MOTION ESTIMATION FOR FUNCTIONAL MEDICAL IMAGING STUDIES USING A STEREO VIDEO HEAD POSE TRACKING SYSTEM

by

William Pak Tun Ma

BSc., Computer Science, University of British Columbia, 2004

BSc., Mathematics, University of British Columbia, 2007

A Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
in the School
of
Computing Science

SIMON FRASER UNIVERSITY
Summer 2009

APPROVAL

Name: William Pak Tun Ma

Degree: Master of Science

Title of Thesis: Motion Estimation for Functional Medical Imaging Studies Using a Stereo Video Head Pose Tracking System

Examining Committee: Dr. Ze-Nian Li
Chair

Dr. Ghassan Hamarneh
Assistant Professor, Computing Science
Co-senior Supervisor

Dr. Greg Mori
Assistant Professor, Computing Science
Co-senior Supervisor

Dr. Mark Drew
Professor, Computing Science
Examiner

Date Approved: July 17, 2009

Declaration of
Partial Copyright Licence

Abstract

Patient motion is unavoidable during long medical imaging scan times. In particular, motion artifacts in functional and molecular brain imaging (*e.g.*, dynamic positron emission tomography in dPET) are known to corrupt the data leading to inaccurate analysis and diagnosis. Most existing motion correction solutions either rely on attaching external markers or on data-driven image registration algorithms. In this work, we propose a new motion correction approach. It alleviates the need for inconvenient external markers and relaxes the dependence on the fragile similarity metrics that are generally incapable of capturing the complex spatio-temporal tracer dynamics in dPET. We develop a hybrid, multi-sensor method that uses a marker-free video tracker, along with image-based registration. The balance between the two is automatically adapted to confide in the more certain measurement. Our quantitative results demonstrate improved motion estimation and kinetic parameter extraction when using our hybrid method.

Keywords: motion correction; registration; functional medical image; dPET; head tracking; markerless; polaris; positron emission tomography

Acknowledgments

My deepest thanks go foremost to my supervisors, Professor Ghassan Hamarneh and Professor Greg Mori. Their continual support has been very valuable during my research. This project and thesis would not have been completed without their guidance.

I am also indebted to the people in UBC-TRIUMP Positron Emission Tomography Group. Katie Dinelle has been a great help in all major parts of the project, and I am grateful for her help in the motion tracking experiment. Thanks also go to Dr. Vesna Sossi for helping during the preparation period of the Medical Imaging Conference, and for presenting our work.

Members of the MIAL have provided numerous ideas. I would especially like to thank Lisa Tang, for her expertise and help with the registration framework, and Ahmed Saad, for his valuable suggestions in kinetic modeling and his help with the synthetic experiment. I have also enjoyed the time with the various members in the Vision and Media Lab, and especially Bo Chen, Johnson Chuang, Jiawei Huang, Mohammad Norouzi, Yan Tan, Yang Wang, Weilong Yang, and Ziming Zhang. Graduate studies would not have as interesting without them.

Lastly, I would like to thank my family for their love and support over the years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The development of various medical imaging modalities has allowed radiologists a wide selection of tools to create images of the human body and diagnose disease. Some of these modalities create a one-time snapshot of the target area in either 2D or 3D, and are mainly used to analyze the internal structure. Example modalities in this category include X-Rays, Computed Tomography (CT), and Magnetic Resonance Imaging (MRI). Other modalities allow physicians to assess body functions, such as the brain in normal and diseased states. This type of modality can vary from techniques that do not create any actual images, such as Magnetoencephalography (MEG) or Electroencephalography (EEG), to techniques that create full 3D volumes, such as functional Magnetic Resonance Imaging (fMRI), Single Photon Emission Computed Tomography (SPECT), or Positron Emission Tomography (PET). These functional modalities usually require a set period of time for retrieval, and, for the modalities that create full 3D volumes, are highly susceptible to patient movement.

The aim of this thesis is to develop a new approach for tracking patient movement and for motion correction. Our development primary focuses on PET brain images. A more detailed overview will be given in section 1.3.

1.1 PET Overview

PET, or dynamic Positron Emission Tomography (dPET), is a type of nuclear medicine imaging technique which involves the detection of gamma rays emitted by the radioactive tracer, such as fluorodeoxyglucose (FDG), administered to the patient. The tracer acts as a natural body compound and accumulates in the appropriate organ. During its normal

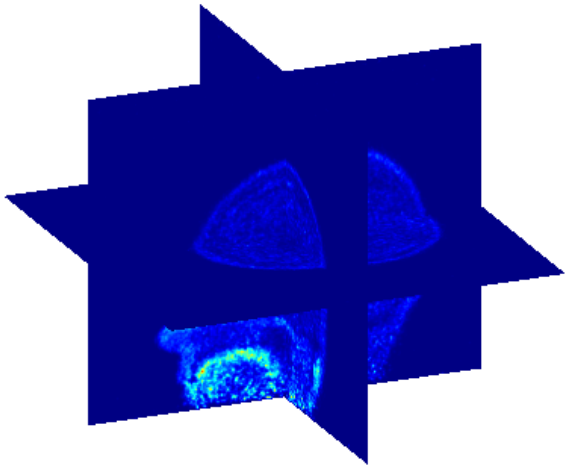


Figure 1.1: An example of a 3D PET image where coloring ranges from blue to red indicating low to high activity.

radioactive decay, positrons are emitted and collide with electrons. The collision annihilates both particles, producing a pair of gamma rays going off in opposite directions at roughly 180 degrees from each other. Two of the detectors within the rings surrounding the patient pick up these two gamma rays, resulting in a known event that can be localized to within the line joining the two detectors, commonly referred to as the Line-of-Response (LOR). After collecting them over a predefined period, these LORs can then be used to create a 3D volume highlighting areas with events detected. Reconstruction can be done via back-projection, or via the more preferred iterative expectation-maximization (EM) algorithm. Figure 1.1 shows an example of a 3D PET image with a coloring ranges from blue (cold) to red (hot), corresponding to low to high activity.

Since the tracer is undergoing radioactive decay, the amount of photons emitted varies with time, usually high at the beginning and rapidly dropping as time progresses. The tracer also interacts with the various tissues within the body differently and accumulates at different speeds. Therefore, PET images are usually taken continuously at a sequence of predefined intervals or time steps such as four 60 seconds volumes, three 120 seconds volumes, eight 300 seconds volumes, and one 600 seconds volume totaling one hour. This creates a series of 3D functional images, with each having different amount of activity at different regions. The dPET analysis then usually involves measuring the time-varying

radioactivity level of the tissue or blood at the target regions of interest (ROI). The time-vary activity level is represented as time activity curves (TACs). Coupled with the knowledge of the tracer behavior, it is possible to use the measured TACs to calculate kinetic parameters describing the relationship between the tracer and tissue physiology. When this relationship is known on a healthy patient, it can be used to identify disease affecting the function of the body. However, this quantification process is sensitive to the patient's head movement, which is unavoidable, given the length of time each scan requires. Therefore, motion correction is needed for accurate diagnoses of therapeutic drugs, or for better understanding neurological disorders such as Parkinsons.

1.2 Motion Corruption

For a healthy subject, the head can drift up to 2mm over the course of the scan, whereas for subjects with Parkinson's disease, drift can go up to 13mm [15]. Most motions during the scan are small, with 1mm-3mm translations or rotations up to 3 degrees, whereas some less common movements, such as using a bedpan, can introduce up to 20mm permanent change in position. In [5], Atkins and Menke found that patients' repositioning due to nurses' interaction could produce up to 7mm difference in position. Similar findings were observed by Ruttimann *et al.* in [82], where up to 20% of the image slides contain rotations in the sagittal plane. With current PET scanners able to achieve a spatial resolution of less than 2.5mm [17] and constantly improving, these movements can prove to be problematic. Many medical staffs have been trying to counter these motions by using head restraints such as a head mark. However, Green *et al.* in [28] demonstrated even with restraints motion cannot be eliminated entirely.

Patient motion can corrupt kinetic modeling (KM) in multiple ways. First, it would shift the LORs used for PET reconstruction. This causes the final PET frames to appear blurred, with activity levels from different regions affecting one another. Some approaches developed for this problem require markers to be attached to patients in order to measure their movements during the scan. This type of tracking has shortcomings, which will be discussed in Chapter 2, and is therefore not widely employed. The second effect of patient motions is the overall shift in brain position inter-frame. Without accounting for this overall movement, the measured TACs will be inaccurate. Techniques developed for this problem involve aligning the frame to some reference, based on some voxel similarity measures using

image registration. However, image registration of functional data is not without its drawbacks. For typical fast decaying tracers (*e.g.* C-11), the rapid uptake early on in the scan requires short sampling time for the initial frames. This means the first few frames suffer from low photon counts and hence are usually very noisy. Furthermore, the spatio-temporal changes in tracer concentration can cause complex intensity pattern changes (not only the intensity values change but their spatial extents change between frames, too). The current trend in medical imaging is to develop tracers that only react to a specific target region, making it impossible to use other areas as a guide for registration. These facts render even state of the art image registration similarity metrics (*e.g.* mutual information-based) incapable of measuring the quality of the alignment.

