pages 43–50, Jan 1998.

[14] D. Caspi, N. Kiryati, and J. Shamir. Range imaging with adaptive color structured light. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):470–480, May 1998.

[15] C. Che and J. Ni. Modeling and calibration of a structured-light optical CMM via skewed frame representation. *Journal of Manufacturing Science and Engineering*, 118(4):595–603, 1996.

[16] C. Chen and A. Kak. Modeling and calibration of a structured light scanner for 3-D robot vision. In *IEEE International Conference on Robotics and Automation*, volume 4, pages 807–815, Mar 1987.

[17] S. Chen and Y. Li. Self recalibration of a structured light vision system from a single view. In *IEEE International Conference on Robotics and Automation*, volume 3, pages 2539–2544, 2002.

[18] X. Chen and J. Davis. Camera placement considering occlusion for robust motion capture. Technical Report CS-TR-2000-07, Stanford Department of Computer Science, 2007.

[19] J. Chihak, J. Cumpelik, L. Krticka, V. Smutny, P. Strnad, R. Sara, F. Vele, M. Veverkova, and V. Zyka. Recognition of breathing pattern by a photogrammetric method. Technical Report CTU-CMP-2005-32, Czech Technical University in Prague, 2006.

[20] C. W. Chu, S. Hwang, and S. K. Jung. Calibration-free approach to 3d reconstruction using light stripe projections on a cube frame. In *Third International Conference on 3-D Digital Imaging and Modeling*, pages 13–19, 2001.

[21] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):781–796, August 2000.

[22] Z. Duric, F. Li, Y. Sun, and H. Wechsler. Using normal flow for detection and tracking of limbs in color images. In *IEEE International Conference on Pattern Recognition*, pages 268–271, 2002.

[23] P. J. Edwards, D. L. G. Hill, D. J. Hawkes, R. Spink, A. C. F. Colchester, A. J. Strong, and M. J. Gleeson. Neurosurgical guidance using the stereo microscope. In *Proceedings of Computer Vision, Virtual Reality, and Robotics in Medicine*, pages 555–564, April 1995.

[24] D. Fofi, E. M. Mouaddib, and J. Salvi. How to self-calibrate a structured light sensor? In *International Symposium on Intelligent Robotic Systems*, 2001.

[25] D. Fofi, J. Salvi, and E. Mouaddib. Uncalibrated vision based on structured light. In *IEEE International Conference on Robotics and Automation*, volume 4, pages 3548–3553 vol.4, 2001.

[26] T. Fox, E. L. Simon, E. Elder, R. H. Riffenburgh, and P. A. Johnstone. Free breathing gated delivery (FBGD) of lung radiation therapy: Analysis of factors affecting clinical patient throughput. *Lung Cancer*, 56(1):69 – 75, 2007.

[27] A. Fusiello, E. Trucco, and A. Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12:16–22, 2000.

[28] G. Gasser, N. Bird, O. Masoud, and N. Papanikolopoulos. Human activities monitoring and bus stops. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 1, pages 90–95, April 2004.

[29] W. E. L. Grimson, G. J. Ettinger, S. J. White, T. Lozano-Perez, W. M. W. III, and R. Kikinis. An automatic registration method for frameless stereotaxy, image guided surgery, and enhanced reality visualization. *IEEE Transactions on Medical Imaging*, 15(2):129–140, April 1996.

[30] L. Haritaoglu and D. H. L. Davis. W4: Real-time surveillance of people and their activities. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):809–830, August 2000.

[31] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2003.

[32] S. K. Hui, J. Kapatoes, and J. Fowler. Feasibility study of helical tomotherapy for total body or total marrow irradiation. *Medical Physics*, 32(10):10, 2005.

[33] S. K. Hui, M. R. Verneris, and P. Higgins. Helical tomotherapy targeting total bone marrow–first clinical experience at the University of Minnesota. In *Acta Oncologica*, October 2004.

[34] D. Huynh. Calibration of a structured light system: a projective approach. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 225–230, Jun 1997.

[35] D. Q. Huynh, R. A. Owens, and P. E. Hartmann. Calibrating a structured light stripe system: A novel approach. *International Journal of Computer Vision*, 33(1):73–86, Sept. 1999.

[36] S. Inokuchi, K. Sata, and F. Matsuda. Range imaging system for 3-d object recognition. In *Proceedings of the International Conference on Pattern Recognition*, 1984.

