Simple Stereo Vision for Image Guided Radiation Therapy Research

Cesare Jenkins

Abstract

A method for simple stereo vision reconstruction is proposed specifically for use in a developmental environment for image guided radiation therapy. The method employs feature matching, fundamental matrix calculation to narrow matches and image rectification as precursors to an auto-calibration implementation. Experiments showed that the method still requires significant improvement before it is useable while also offering important insight into the necessity of increasing the match-able features in the acquired images.

1. Introduction

Radiation therapy has been utilized for decades in treating cancer. The most common form of radiation therapy delivery is known as external beam radiation therapy in which a medical linear accelerator is used to deliver high-energy photons to a targeted region within a patient's body. The delivery of the radiation is carefully planned to deliver sufficient radiation to eradicate malignant cells while sparing healthy tissue as much as possible. The success or failure of therapy depends heavily on the patient being in the correct position during treatment. To this end, a host of techniques have been employed to ensure proper alignment between the patient's actual and planned location. These techniques include patient immobilization, x-ray imaging and most recently optical tracking technologies.

Over the last two years an optical tracking system known as Vision RT has been clinically implemented. The system utilizes a set of ceiling mounted 3D cameras to create a surface corresponding to the patient. This surface is then registered (aligned) to the patient planning CT and the position is compared to ensure that the patient is in the correct location and remains there throughout the duration of the treatment.

As we work to extend the capabilities of these types of systems it would be extremely useful to have available a simple 3D reconstruction algorithm that would be able to reconstruct stereo views without extensive calibration. This would be particularly useful as we work with different systems within the room. Therefore the goal of this project is to develop a method to quickly and simply reconstruct stereo views from previously recorded image sequences.

2.1 System Requirements and Challenges

The main role of this algorithm is to allow us to quickly asses the value of some of the techniques for monitoring radiation therapy that we are investigating in our lab. Since time in the therapy rooms is limited to after hours and weekends, it is necessary for us to be able to set up an experiment, collect data and remove our equipment from the room all within a matter of a few hours (unless we want to stay either all night or all weekend). Additionally, in most cases, we only have one camera available. This means that we have to set up a scene, image it from one location, move the camera and image it again. In addition to the necessary mobility of the setup, it is also very constrained in terms of space, as the treatment rooms are relatively small and somewhat cluttered. Taken together, these factors make it infeasible to perform calibration using many of the techniques generally employed in the computer vision world. Many of these rely on simultaneous capture of images from two cameras and the use of carefully placed calibration grids. It would be ideal if an algorithm could estimate a 3D reconstruction from two views without any additional information.

The major challenges of working with camera systems, specifically lighting and image quality, are greatly magnified when working in radiation therapy rooms. In most instances, light levels are kept relatively low for the comfort of patients. In general therapists prefer to be able to adjust the light levels for each patient, introducing additional variation to the lighting levels in the room. Additionally, for some of the specific applications in our lab, it is necessary for us to detect very small amounts of light. The result is that we often operate with an extremely large aperture setting (sometimes as low of f/0.95) and high gain on our cameras. The result is that our images often have a very narrow depth of field and contain significant amounts of noise. Also, factors other than overall image quality often dictate the exposure settings used for the camera. Hence, it is not uncommon for the images to generally be either slightly under-exposed. This combined with the fact that the scenes we are interested in are either highly uniform or have repetitive patterns makes any sort of image processing difficult, but especially attempts to perform reconstruction.

2.1 Review of Previous Work

As stated above, Vision RT is a commercially available system made up of two ceiling mounted 3D cameras and software to locate and track patients during radiation therapy. While the system is certainly capable of much more, the current clinically available functions include patient alignment and tracking of patient motion to ensure that the patient remains within a specified envelope during treatment. If the patient moves outside of this envelope (a common example being that the patient starts coughing) the system detects this motion and halts the radiation therapy until the patient has returned to the specified envelope.

Other optical tracking systems have also been employed to track the breathing motion of patients receiving therapy to regions of the body that experience significant movement during a breathing cycle (i.e. the lungs or adjacent organs, interestingly enough, lung cancer is one of the most common cancers treated with this type of therapy). Many of these systems include a special marker/fiducial that is placed on the patient's chest. Some markers include infrared LEDs or highly reflective spots that enable them to be easily tracked by one or more cameras in the room. Since the goal of these systems is to simply estimate where the patient is in his breathing cycle (i.e. are the lungs full or empty), there is no need for 3D reconstruction.