1.3 Contribution

In this thesis, we purpose a markerless video-based framework for tracking a patient's head movement during PET scan, which eliminates some of the shortcoming with marker based tracker. To accomplish this task, the work makes use of stereo vision, feature extraction, and the Unscented Kalman Filter (UKF) [39]. The algorithm uses features directly available on the patient's face, captured by a set of calibrated stereo video cameras, to determine the motion of the patient's head. Our tracking differs from normal computer vision approaches in that, since the patient's head will be surrounded by the machine during the scan, video sequences will be taken at atypical angles with limited view of the patient's face. A markerless approach also allows us to align brain images of the same patient under a long study, where the patient can be scanned multiple times over the course of several months or years. Features points on the patient are less likely to change position compared to the manually attached markers, and therefore can be used to align scans from different time period.

We also explore a hybrid approach which combines the external tracker information with the pose estimated by registration algorithm, thus bridging the two different methods used for motion correction. We apply an algorithm which favors the tracker information early on when the PET images have a low signal-to-noise ratio (SNR), and gradually switches over to use the registration result when registration is able to produce accurate result due to similarity between these frames and the reference frame. We test this approach by comparing the final retrieved TACs and KM parameters on a set of synthetic dPET brain images under different noise level.

1.4 Chapters Overview

In Chapter 2, an overview of the different motion correction method is presented. Works from both the medical imaging community and the computer vision are covered. The markerless video tracker is purposed in Chapter 3, whereas the hybrid method is described in Chapter 4. Chapter 5 gives an analysis of the performance of both the video tracker and the hybrid algorithm. The thesis is concluded in Chapter 6 and some possible future directions are mentioned.

Chapter 2

Related Work

Two classes of motion estimation and correction methods exist:

1. Extrinsic
2. Intrinsic

"Extrinsic" methods measure motion during the scan, and the correction approaches can be further broken into four groups: Multiple Acquisition Frame (MAF), Deconvolution, LOR rebinning, and Expectation-Maximization (EM). Most current "extrinsic" approaches require markers to be attached to the patient. This is usually inconvenient and uncomfortable for the patients, and time consuming for the staff. It is also difficult to rigidly fasten markers to the head, so they slide or slip, thus producing inaccurate motion estimates. "Intrinsic" methods estimate motion without prior knowledge of head motion during the scan. In this class of methods, motion correction is performed via voxel-based 3D image registration on reconstructed image volumes. As mentioned in section 1.2, these approaches rely on functional images with low spatial resolution and low SNR, and depend on assumptions that may not always hold true, so a change in the observed location of functional activation can not be reliably attributed to either brain motion or change in activation.

Several people have written extensively on this subject. For example, Rahmim [76] wrote a review article, giving an overview of motion correction for PET, with the focus on external methods. Bannister [7] wrote a thesis covering various approaches in motion correction on fMRI. fMRI differs slightly from PET; instead of LOR rebinning or using the EM algorithm, fMRI has its raw image data stored in k-space which are used later for reconstruction, and

motion correction can be done on the k-space data. Lerner also provided a thesis on a similar topic in [41].

2.1 Extrinsic Method

As mentioned in the previous section, extrinsic methods measure motion during the scan, and employ one of the four different techniques for correction. Some of these techniques change the way the image is acquired. Some try to correct the image after reconstruction, and some try to change the way images are reconstructed. The simpler approaches usually make some assumptions or ignore some of the motions and are therefore usually not as accurate as the more complete approaches, but are usually much easier to implement into existing systems and run much faster. For example, in [24], Fulton *et al.* compared the accuracy of MAF and LOR rebinning methods. A physical phantom was used and tracked by the Polaris tracker [43]. The corrected images were compared with the motion-free images by calculating the summed squared error. Results show that LOR rebinning is more accurate than MAF.

2.1.1 Camera System

All extrinsic methods require a head tracking setup. Currently, systems that involve attaching markers to the patient's head proved to the most accurate. These systems either use infrared or charge coupled device (CCD) cameras, and they also range from single camera setup to two cameras setup. Below provides an overview of these systems.

In [68], Picard and Thompson developed a CCD camera-based surveillance system which was capable of monitoring both the patient's head position and movement. This system required three light emitting diodes to be fixed on the patient's face: on the nose, between the eyes, between an eye and an ear. Goldstein *et al.*'s optical motion detector used incandescent lights, but instead of detection via CCD cameras, their work required two electro-optical position sensitive detectors [27]. This system has a reported accuracy of 0.07° and 0.2mm. In [12], Buhler *et al.* used an alternate infrared system called ARTtrack 1 to track a set of retro-reflecting spherical markers for LOR rebinning, whereas in [37], Hu *et al.* used a twin CCD cameras based system, coupled with a less intrusive attachment (three round dots on the forehead).

Similar systems can also be used for small animal PET. It differs from normal PET in

that small animal PET has a higher spatial resolution. Also, small animals usually move much more rapidly than normal human, so pose estimations must be highly accurate [99]. Kyme *et al*. [40] explored the usage of Micron Tracker model S60 in this context. The tracker consisted of two CCD cameras tracking a set of checker-like markers. In addition to attaching the markers to the target animal, the authors also attached a set of reference markers to the scanner. This setup allowed the cameras to be easily recalibrated to the scanner when the cameras needed to be moved. A root mean square (RMS) error of 0.46mm was reported.

Some researchers have explored the possibility of using a single camera. In [58], Muraishi *et al*. [58] designed a new solid marker such that using a single camera is possible. This made the system suitable for a scanner with long narrow patient port space, where a stereo setup might be difficult. This system was further improved in [31].

By far, the most popular tracker is the commercially available system Polaris [43] by Northern Digital Inc. This system uses a pair of infrared cameras tracking either a set of infrared-emitting diodes in active mode, or a set of retroreflective disks or spheres in passive mode. The manufacturer reports an accuracy of 0.35mm RMS.

Others have also researched methods for a markerless approach. In [26], Gao *et al*. were developing a system for PET which used facial features, and this work is continued in [3] and [4]. The head was first located by segmenting the input image and locating skin colored segments. Feature detector was used to locate the two eye corners and the nose (totaling 3 landmarks), and the patient's head pose was calculated by finding the linear and angular relations between two sets of points from two video frames. Unfortunately, detailed results on the tracker's performance were not available. Gao *et al*. tested their feature detectors on a sequence of 12 video frames, claiming they can locate the facial landmarks with accuracy of 1 ± 0.64 pixels. In terms of the motion estimation accuracy, their test was still in the preliminary stage where they were testing on a 3D model of the head showing a promise of preserving one degree precision on rotation angle. Their work differs from our video tracker in three ways. First, they used segmentation to locate the head while our tracker requires a one time user interaction to identify the head region. Second, instead of using all possible features found on the face, their work used three predefined features. Third, their work did not use any temporal coherency and motions were only estimated by comparing with the initial frame.

2.1.2 Multiple Acquisition Frame

Multiple Acquisition Frame [69, 25, 32] is a simple technique for incorporating head tracking into the PET framework. Purposed by Picard and Thompson in 1997, the MAF method works by connecting the tracker to the PET scanner. When the motion tracker detects motion above some predefined threshold, it tells the scanner to begin acquiring a new PET frame. Each new frame triggered is associated with the head pose at the start of each frame and is used later for alignment. Performance of MAF depends on the size of the threshold. Setting a threshold that is too small will trigger too many low statistic frames, whereas setting a threshold that is too high will neglect motions within frames.

2.1.3 Deconvolution

Another direction would be to deconvolve the PET image after reconstruction as a post-processing step [53, 18, 73]. Information from the motion tracker is used to construct the deconvolution filter [53] or to be used in an iterative algorithm [18, 73]. As with any post-processing deconvolution method, without prior knowledge of the original motionless image, this method tends to amplify noise and introduces new artifact into the image. The advantage is that it does not require the detailed specification of the scanner, or alter the reconstruction algorithm [18].