[37] V. Isler, S. Kannan, K. Daniilidis, and P. Valtr. VC-dimension of exterior visibility. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):667–671, 2004.

[38] O. Jokinen. Self-calibration of a light striping system by matching multiple 3-D profile maps. In *International Conference on 3-D Digital Imaging and Modeling*, pages 180–190, 1999.

[39] X. Kai, L. W. Yu, and P. Zhao-Bang. The hybrid calibration of linear structured light system. In *IEEE International Conference on Automation Science and Engineering*, pages 611–614, Oct. 2006.

[40] A. Kelly. Precision dilution in triangulation-based mobile robot position estimation. In *Intelligent Autonomous Systems*, 2003.

[41] H. S. Kim, S. Ishikawa, Y. Ohtsuka, H. Shimizu, T. Shinomiya, and M. A. Viergever. Automatic scoliosis detection based on local centroids evaluation on moiré topographic images of human backs. *IEEE Transactions on Medical Imaging*, 20(12):1314–1320, December 2001.

[42] T. P. Koninckx, P. Peers, P. Dutr, and L. V. Gool. Scene-adapted structured light. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 611–618, June 2005.

[43] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, February 2008.

[44] D. Lanman, D. Crispell, and G. Taubin. Surround structured lighting for full object scanning. In *Proceedings of the IEEE Conference on 3-D Digital Imaging and Modeling*, 2007.

[45] M. W. Lee and I. Cohen. A model-based approach for estimating human 3D poses in static images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):905–916, June 2006.

[46] Y. Li and S. Chen. Automatic recalibration of an active structured light vision system. *IEEE Transactions on Robotics and Automation*, 19(2):259–268, Apr 2003.

[47] J. Liao and L. Cai. A calibration method for uncoupling projector and camera of a structured light system. In *IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pages 770–774, July 2008.

[48] N. Linthout, D. Verellen, K. Tournel, and G. Storme. Six dimensional analysis with daily stereoscopic x-ray imaging of intrafraction patient motion in head and neck treatments using five points fixation masks. *Medical Physics*, 33(2):504–513, February 2006.

[49] M. A. Livingston. *Vision-Based Tracking with Dynamic Structured Light for Video See-Through Augmented Reality*. PhD thesis, University of North Carolina at Chapel Hill, 1998.

[50] W. Lorensen, H. Cline, C. Nafis, R. Kikinis, D. Altobelli, and L. Gleason. Enhancing reality in the operating room. In *Proceedings of the IEEE Conference on Visualization*, pages 410–415, 1993.

[51] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 20:91–110, 2003.

[52] R. I. MacKay, P. A. Graham, J. P. Logue, and C. J. Moore. Patient positioning using detailed three-dimensional surface data for patients undergoing conformal radiation therapy for carcinoma of the prostate: a feasibility study. *International Journal of Radiation Oncology Biology Physics*, 39(1):225–230, 2001.

[53] A. McIvor. Calibration of a laser stripe profiler. In *International Conference on 3-D Imaging and Modeling*, pages 92–98, 1999.

[54] A. M. McIvor. Nonlinear calibration of a laser stripe profiler. *Optical Engineering*, 41(1):205–212, 2002.

[55] S. L. Meeks, W. A. Tome, T. R. Willoughby, P. A. Kupelian, T. H. Wagner, J. M. Buatti, and F. J. Bova. Optically guided patient positioning techniques. In *Seminars in Radiation Oncology*, volume 15, pages 192–201, July 2005.

[56] J. P. Mellor. Enhanced reality visualization in a surgical environment. Technical Report 1544, Massachusetts Institute of Technology Artificial Intelligence Laboratory, January 1995.

[57] B. D. Milliken, S. J. Rubin, R. J. Hamilton, L. S. Johson, and G. T. Y. Chen. Performance of a video-image subtraction-based patient positioning system. *International Journal of Radiation Oncology Biology Physics*, 38(4):855–866, 1997.

[58] A. Mittal and L. S. Davis. A general method for sensor planning in multi-sensor systems: Extension to random occlusion. *International Journal of Computer Vision*, 76(1):31–52, 2008.