Finally, the simplest vision system that is actually required by law to be operational whenever a therapy session is active, is a basic video and audio feed of the room to the control console. This enables the therapists (the individuals who actually operate the machines) to monitor what is going on inside. There is some work being done (mentioned at a recent symposium, but not yet published) to implement computer vision techniques at this "whole room" level to identify and track things in the room in order to identify problems that a person might miss. Errors include positioning the patient incorrectly, not installing or installing the incorrect accessory on the linear accelerator or leaving an object in a location where the linear accelerator will collide with it. Similar to the Vision RT system, this system would also be implemented using depth cameras.

On the computer vision side of things, auto-calibration has been a field of interest for sometime (see Faugeras or Hartley). Most methods involve either a structure from motion approach or utilizing a priori knowledge about the scene (for example utilizing known flat surfaces). However, all of these methods require relatively precise point location extraction and matching across views. Given the specific challenges of our operating environment it seems safe to assume that the major challenge lies in obtaining sufficiently clean data to make use of one of the previous implementations.

2.2 Improvements

Our method will seek to identify robust and repeatable matches in the face of shallow depth of field and high noise images. Originally within the scope of the project was the actual implementation of the auto-calibration. However, as will become clear in the experiments section of this report, it became very clear that the project would be limited to feature extraction and matching.

3.1 Overview of the Technical Implementation

The general approach taken for feature extraction and matching was to

1. Load and preprocess images in order to minimize the effects of noise and exposure settings
2. Identify SURF features in the image

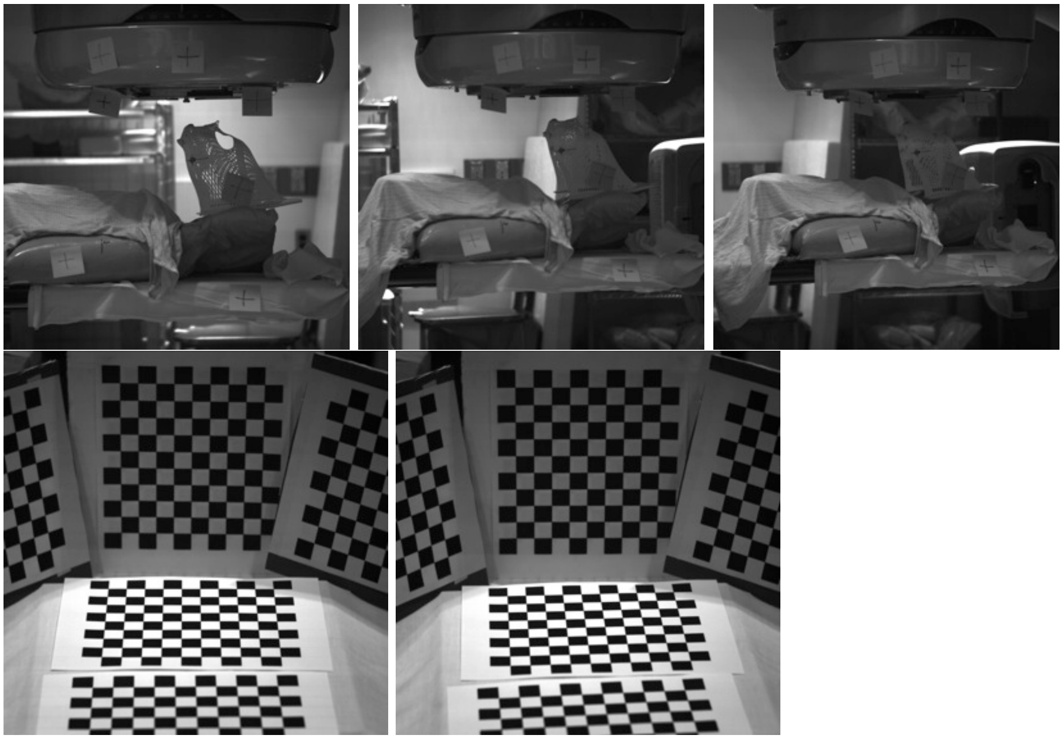
3. Perform a first pass feature matching
4. Utilize RANSAC and Fundamental Matrix estimation to refine matches
5. Perform image rectification to evaluate the sufficiency of the matches

While a more robust approach would certainly be to perform the full auto-calibration and compare results to a ground truth estimate, it quickly became evident that the algorithm was not performing well enough to justify this extended implementation.