2.1.4 LOR rebinning

The LOR rebinning method [53, 23, 10, 12, 93, 99] is one of the two that tries to alter how the images are reconstructed. It works directly on the list-mode data. In short, list-mode data is a sequence of data containing information about all the events that occurred during the scan. Typically, each event is stored in a 32-bit data word indicating which pair of detectors is involved, although the actual length of the data packet and the format of the word vary from scanner to scanner depend on the model. Every millisecond, a time tag is inserted into the list-mode data, so with both pieces of information combined, we get a full picture of what event occurred at what time. LOR rebinning is then a straight forward correction of these events. The motion information collected from a marker-based tracker, such as Polaris, is used to interpolate which pair of detectors would actually be triggered if there were no motion. The corrected events can then be converted for reconstruction in histogram-mode, or passed directly to the list-mode EM reconstruction algorithm. Additional care must be

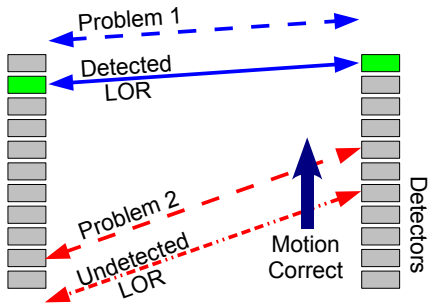


Figure 2.1: A simple 2D illustration of the problem with LOR rebinning. (1) Events that should not have been detected when there is no motion will move outside FOV after motion correction, and (2) events that should have been detected when there is no motion are lost.

taken to ensure the normalization factor associated with the original LOR is used, instead of the new corrected LOR.

Two problems exist with this method: (1) detected events might move outside the scanner field-of-view after motion correction, and (2) events that should have been detected but are lost due to the motion cannot be recovered. These two problems are illustrated in figure 2.1 as a simple 2D example. Most researchers opt to simply discard the events in (1), as these events would not be there if there were no motion. However, neglecting (2) will lead to underestimation of the amount of activity in the affected regions and also produce artifacts. Methods have been purposed, such as [12], which involve scaling the normalization correction factor, but they are generally computationally intense.

2.1.5 EM Algorithm Modifications

The last known method attempts to incorporate the motion data directly into the EM algorithm used for reconstruction [71, 77, 72, 79, 78]. This approach has been applied to either the histogram-mode EM algorithm [77] and the list-mode EM algorithm [71, 77]. For the histogram-mode, this is incorporated using by the probability system matrix, whereas for the list-mode, this is implemented by modeling the motion into the likelihood function. The approaches for the two modes are similar, but list-mode has the advantage that instead of working on discrete sinogram bins, the motion corrected coordinates can be used as a continuous variable, and is therefore more accurate [76].

2.2 Intrinsic Method

In contrast to extrinsic methods, intrinsic methods only use the information available on the reconstructed image without relying on external information. Registration algorithms are used to bring one image (the moving image) into alignment with another image (the fixed image). This is usually done by matching regions from one image to another, such that a specific criterion is minimized or maximized. Techniques have been developed for a wide variety of situations, covering differences in dimension, modality, transformation, and even differences in patient. When we align two different images, we first need to consider the problem of interpolation. This is the case, for example, when we are aligning two images of different modalities, two images of different scales, or when the alignment brings points from the moving image to non-grid positions on the fixed image. As show in [33], different interpolation methods can produce very different results. A long list of literature exists on the topic of registration, and many people have done extensive survey and comparison on the various algorithms [92, 49, 34, 100]. Some algorithms involve a voxel intensity metric which they try to minimize via an optimization technique [66], while others are landmarks based [13, 86], or surface matching, segmentation based [85, 38]. However, aside from a few exceptions [70, 101], majority of the monomodal PET-to-PET registration articles are voxel intensity based [49].

The most popular registration package for PET-to-PET images is Automated Image Registration (AIR) [95, 96]. For rigid registration, AIR provides three possible cost functions to optimize: ratio image uniformity (RIU), least-squared difference image (LS), and scaled least-squared difference image (SLS). AIR minimizes one of these cost functions via Newton-type minimization. As of the current version, AIR has been expanded to also handle intermodality and intersubject registration.

In [54], Minoshima *et al.* purposed using stochastic sign change (SSC) as the metric for registering PET images, with special focus on asymmetrical images caused by lesion. Malandain *et al.* [50] described a potential minimization technique that is less sensitive to local minima, whereas Eberl *et al.* [16] used the sum of the absolute pixel-by-pixel differences (SAD) as the metric. Cross-correlation (CC) can also be used as a registration metric, as in the case by Maintz *et al.* [48]. On the other hand, others have developed methods using Mutual Information as a criterion [47, 52, 81]. Most of these techniques iteratively locate local minima/maxima which might or might not be global. Therefore, approaches

which always return a solution at the global optimum have been purposed [19, 64]. These methods use Fast Fourier Transform (FFT) and differ from cross-correlation in that the global optimum is clearly defined.

2.2.1 Registration of Image Sequence

All of the mentioned registration approaches simply register two images. However, there is a unique feature which is specific to functional modality which is not focused on in the above methods. Functional imaging is usually 4D in nature, creating a sequence of images that needed to be aligned. Registering the sequence means a reference must be chosen to which all other images are aligned. In [35], where Hoh *et al.* compared the performance of registration using the SAD and SSC metric, the authors used a motionless image sequence as the reference. Images from another sequence were then registered to the corresponding images in the motionless sequence. A similar approach was taken by Anderson [2], only instead of using SAD or SSC as the metric, the author calculated the CC at the edge of the brain to decrease computation time.

A different reference was chosen by Lin *et al.* in [42] when they tried to evaluate different metrics. Given an 18 frame FDOPA image sequence, the authors used the last frame, 18, as the reference frame and registered frames 10-17 to it (ignored frames 1-9). They also tested both the unidirectional and bidirectional approaches. Unidirectional means the standard registration where one image is the reference. Bidirectional [94], on the other hand, means doing two registrations, once with one image as reference, and once with the other image as reference. In this case, the average of the two runs is computed to be the final registration result.

In [1], Anderson registered each frame in the sequence with the previous frame. All frames can then be aligned to the first frame by multiplying the transformation matrix. The author noted that it is important to use a registration algorithm with no bias. If an algorithm has a bias of 0.5mm, then after registering 25 frames, a total of 12.5mm error would have been accumulated.

Perruchot *et al.* [67] did a comparison of six different reference frame: (1) a time average volume created from attenuation corrected volumes, (2) a time average volume created without attenuation correction, (3) a MRI image of the same patient, (4) a PET image constructed from the MRI image, (5) the attenuation map of the patient, and (6) a variant of (5). They showed among these 6 choices, the optimal reference frame was (2).

Lastly, a reference image can be chosen by an expert to avoid images with high amount of motion or low contrast, as was done by Pascau *et al.* in [65].

2.3 Computer Vision

Most markerless head pose estimation approaches are developed in the field of computer vision. They are often designed for user interaction purposes instead of motion correction in medical imaging. These methods usually assume some prior model of the object being tracked and have a clear frontal view of the target face. An extensive survey on this subject was written by Murphy-Chutorian and Trivedi in [59]. In general, computer vision approaches can be classified into either feature based or optical-flow based. Optical-flow based methods try to use the observed motion vectors to determine the movement of the modeled object [9, 8, 14]. Feature based methods involve the tracking of specific features on the target object. For example, Azarbayejani *et al.* [6] used the Extended Kalman Filter to track features such as the eyes or the mouth.

Just like in the medical community, some people have tried tracking with a single camera. For example, in [36], Horprasert *et al.* used a single camera to determine head orientation by tracking eye corners and the nose, while Ohayon and Rivlin [63] matched 2D feature points to a sparse 3D model. However, high accuracy is usually difficult when tracking with only one camera, so in [51], Matsumoto and Zelinsky considered using stereo vision. The work was further extended in [61] by Newman *et al.* Their method required the user to manually select up to 32 features on three different orientations of the head. Similar work was conducted by Yang and Zhang [97], requiring a one time model acquisition of the user's head, and manual selection of seven landmark features. Pose estimation was calculated by matching the model with the features. Niese *et al.* [62], on the other hand, used the depth information and an assumption of skin color for tracking, and required a model of the person's face to be constructed via range scan.

Morency *et al.* [56] attempted to remove the requirement of user interaction at startup by using a frontal face detector to locate features. This work results in a sequence of papers [74, 57, 75] aiming at reducing drift in long video sequences. This is accomplished by storing a set of representative frames in predefined orientations, and head pose is calculated from the closest representative frame.

Our markerless video head tracker is developed with techniques similar to those outlined

in this section, such as using face features for tracking [56] or using the Kalman Filter [6]. However, as mentioned, none of the above are developed for medical imaging purposes, and the main difference is that we have an atypical view of the head with limited view of the face. Nonetheless, it is possible to use these head tracking algorithms in place of ours and gain better performance, depending on the algorithm's adaptability to such a unique view.