[59] C. J. Moore and P. A. Graham. 3D dynamic body surface sensing and ct-body matching: a tool for patient set-up and monitoring in radiotherapy. *Computer Aided Surgery*, 5:234–245, 2000.

[60] G. Olague and R. Mohr. Optimal camera placement to obtain accurate 3D point positions. In *International Conference on Pattern Recognition*, pages 8–10, 1998.

[61] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs. The office of the future: A unified approach to image-based modeling and spatially immersive displays. In *International Conference on Computer Graphics and Interactive Techniques*, pages 179–188, 1998.

[62] G. T. Reid. Automatic fringe pattern analysis–a review. *Optics and Lasers in Engineering*, 7(1):37–68, 1986–1987.

[63] I. D. Reid. Projective calibration of a laser-stripe range finder. *Image and Vision Computing*, 14(9):659 – 666, 1996.

[64] J. Salvi, J. Pagés, and J. Batlle. Pattern codification strategies in structured light systems. *Pattern Recognition*, 37(4):827–849, April 2004.

[65] G. Sansoni, M. Carocci, and R. Rodella. Calibration and performance evaluation of a 3-D imaging sensor based on the projection of structured light. *IEEE Transactions on Instrumentation and Measurement*, 49(3):628–636, Jun 2000.

[66] T. C. Shermer. Recent results in art galleries. *Proceedings of the IEEE*, 80(9):1384–1399, September 1992.

[67] D. Shin and J. Kim. Point to point calibration method of structured light for facial data reconstruction. In *International Conference on Biometric Authentication*, pages 200–206, 2004.

[68] M. Siddiqui and G. Medioni. Real time limb tracking with adaptive model selection. In *IEEE International Conference on Pattern Recognition*, 2006.

[69] R. Sundareswara and P. Schrater. Bayesian modelling of camera calibration and reconstruction. In *Proceedings of the International Conference on 3-D Digital Imaging and Modeling*, pages 394–401, June 2005.

[70] H. Takasaki. Moiré topography. *Applied Optics*, 9(6):1467–1472, June 1970.

[71] K. Tarabanis, R. Tsai, and A. Kaul. Computing occlusion-free viewpoints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(3):279–292, March 1996.

[72] G. Taubes. Surgery in cyberspace. *Discover*, 15(12):84–94, December 1994.

[73] O. Tekdas and V. Isler. Sensor placement algorithms for triangulation based localization. In *IEEE International Conference on Robotics and Automation*, pages 4448–4453, April 2007.

[74] V. E. Theodoracatos and D. E. Calkins. A 3-D vision system model for automatic object surface sensing. *International Journal of Computer Vision*, 11(1):75–99, 1993.

[75] P. Treleaven and J. Wells. 3D body scanning and healthcare applications. *Computer*, 40(7):28–34, July 2007.

[76] H. Veeraraghavan, S. Atev, N. Bird, P. Schrater, and N. Papanikolopoulos. Driver activity monitoring through supervised and unsupervised learning. In *Proceedings of the IEEE Conference on Intelligent Transportation Systems*, pages 580–585, September 2005.

[77] H. Veeraraghavan, N. Bird, S. Atev, and N. Papanikolopoulos. Classifiers for driver activity monitoring. *Transportation Research Part C*, 15(1):51–67, February 2007.

[78] J. Wakitani, T. Maruyama, T. Morita, T. Uchiyama, and A. Mochizuki. Wrist-mounted laser rangefinder. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 3, pages 362–367 vol.3, Aug 1995.

[79] C. W. Wang, A. Ahmed, and A. Hunter. Vision analysis in detecting abnormal breathing activity in application to diagnosis of obstructive sleep apnoea. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4469–4473, 2006.

[80] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[81] K. Yamauchi, H. Saito, and Y. Sato. Calibration of a structured light system by observing planar object from unknown viewpoints. In *International Conference on Pattern Recognition*, pages 1–4, Dec. 2008.

[82] B. Zhang and Y. Li. Dynamic calibration of a structured light system via planar motion. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 133–138, Aug. 2005.

[83] B. Zhang, Y. F. Li, and Y. H. Wu. Self-recalibration of a structured light system via plane-based homography. *Pattern Recognition*, 40(4):1368–1377, 2007.

[84] F. Zheng and B. Kong. Calibration of linear structured light system by planar checkerboard. In *International Conference on Information Acquisition*, pages 344–346, June 2004.