3.2 Technical Implementation Details

Image Collection

Three images of a simulated treatment setup were acquired using a machine vision camera (Basler Inc.) equipped with a 50mm f/1.4 lens (Pentax). The camera is equipped with a high speed 1” class CMOS sensor that has a resolution of 2048 x 2048 pixels. The images were captured using the Pylon Viewer software provided with the camera (Basler Inc.). The images used for the project are shown in Figure 1.



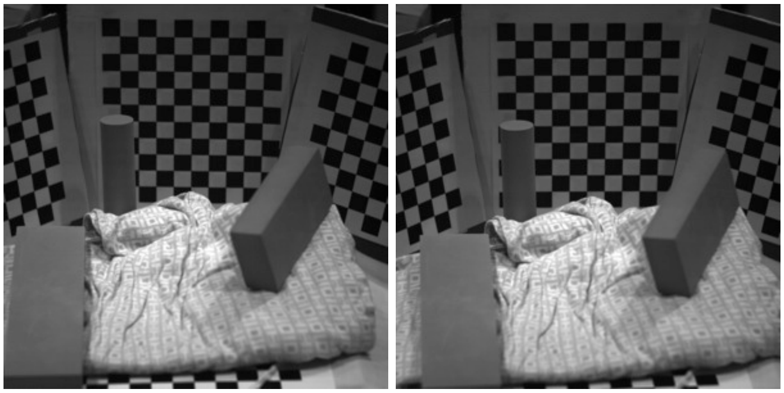


Figure 1. Images acquired for the project.

Initial Evaluation Algorithm Development

The first step taken in the project was to manually identify several correspondences across the acquired images and compute the fundamental matrix and epipolar lines. By looking at the distance between the points and the epipolar lines, as well as observing the convergence of the calculated epipolar lines would offer us a quick check of how well later correspondences were performing. We also implemented an affine structure from motion calculation using the factorization method presented by Tomasi and Kanade as it was our intent to utilize this later as a method to get a first of the camera matrices.

Current Implementation

The following breaks down the method currently implemented in the project code. The images are loaded and have the histograms equalized. This increased the contrast of the images and improves feature extraction. Next the images undergo a median filter to minimize noise. There is a clear tradeoff between the minimization of noise the blurring of useful features in the image.

Next SURF features are extracted from the image. These features are then matched using a nearest neighbor algorithm that includes the nearest match ratio as a filtering parameter. At this point we estimate the fundamental matrix and utilize RANSAC to eliminate outlier matches across the two images.

Once the fundamental matrix has been calculated we can use it to perform an estimated rectification of the images. We can then plot the correspondences and evaluate the performance of the algorithm by looking at the rectification and parallelism of the plotted correspondences.

Once the images are rectified we search for additional correspondences using a semi-local block matching algorithm. The results are displayed as a disparity map. This offers a quick assessment of how well the algorithm can subsequently identify additional matches.

Experimental Implementations

In order to increase the number of high quality matches passed to the RANSAC algorithm we implemented a stepped thresholding algorithm. Essentially the images would be each be converted to binary images with all pixels above a certain value being converted to white and all those below converted to black. We then extract features from these binary images (we tried SURF features and various corner extractors). This process is repeated for several different threshold values and all of the points are passed to the matching algorithm.

Another method that was implemented was an iterative approach to finding matches. The current implementation above was essentially repeated using the disparity matches map matches as the input instead of the nearest neighbor SURF feature matches.

4. Experiments

Figure 2 shows the results of the initially implemented “quick-check” algorithm using the manually identified correspondences. Figure 3 is a simple representation of the calculated affine structure from motion 3D reconstruction of the images. At this point in the process things looked fairly promising. The epi-polar lines in Figure 2 appear quite reasonable and the 3D reconstruction, though rough appears true to life. Average distances between points and epipolar lines were approximately 4 pixels.