Chapter 3

Stereo Video Tracker

We developed a method for tracking head pose that eliminates the tracker dependence on attaching markers to the head. In particular, we use a stereo video tracking system, in which left and right (L/R) high resolution video cameras record head movement, and computer vision methods calculate the head's 3D position. For 3D head pose estimation, non-collinear pairs of corresponding head/facial feature points (in L/R images) are identified and tracked throughout the video using feature point detection (with Scale-invariant feature transform (SIFT) [44]) and object tracking (with the Unscented Kalman filter [39]). As the head is mostly surrounded by the scanner gantry, pose estimation will be based on tracking facial features as seen by looking from inferior to posterior (Figure 3.1).

3.1 Camera Calibration

As mentioned, this framework approaches the problem by using video sequences taken from a stereo camera setup, and since the patient's head will be surrounded by the machine during the scan, the video sequences will be taken at atypical angles with limited view of the patient's face. Taking video at such angles also minimizes any inconvenience to the patient or medical staff. The cameras are calibrated using the Bouguet's Camera Calibration Toolbox for Matlab [11]. The process involves taking a sequence of photos of a checkerboard at different positions and orientations (Figure 3.2). These photos are input into the toolbox, and for each photo, the user manually selects the four corners of the checkerboard, and ensures the same corners are selected in both the left and right images. The Matlab code automatically identifies all the checkers' corners, and coupled with the real width and height

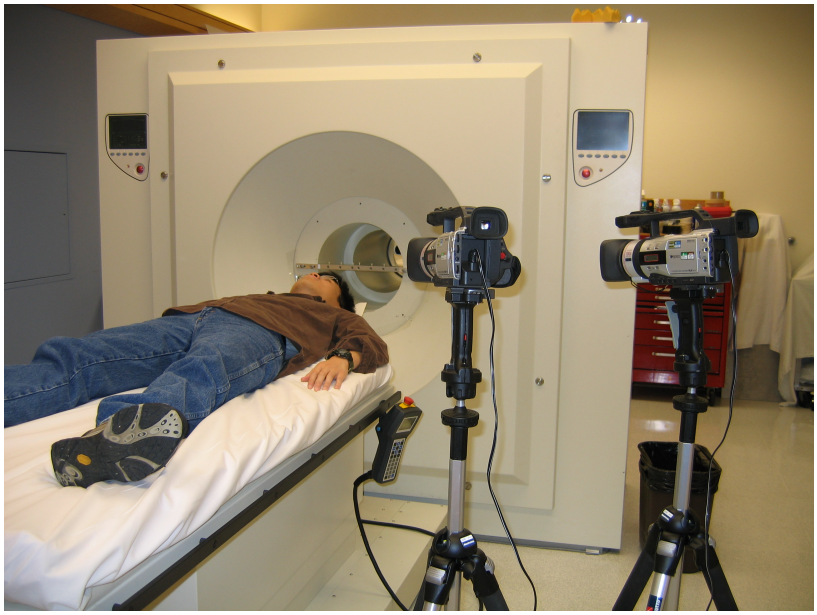


Figure 3.1: Example setup of the stereo system, looking inferior to posterior. The system can also be set further away or be mounted on the ceiling to minimize any inconvenience to the medical staff.

of each checker, the toolbox can find the parameters that characterize the cameras.

The toolbox also provides a functionality to calculate the 3D position relative to one of the cameras by passing two matched points from both views, a common technique called stereo triangulation [20, 45]. This work makes use of the functionality by matching features from the two cameras and locates unique 3D points on the patient's face.

3.2 Feature Matching

Since not all facial features are ideal for motion estimation, the users begin the process by selecting regions on the base L/R images, where feature points will be extracted and matched using SIFT. SIFT will be discussed in more detail in section 3.2.1. Example regions for features extraction might be the nose, the eyes, or the ears. Features points between the L/R images are matched, and only those that satisfy the epipolar constraint are kept. This guarantees the matches will produce points in 3D space via stereo triangulation. Feature points are also extracted from each pair of subsequent L/R video frames, and matched

(a) Left Camera (b) Right Camera

Figure 3.2: The checkerboard images used for camera calibration. The Bouguet's Camera Calibration Toolbox for Matlab [11] is used to automatically calibrate the cameras after manual selecting the four corners of each checkerboard on each image.

with the features from the base images. These matches must satisfy the circular constraint outlined in section 3.2.2.

After the preliminary step, we have a set of points S_t in 3D (via stereo triangulation on the matched feature points) for each pair of L/R video frames at each time step t, with the correspondence to the base 3D points S_0 known. The second step involves finding the actual motion of the head. For this work, we used the exponential map to represent rotation in 3D [46].

The exponential map describes rotation with a 3D vector $\vec{w}$ where $\vec{w}$ is the axis of rotation, and $||\vec{w}||$ is the rotation angle. Given $\vec{w}$, the rotation matrix is given by:

$$R = I + \frac{\widehat{w}}{||\vec{w}||}\sin||\vec{w}|| + \frac{\widehat{w}^2}{||\vec{w}||^2}(1 - \cos||\vec{w}||) \tag{3.1}$$

where

$$\widehat{w} = \begin{bmatrix} 0 & -w_3 & w_2 \\ w_3 & 0 & -w_1 \\ -w_2 & w_1 & 0 \end{bmatrix}. \tag{3.2}$$

The above representation gives us 3 rotation and 3 translation parameters that need to be calculated for each frame. We know that the orientations of the head between adjacent frames are very similar, and we take advantage of this fact by calculating the head orientation using the UKF outlined in section 3.3.

3.2.1 SIFT

The feature detector we use is SIFT. SIFT is a well known features detector that is capable of finding local image features that are invariant to changes in translation, rotation, and scale, and also partially invariant to illumination and affine changes.

SIFT achieves these by applying a difference of Gaussian function at each scale of the image, and selecting keypoints at locations with the maximum or minimum value. Each keypoint is assigned an orientation by computing the peak of the local gradient direction histogram. Relative orientations are also computed for the 16×16 Gaussian weighted region around each keypoint. These 16×16 orientations are divided into 4×4 subregions with each subregion summarized by a 8 orientations histogram, creating the final $4 \times 4 \times 8 = 128$ dimensions descriptor for each keypoint/feature.

Keypoints between two images are matched via nearest neighbor. In a video sequence recorded from setup such as figure 3.1, we can generally find roughly 20 features matched per frame after applying the epipolar constraint.

In this work, we use the open implementation of SIFT available from UCLA [89] for feature extraction and matching. Descriptors are first generated on the left base image and right base image separately, and features that are outside the user defined regions are discarded. The remaining descriptors on the two images are matched against each other to find the correspondence.

3.2.2 Circular Constraint

To improve the accuracy of the correspondences between the base images and another L/R input images, we apply what we termed as the circular constraint to the matched features. Each frame at time t undergoes the same features matching between left input image and right input image to generate the feature sets F_{lt}, F_{rt} that satisfy the epipolar constraint. Features found at each time step are also matched with the base features F_{l0}, F_{r0} respectively. In general, if f_{l0} is a base feature on the left camera that is matched to f_{lt} at time t, then this match must be confirmed by the circular matches $f_{l0} \leftrightarrow f_{lt}, f_{l0} \leftrightarrow f_{r0}, f_{lt} \leftrightarrow f_{rt}$, and $f_{r0} \leftrightarrow f_{rt}$ (see figure 3.3). Features that do not satisfy this constraint are removed from the sets.

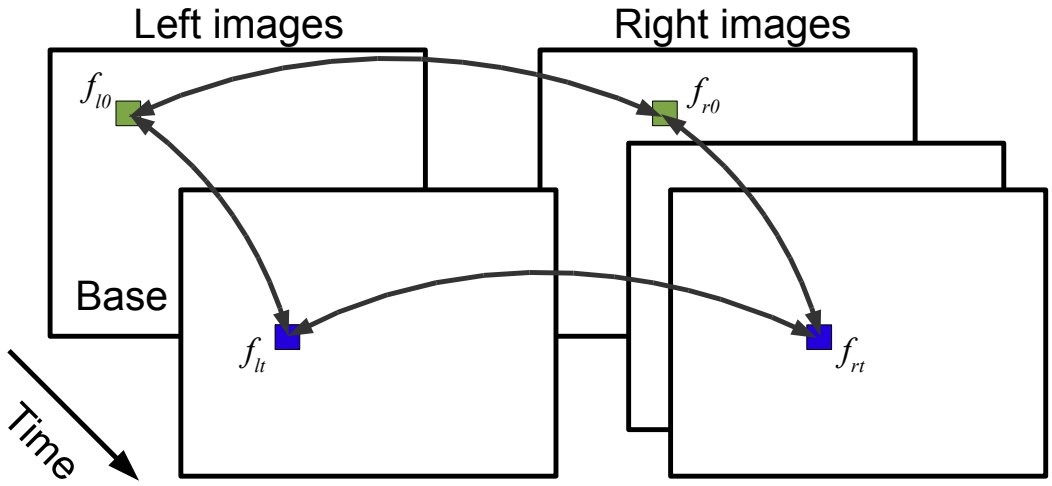
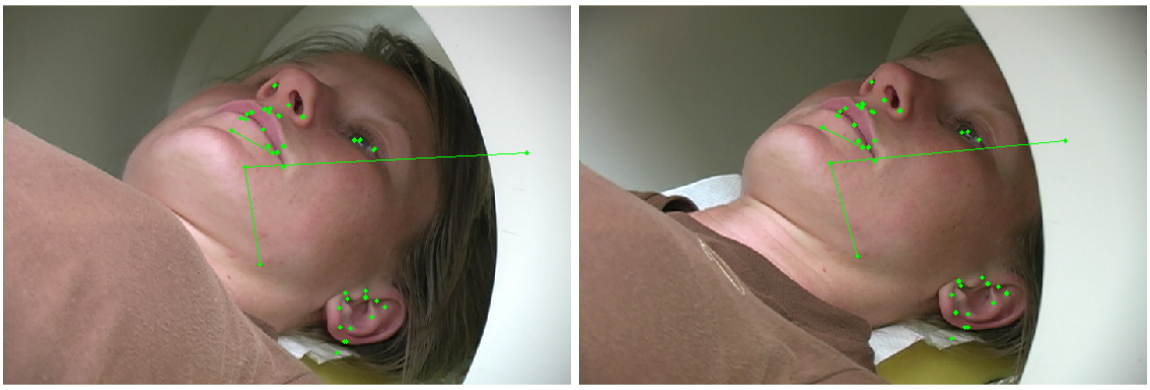


Figure 3.3: The circular constraint requires features to match across the two base images and two input images.

3.2.3 Extra Constraints

Unfortunately, even after applying the circular constraint, it is possible that the same feature on the left and right input frame is matched to the same incorrect feature on the base frames as shown in figure 3.4. This type of match will satisfy both the epipolar constraint and the circular constraint, but creates a high amount of noise that will deteriorate the performance of the UKF. In this work, we applied several additional constraints to combat this type of noise.

The first measure involves using Random Sample Consensus (RANSAC) [21] to prune away points that do not agree with the rest of the group. Three points are chosen at random from S_t, and the least square method [87] is used to find the rigid transformation matching these points to the corresponding points in the base set S_0. The transformation is then applied to the rest of the points in S_t, and the number of points which are within a fixed distance D to their corresponding base points are counted and stored. This stored set is considered as the set of inliers associated with this iteration of RANSAC. Since calculating the transformation at this stage with only 3 points tends to be noisy, we empirically set D to a relatively large tolerance of 5.5mm. The whole process is repeated 70 times, each time



(a) Left Camera (b) Right Camera

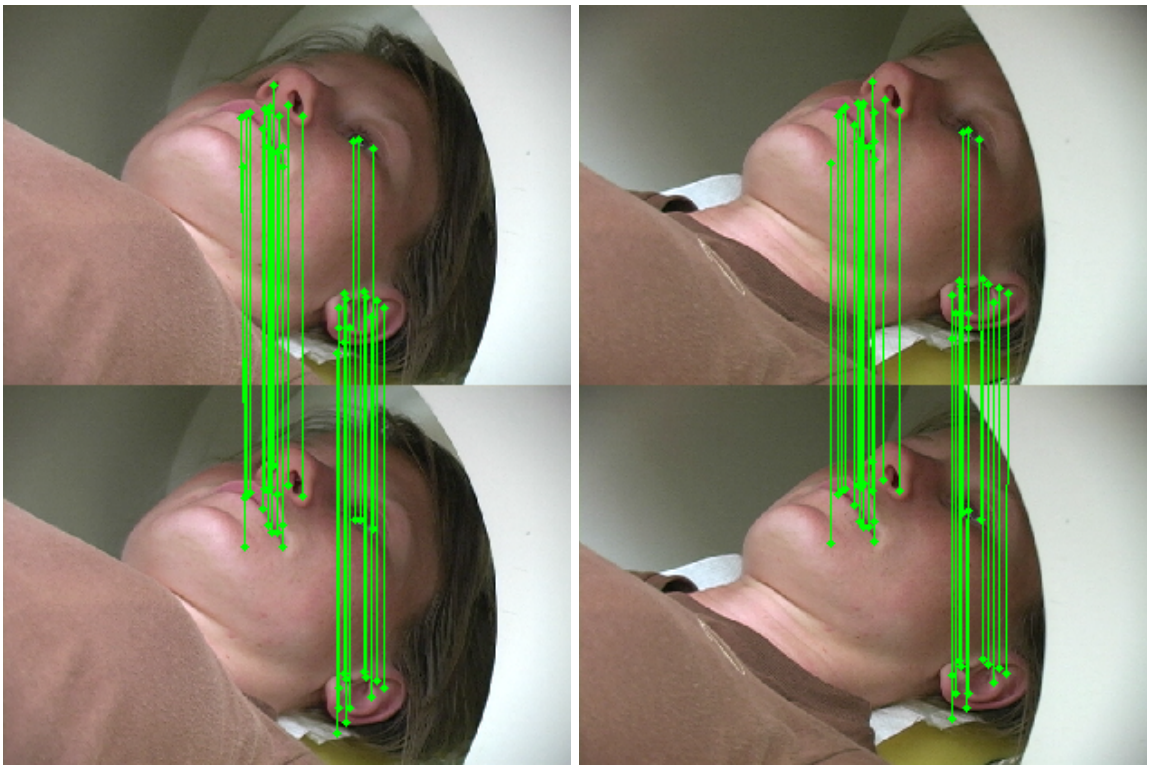
Figure 3.4: Features matched between the base frames and a pair of input frames. Each pair of connected points shows a feature that satisfied the epipolar constraint and the circular constraint. In this case, there are features that get matched to the same incorrect base features (shown with the long lines, indicating large displacement between the two matched features).

the algorithm starts by randomly choosing 3 points. The points in the largest set of inliers are kept as our observation. The number of iterations is chosen by assuming 50% of the points are outliers (an overestimation). Solving for k in the following inequality

$$0.01 > (1 - 0.5^3)^k \tag{3.3}$$

shows that 35 iterations are needed to keep the failure rate below 1%. The failure rate refers to the chance that RANSAC is unable to find the correct set of inliers after completing all iterations. In this work, the number of iterations is doubled to keep the failure rate even lower. In the end, the first constraint will remove most of the incorrect matches (figure 3.5). The algorithmic detail is shown in algorithm 3.1.

The second measure tries to remove points whose distances to other points differ greatly to their correspondence in S_0. Unlike the previous measure, the second step is used to remove points with small amounts of error which are usually difficult to spot just by looking at the SIFT matches, and does not depend on estimating a rough transformation. For each point p in S_t we calculate its distances to the other points. Each distance is compared to the distance calculated between the same pair of points in S_0. When the difference is less than 2mm, the distance is considered a good match. If $^2/_3$ of p's distances to other points



(a) Left Camera (b) Right Camera

Figure 3.5: Final features matched for the left and right input image (bottom row is the base frame) after using RANSAC to remove outliers.

are good matches, then p will be kept in the final calculation. In order to avoid keeping only points in a small tight cluster, which would make finding the correct transformation more difficult, the second constraint is only applied when the range of points τ_t in the final set is over 45mm. The range τ_t is defined as the length of the vector formed by the points' standard deviation in the 3 axes. Algorithm 3.2 shows the details of this second measure.

To differentiate points sets which pass the above constraint and the range test, from points sets which do not pass the range test, an additional match quality variable γ_t is kept for each set:

$$\gamma_t = \begin{cases} 1 & \text{if } \tau_t > 45\text{mm} \\ \frac{l_t}{L_t} min(1, \frac{\tau_t}{45}) & \text{if } \tau_t \leq 45\text{mm or } l_t \leq 3 \end{cases} \tag{3.4}$$

Algorithm 3.1 RANSAC algorithm used to prune bad matches

```
for i = 1 to 70 do
  draw 3 points from S_t at random
  find the transformation from S_t to S_0 with the 3 points
  set M_i as an empty set
  for each point p in S_t do
    apply transformation to p
    calculate distance of the transformed p to its correspondence in S_0
    if distance ≤ 5.5mm then
      add p to M_i
    end if
  end for
end for
find largest set M_k
overwrite S_t with M_k
```

where l_t is the number of points with good relative distance, and $L_t = |S_t|$ is the number of points originally found. The variable is calculated such that the match quality is higher when the number of good matches is high, and when the point range is high. This variable is then used in the Kalman Filter to alter the measurement noise level.