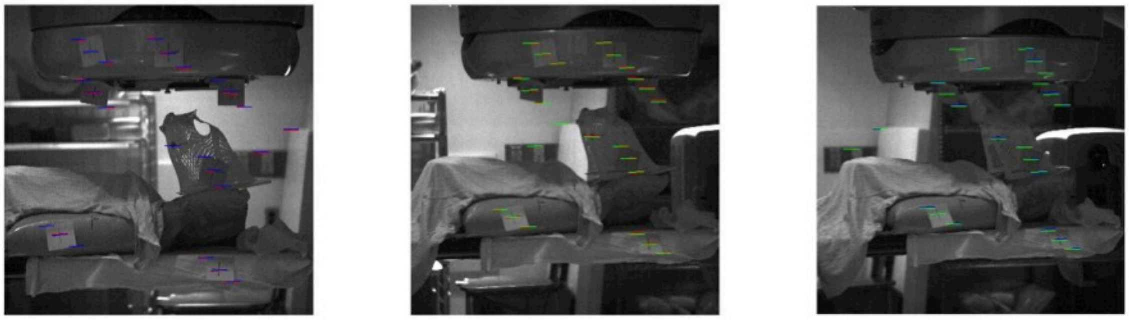


Figure 2. Correspondences and epipolar lines.

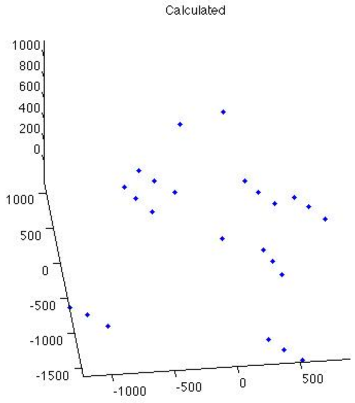


Figure 3. 3D points reconstructed from correspondences.

Figure 4 is an image of the initial SURF feature correspondences identified by the nearest neighbor search on the same images that the manual correspondences were extracted from.

Figure 5 is a rectified view of the images showing the remaining correspondences after performing RANSAC. It is fairly clear from these images that the number of surviving correspondences is too low. It should also be noted that even at this highly selective level it is not uncommon to see a spurious match. When one does occur, the outlier unacceptably skews the calculated rectification. Also, the fundamental matrix obtained by this method differed significantly (both in numbers and sign) from the one obtained with the initial test algorithm.

Figure 6 is an image of the initial SURF feature correspondences for an image that contains calibration grids. Figure 7 is a rectified version of the images similar to that shown in Figure 5. It is fairly clear that the changes to the scene, as well as a smaller baseline for the cameras, dramatically increases the ability of the system to identify correspondences.

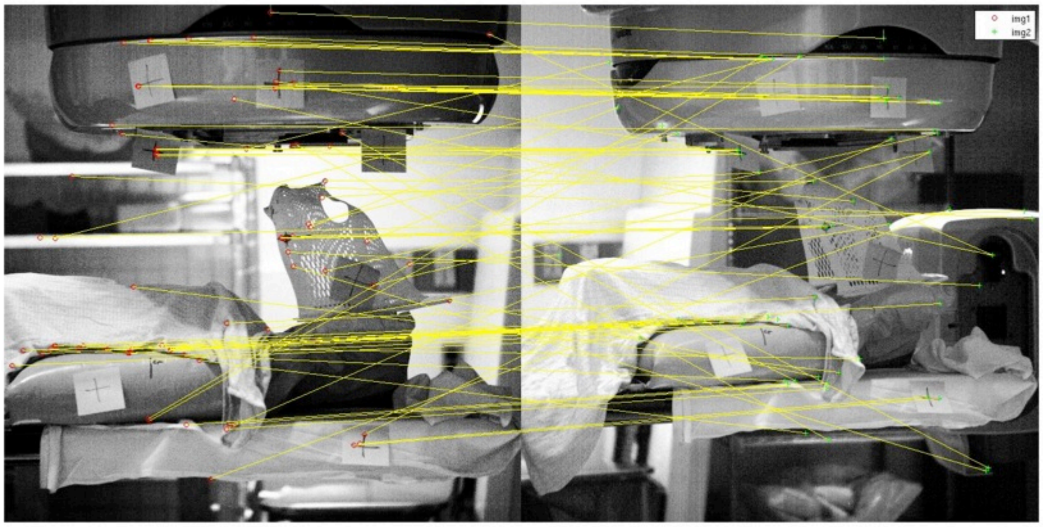


Figure 4. Initial nearest neighbor matches.