3.3 Predictive Filters

After determining the sets of matched features F_{lt}, F_{rt} and the corresponding 3D points set S_t, the next step is to find the actual rotation and translation. The basic method is to apply Singular Value Decomposition (SVD) to find the best rigid transformation by looking at points in a pair of frames [87]. However, SVD does not look at information from any other frames, and will lose the temporal coherency information where the rotation and translation at time t should be very similar to the transformation at time $t + 1$.

3.3.1 Kalman Filter

One method to take into account the temporal coherency is to use the Kalman Filter [91]. The Kalman Filter is a type of predictive filter that estimates the state $x_t \in \mathbb{R}^n$ of a discrete time controlled process described by the linear equation

$$x_t = Ax_{t-1} + q_{t-1} \tag{3.5}$$

Algorithm 3.2 Relative distance constraint

initialize D_t for storage
for each point p in S_t **do**
  for each point q in $S_t - \{p\}$ **do**
    calculate distance between p and q
    store distance in $D_t(p, q)$
  end for
end for
do similar calculation for D_0 of S_0
set M as an empty set
for each point p in S_t **do**
  compare p's distances in $D_t(p, *)$ with corresponding distance in D_0
  if $^2/_3$ of p's distances in $D_t(p, *)$ within 2mm of those in D_0 **then**
    add p to M
  end if
end for
calculate range τ_t of M
set l_t to $|M|$
set L_t to $|S_t|$
set γ_t as in (3.4)
if $l_t > 3$ and $\tau_t > 45$mm **then**
  overwrite S_t with M
end if

with the observation $y_t \in \mathbb{R}^m$, described by another linear equation

$$y_t = Hx_t + v_t. \tag{3.6}$$

In (3.5) and (3.6), q_{t-1} and v_t are the process model's noise, and the observation's noise respectively. Kalman filtering operates in two steps: time update and measurement update (figure 3.6). In the time update step, the new state $\hat{x_t}$ is estimated using only the past state x_{t-1} with

$$\hat{x_t} = Ax_{t-1} \tag{3.7}$$

$$\hat{P_t} = AP_{t-1}A^T + Q_{t-1} \tag{3.8}$$

where P_{t-1} and $\hat{P_t}$ are the error covariance of x_{t-1} and $\hat{x_t}$ respectively, and Q_{t-1} is the covariance of process noise q_{t-1}. In the measurement update step, the Kalman Filter uses

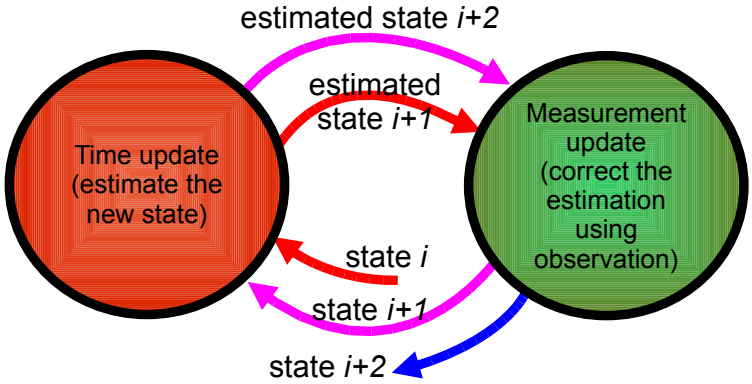


Figure 3.6: The two steps of Kalman Filtering: time update and measurement update. Each step of the Kalman filter takes into account all past states.

the actual observation y_t to correct its estimation as follows

$$x_t = \hat{x}_t + K_t(y_t - H\hat{x}_t) \tag{3.9}$$

$$P_t = (I - K_t H)\hat{P}_t \tag{3.10}$$

where K_t is called the Kalman gain and is calculated using the covariance V_t of observation noise v_t

$$K_t = \hat{P}_t H^T (H\hat{P}_t H^T + V_t)^{-1}. \tag{3.11}$$

For the head pose problem, the state x_t should contain the rotation vector $\vec{w}$ and the translation T. This creates a 6-dimensional state vector $x_t = \{w_1, w_2, w_3, T_x, T_y, T_z\}$. The observation, on the other hand, will include all 3D points location of the matched features, *i.e.*, the set S_t as a single vector. However, the linearity assumption of the Kalman Filter breaks down here. While we use the common assumption of constant position – the location of the head at time t should be near that at time $t-1$, thus leading to the assumption of a linear relationship (identity) for the state transition, the relationship between the state and observation is governed by a non-linear equation $h(x_t)$ defined as

$$h(x_t) = R_t(x_t)S_{0,t} + T_t \tag{3.12}$$

where R_t is calculated by (3.1) using the w_1, w_2, and w_3 components of x_t, and T_t is the translation matrix from T_x, T_y, and T_z. $S_{0,t}$ is a set of points from S_0 that actually have

matches to the points in S_t. This reduced set is necessary since not all features from the base frames will be matched at each time step.

To allow non-linear relationship in the Kalman Filter, the general approach is either to use the Extended Kalman Filter (EKF) or the Unscented Kalman Filter (UKF). Merwe *et al.* proved the UKF is a substantial improvement over other non-linear filters [88]. Moghari and Abolmaesumi [55] showed that the UKF performed much better than the EKF for rigid-body transformation problem. Unlike the EKF, which linearizes the non-linear relationship using Jacobian matrices, and is only accurate to the first-order, the UKF does not require any linearization, and is accurate to at least the second-order. Therefore, for this work, the Unscented Kalman Filter is chosen as the predictive filter.

3.3.2 Unscented Transform

The UKF attacks the non-linear problem by trying to approximate the probability distribution. This is accomplished using the Unscented Transform [39]. Given x_t and P_t, the Unscented Transform first requires the generation of a set of $2n+1$ points termed sigma points, where n is the dimension of the variable x_t. These sigma points have the properties that their mean and covariance matrix are equal to x_t and P_t. One common set of sigma points and the associated weights are generated as follows:

$$\begin{aligned}
\tilde{x}_{k,t} &= \begin{cases} x_t & \text{if } k = 0 \\ x_t + (\sqrt{(n+\lambda)P_t})_k & \text{if } k = 1, \ldots, n \\ x_t - (\sqrt{(n+\lambda)P_t})_k & \text{if } k = n+1, \ldots, 2n \end{cases} \\
w_0^m &= \frac{\lambda}{\lambda + n} \\
w_0^c &= \frac{\lambda}{\lambda + n} + (1 - \alpha^2 + \beta) \\
w_k^m &= w_k^c = \frac{1}{2(\lambda + n)} \qquad k = 1, \ldots, 2n
\end{aligned} \tag{3.13}$$

where $\lambda = \alpha^2(n + \kappa) - n$ and $(\sqrt{(n+\lambda)P_t})_k$ is the kth row or column of the matrix square root calculated with stable methods such as Cholesky decomposition. κ, α, and β are constants, and can be set at $0, 1e-3$, and 2 respectively as suggested in [88]. These sigma points are then propagated with the non-linear function to generate a new set of points:

$$\tilde{y}_{k,t} = h(\tilde{x}_{k,t}) \tag{3.14}$$

The final propagated mean and covariance are calculated as follows:

$$\hat{y}_t = \sum_{k=0}^{2n} w_k^m \tilde{y}_{k,t} \tag{3.15}$$

$$P_{y_t} = \sum_{k=0}^{2n} w_k^c (\tilde{y}_{k,t} - \hat{y}_t)(\tilde{y}_{k,t} - \hat{y}_t)^T. \tag{3.16}$$

3.3.3 Additive Unscented Kalman Filter

The Unscented Kalman Filter generally requires augmenting the state vector with the noise covariance, and generates the sigma points and applies the Unscented Transform on this new state. However, when the process noise and measurement noise are additive, the UKF can be simplified to work directly on the original state vector [30] with the Additive Unscented Kalman Filter.