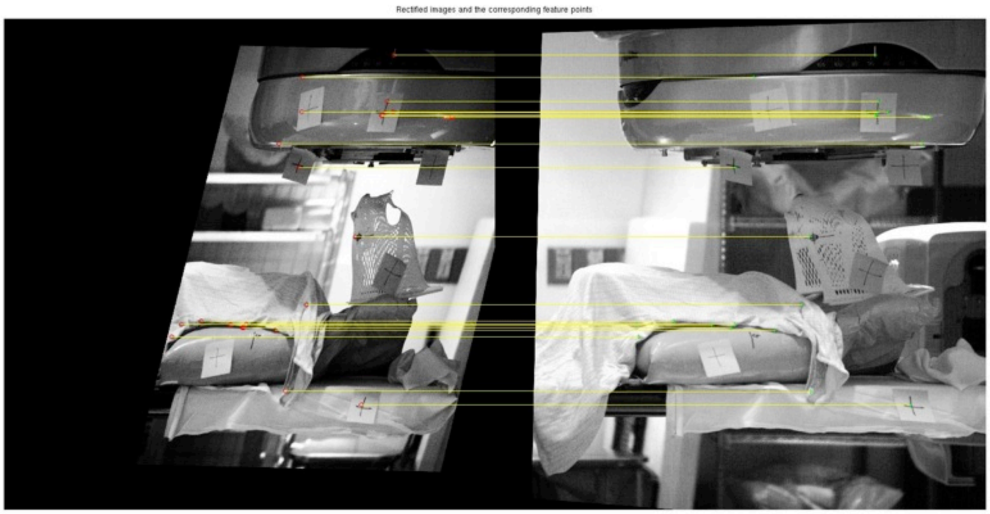


Figure 5. Estimated rectification and correspondences.

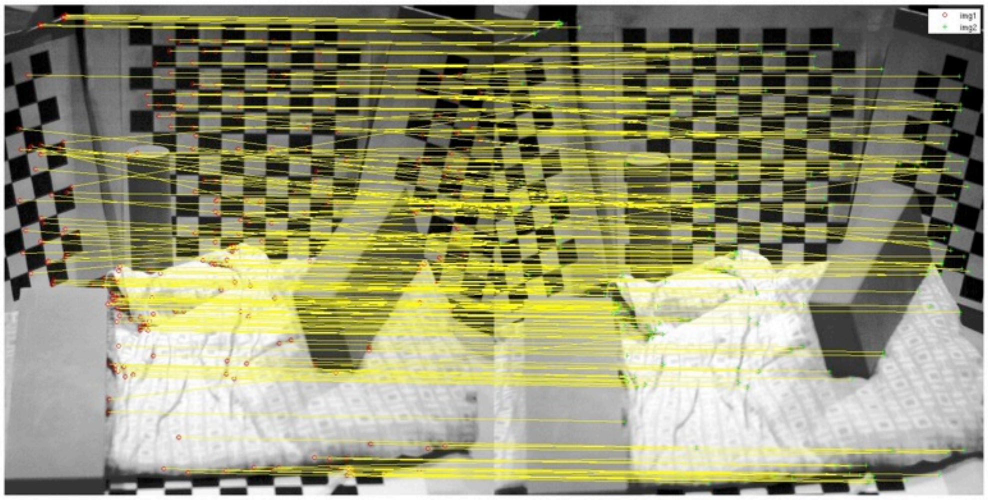


Figure 6. Initial nearest neighbor matches.