For the head tracking problem, the only non-linear relationship is the transition from states to observations. In this case, the problem can be further simplified. For the time update step, the algorithm can use the original Kalman Filter for estimating the new state and error covariance. Since the state transition function is the identity, this gives the following time update calculation:

$$\begin{aligned} \hat{x_t} &= x_{t-1} \\ \hat{P}_t &= P_{t-1} + Q_{t-1}. \end{aligned} \tag{3.17}$$

The set of sigma points $\tilde{x}_{k,t}$ is then generated using the estimated state $\hat{x_t}$ and error covariance $\hat{P}_t$, and transformed using the non-linear equation (3.12). This gives a set of transformed points $\tilde{y}_{k,t}$ and their mean $\hat{y}_t$. The transformed covariance is modified to include the measurement noise, with the following calculation from the Additive Unscented Kalman Filter:

$$P_{y_t} = V_t + \sum_{k=0}^{2n} w_k^c (\tilde{y}_{k,t} - \hat{y}_t)(\tilde{y}_{k,t} - \hat{y}_t)^T. \tag{3.18}$$

The measurement update is completed with the following equations:

$$x_t = \hat{x_t} + K_t(y_t - \hat{y}_t) \tag{3.19}$$

$$P_t = \hat{P}_t - K_t P_{y_t} K_t^T \tag{3.20}$$

where $K_t = P_{x_t,y_t} P_{y_t}^{-1}$ is the new Kalman gain used by the UKF and P_{x_t,y_t} is given by the following cross covariance:

$$P_{x_t,y_t} = \sum_{k=0}^{2n} w_k^c (\tilde{x}_{k,t} - \hat{x_t})(\tilde{y}_{k,t} - \hat{y_t})^T. \tag{3.21}$$

3.3.4 Kalman Smoother

Since this framework is not required to run in real-time, we can get better performance by running the UKF both forward and backward, a technique commonly known as the Kalman smoother [90]. Generally, running the UKF backward requires a different state transition function that relates x_t to x_{t-1}. However, since here the state transition function is the identity, the backward UKF can be implemented in the same way as the forward UKF by simply passing observations in reverse order. Merging the solution from the two UKF runs requires the following calculations:

$$(P_t^s)^{-1} = (P_t^f)^{-1} + (P_t^b)^{-1} \tag{3.22}$$

$$x_t^s = P_t^s[(P_t^b)^{-1} x_t^b + (P_t^f)^{-1} x_t^f] \tag{3.23}$$

where $x_t^s, P_t^s, x_t^f, P_t^f, x_t^b, P_t^b$ are the state and error covariance of the Kalman smoother, the forward UKF, and the backward UKF respectively. The initial state and error covariance for the backward UKF are taken to be the final state and doubled error covariance outputted from the forward UKF [22].

3.3.5 Varying Measurement Noise

As mentioned in section 3.2.3, a variable γ_t is kept for each points set indicating the quality of the matches. The variable can be used to change the magnitude of the measurement noise v_t with

$$v_t = v_0 - 0.8 v_0 \gamma_t \tag{3.24}$$

where v_0 is the default level of measurement noise set empirically.

Chapter 4

Hybrid Approach

All of the approaches described in chapter 2 use only one of the two available pieces of information: the value from the motion tracker, or the result from registration. In this work, we consider the possibility of bridging this gap by building a simple hybrid method which takes advantage of both sources (figure 4.1). We begin by considering the case when only one motion is associated with one PET volume, and we develop a hybrid algorithm which uses the external video tracker's information when PET registration results seem to be poor. The possibility arises from the fact that PET uses non-uniform time sampling to capture the uptake of the tracer over time. This results in low SNR in the early frames when tracer uptake is rapid, requiring short sampling time. The low SNR and the changing activity level cause the first frame to differ greatly with the last frame in the PET image sequence. An external tracking system, on the other hand, is not restricted by these limitations, thus making a hybrid approach feasible. The hybrid approach also alleviates the complexity involved with rebinning the LOR and altering the EM reconstruction algorithm, both requiring significant knowledge of the scanner and expertise in the reconstruction framework. This fact, combined with the usage of the markerless video tracker, make it possible to avoid the two major obstacles hindering wide adaptation of the LOR rebinning and EM reconstruction algorithm methods.

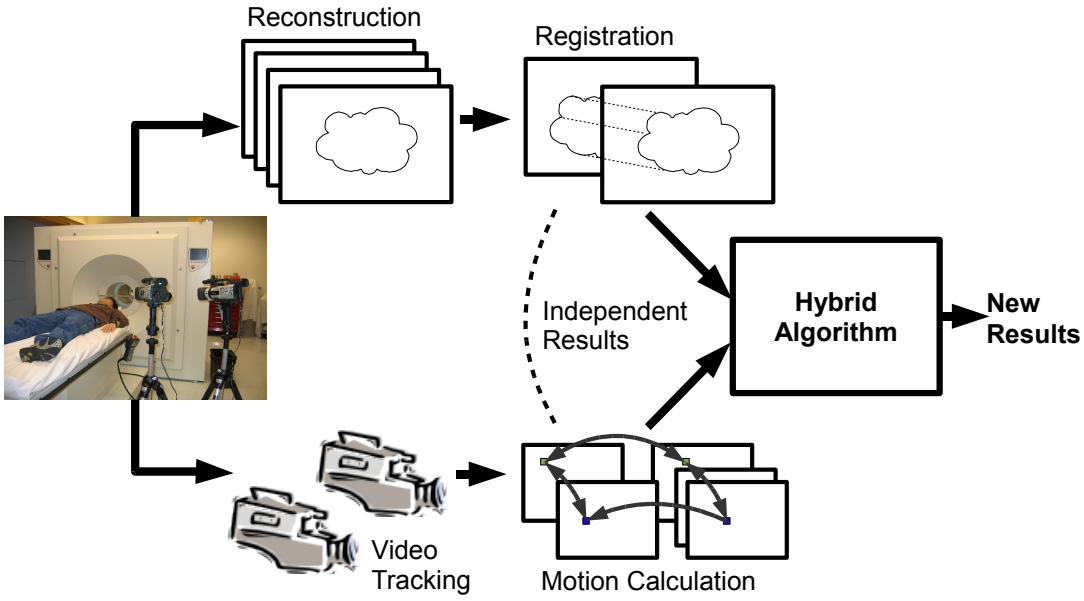


Figure 4.1: A hybrid approach using two independent inputs to generate new result. Since the two sources are independent, each has advantages and disadvantages over the other, and the hybrid method attempts to combine the advantages from both source while avoid the disadvantages.

4.1 Registration Framework

We built our registration framework using the Insight Toolkit (ITK) [98]. ITK is an open-source cross-platform application development framework funded by the US National Library of Medicine. Implemented in C++, ITK allows for the development of segmentation and registration software using preexisting leading-edge algorithms.

Our registration algorithm uses the VersorRigid3DTransformOptimizer from ITK to iteratively find the best transformation. The VersorRigid3DTransformOptimizer is a specially built gradient decent based optimizer using versor composition to update the rotation parameters and standard vector addition to update the translation parameters. In addition, we deployed a multi-resolution framework to improve the speed and accuracy of the algorithm.

4.1.1 Image Similarity Metric

We have compared a few standard metrics available in ITK, such as mean square or normalized correlation, and found mutual information outperforms both metrics. The main theory behind mutual information is the concept of information measurement termed entropy purposed by Claude E. Shannon in 1948 [84]. By definition, the entropy H is defined as

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{4.1}$$

where n is a set of symbols with probabilities given by $p_{1...n}$. In terms of medical imaging, the entropy $H(A)$ of an image A can be defined as

$$H(A) = -\sum_{i=1}^{n} p(a_i) \log p(a_i) \tag{4.2}$$

where n would be the number of bins in the image intensity histogram and $p(a_i)$ would be the value at bin i. From this definition, the entropy $H(A)$ is minimized when image A is a uniform image of a single intensity. In this case, the histogram will have a single sharp peak. When the image contains noise, the intensity values will spread around the peak smoothing the histogram and increase the entropy.

Since registration involves two images, we need to calculate the joint entropy to measure the combined information. The joint entropy of two images A and B is defined as

$$H(A, B) = -\sum_{i} \sum_{j} p(a_i, b_j) \log p(a_i, b_j) \tag{4.3}$$

where in terms of medical imaging, the values for $p(a, b)$ can be taken from the joint histogram of the two images. When A and B are unrelated and independent, we have

$$p(a, b) = p(a)p(b) \tag{4.4}$$

in which case

$$H(A, B) = H(A) + H(B). \tag{4.5}$$

However, when A and B become more similar and less independent, we have

$$H(A, B) < H(A) + H(B). \tag{4.6}$$

This means that the joint entropy $H(A, B)$ is minimized when the two images A and B are exactly the same. In analogy with the entropy example, the joint histogram will become

sharper when the two images are similar, and become smoother when the two images are dissimilar.