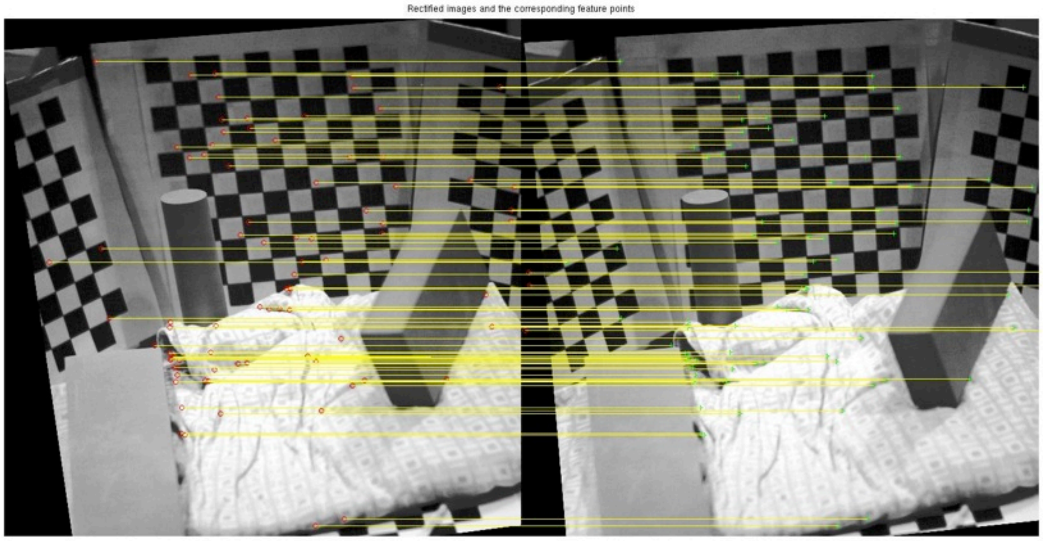


Figure 7. Rectified images with correspondences identified. These images are consistent with rectified images produced by the Camera Calibration Toolbox for Matlab.

It was found that a window size of 2 was optimal for the median filter. Window sizes larger than this dramatically decreased the identification of SURF features. The thresholding method did not yield a significant increase in the number of matches. The iterative approach was extremely time consuming, but also did not yield a significant (if any) improvement in performance.

Further experiments were performed to establish a baseline calibration for the images seen on the second row of Figure 1. Using the Camera Calibration Toolbox for Matlab, the checkerboard patterns were manually identified and a calibration was performed. The results were reasonable but included uncertainties of greater than 10% for each parameter. This clearly indicates a need for further work improving either image acquisition or preprocessing.

Also of note is that the affine structure from motion estimation consistently fails when applied to the correspondences identified with this method. The results behave as if all of the correspondences lie on a single plane (only two "large" singular values are produced). While we have been unable to determine whether this is a result of insufficient feature matching or a bug in the SFM code, it is clear that this portion of the code will need additional attention if it is to be used in further development.

5. Conclusions

It is perhaps a minor understatement to say that the most prominent conclusion of this project is that it proved to be far more difficult than originally anticipated. Initial experiments seemed to indicate that the method would perform well. However when an attempt was made to automate the process of identifying correspondences, as was done in the initial experiments, it became quite clear that a great deal of additional attention needed to be focused on this portion of the process.

While some progress has been made, we have yet to find a robust means of consistently identifying correspondences in an entirely unmodified scene. Hence, for us, the biggest takeaway is that it is unfortunately necessary for us to increase the contrast of our scenes such that an algorithm can perform the identification (the alternative of course is to perform this manually as we did in the initial experiments). We also plan on performing additional experiments to determine the range of baseline distances we can tolerate and which ones would be optimal.

Going forward we plan on continuing to develop this project. We may continue to experiment with additional preprocessing methods and finish implementing a full auto-calibration method. One potential method we may harness going forward is to carefully calibrate the camera in another setting using a checkerboard pattern in order to obtain a good estimate of the camera intrinsic parameters, and then utilize this to aid us in our auto-calibration routine. One foreseeable challenge with this method is that changes in the focus of the camera lens cause changes in the intrinsic parameters. We may consider calibrating the camera at different focus settings or simply lock the focus, calibrate the camera and adjust its location in the room to obtain the focus we need.

6. References

N. Smith, I. Meir, G. Hale, R. Howe, L. Johnson, P. Edwards, D. Hawkes, M. Bidmead, D. Landau. Real-Time 3D Surface Imaging for Patient Positioning in Radiotherapy. International Journal of Radiation Oncology Biology Physics: 57(2):S187; October 2003

P.J. Schöffel, W. Harms, G. Sroka-Perez, W. Schlegel and C.P. Karger. Accuracy of a commercial optical 3D surface imaging system for realignment of patients for radiotherapy of the thorax. Physics in Medicine and Biology; 52: 3949-3963; June 2007.

J L Peng, D Kahler, J G. Li, et al. Characterization of a real-time surface image-guided stereotactic positioning system. Medical Physics, 37(10); pp. 5421-5433; October 2010.

O.D. Faugeras, Q.T. Luong, S.J. Maybank (1992). "Camera Self-Calibration: Theory and Experiments"

R. Hartley and A. Zisserman (2003). Multiple View Geometry in computer vision. Cambridge University Press.

C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. IJCV, 9(2):137-154, November 1992.

Hirschmuller, H., Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information, International Conference on Computer Vision and Pattern Recognition, 2005.