It is possible to assess the quality of a registration transformation using the joint entropy as the metric. However, as noted in [34], this will cause the algorithm to favor maximizing the amount of overlap in the air region and might create undesirable results. An alternative is to include both the entropy of the overlapping region (marginal entropy) and the joint entropy, as is done in mutual information $I(A, B)$, defined as

$$I(A, B) = H(A) + H(B) - H(A, B) \tag{4.7}$$

$$= \sum_i \sum_j p(a_i, b_j) \log \frac{p(a_i, b_j)}{p(a_i)p(b_j)}. \tag{4.8}$$

A registration algorithm would aim to maximize $I(A, B)$, which simultaneously maximizes the marginal entropies $H(A)$ and $H(B)$, and minimizes the joint entropy $H(A, B)$. The version of mutual information based registration algorithm available on ITK, which we are using in our registration framework, was developed by Mattes *et al.* [52]. Instead of maximizing $I(A, B)$, the algorithm minimizes $-I(A, B)$.

4.1.2 Reference Volume

Given a sequence of PET images, one must decide which image or images to use as the reference volume. After consulting with medical imaging experts, we decided to use the last frame in the sequence as the reference frame, similar to [42]. This means the last frame will be considered as free of motion, and the registration algorithm will be used to find the relative transformation that transforms any other frame to this last reference frame.

4.2 Hybrid Algorithm

Much of the complexity involved in combining the two independent sources of information comes from deciding which source is providing a more accurate result at any given time. During our testing, we found that the ability to estimate the performance of the video tracker is severely limited. In order to find a suitable variable to predict the performance, we calculated the correlation ρ between our selected variable and video tracker's error on the training set. However, we have tested variables such as the size of rotation and translation measured, the covariance matrix from the Kalman Filter, the number of SIFT matches,

the difference in image intensity between the base video frame and the current frame, the match quality variable γ_t from (3.4), and so forth, and most only correlate with the actual performance of the tracker at roughly $\rho = 0.2$. Attempts at using techniques such as multiple regression or canonical correlation analysis to linearly combine all these variables into one with high correlation resulted in overfitting. Overfitting occurred mostly due to the limited availability of actual videos of patients in the PET scanners with known ground truth motion.

To compensate for the lack of information regarding the performance of the video tracker, the hybrid algorithm uses a simple two step approach using information from the registration algorithm and the known time interval used to collect events for each PET volume during the reconstruction phase. The first step decides whether the registration algorithm is returning a trustworthy result while acknowledging the existence of the unknown video tracker's performance. If the registration algorithm is determined to be good, the hybrid algorithm simply returns the registration result as its final solution. If, on the other hand, the registration result is deemed to be not trustworthy, it is linear combined with the video tracker's result using a time dependent weight.

4.2.1 Final Metric Value

To determine whether the registration is performing poorly, the final mutual information metric value returned by the registration algorithm is used. The assumption is that the final metric value should reflect how similar two images are, and when the two images are highly dissimilar, registration has a higher chance of having poor performance. We found that the final metric value correlates with the actual registration performance at $\rho =\sim 0.8$ on our simulated training dataset, considerably better than what we have found for the video tracker. However, since one must consider the unknown performance of the video tracker, it is not a good idea to simply set a threshold and use the video tracker information whenever the final metric value is higher. Instead, the metric value is used as a probability that determines how much chance we trust the registration result fully. Since the registration algorithm minimizes the negative mutual information, we define $r_m(t)$ to be the final metric value shifted up by 1, such that $r_m(t) = 1$ when the two images (frame or volume t and the reference) are completely independent. The probability $p(t)$ of trusting the registration result is then

$$p(t) = 1 - 0.8r_m(t). \tag{4.9}$$

The factor of 0.8 allows the algorithm to slightly favor the registration result as the registration framework tends to be accurate. A random number between 0 and 1 is generated, such that the registration result is trusted when the number is lower than $p(t)$. This acknowledges the fact that even when the registration result is poor, the video tracker's performance might be worse, and even when the registration result is good, the video tracker's performance might be better. In practice, this random method on average produces much better result than relying on the low correlation variables from the video. Also, since registration usually only performs poorly on the initial PET volumes, on many later frames we have $r_m(t) < 0$. This means in such cases the registration results have probabilities $p(t) = 1$ of being trusted (*i.e.*, we completely trust the registration results).

4.2.2 Time Dependent Weight

If the registration performance is not fully trustworthy after examining (4.9), we move onto the second step of the hybrid algorithm. We do not simply use the video tracker's pose estimate as the final head pose. Instead, the registration result is interpolated with the result we gained from the video tracker to generate a new pose estimate. Since the reference frame is the last frame in the PET images sequence, the registration performance will be worst on the initial frames and most accurate at the second to last frame. This means a time dependent weight $a(t)$, such that it is largest at the beginning and smallest at the end, will be optimal. With such weight, we could combine the two sources of information via linear combination:

$$(1 - a(t))r_f(t) + a(t)v_f(t) \tag{4.10}$$

where $r_f(t)$ is the transformation from registration for frame t, and $v_f(t)$ is the transformation from the video tracker. This way, even if the estimation of the registration performance was wrong from the previous step, it is still possible to recover a portion of its performance with this weighting.

A simple definition for $a(t)$ would be the inverse exponential function. However, such definition would not take into account the different frame timing involved in different type of scan. Instead, we define $a(t)$ as

$$a(t) = max(0, r_m(t) - \lambda(t)) \tag{4.11}$$

where $\lambda(t)$ is the ratio of time used to collect events for frame t, relative to the last frame. For example, if we have a PET sequence where frame 1-4 used 60 seconds, frame 5-7 used 120

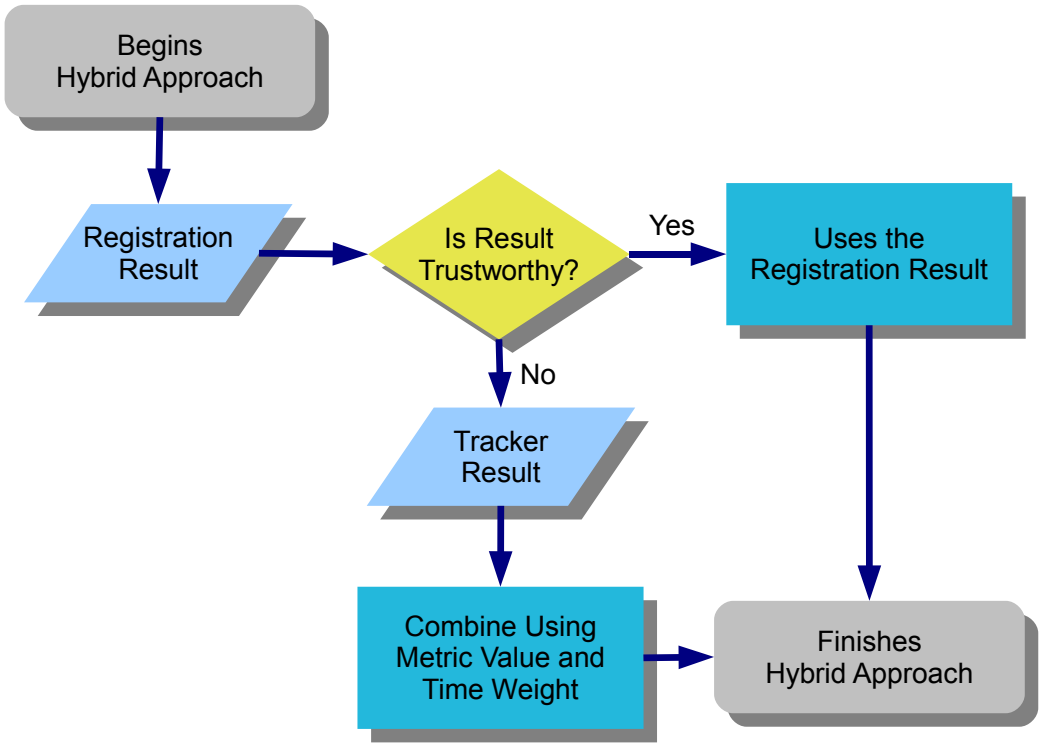


Figure 4.2: Flow chart of the hybrid approach. The time weight and linear interpolation with the video tracker's results are only used when the registration results is not trustworthy.

seconds, frame 8-15 used 300 seconds and frame 16 used 600 seconds, then $\lambda(1 \dots 4) = 0.1$, $\lambda(5 \dots 7) = 0.2$, $\lambda(8 \dots 15) = 0.5$ and $\lambda(16) = 1$. By including the metric value into $a(t)$, the hybrid algorithm also takes into account the actual registration performance instead of blindly using the time steps. Combining this interpolation method with the first step, where registration is trustworthy, gives us a general flow chart summarizing the hybrid approach (figure 4.2).

Chapter 5

Experiments and Results

The purpose of this chapter is to describe the methods used to evaluate the stereo-video head tracker and the hybrid approach.

5.1 Polaris

Since the markerless video head tracker tracks features on a real person's face, a device which can also simultaneously measure the head motion accurately is needed. To achieve this, we used the Polaris tracking system mentioned briefly in section 2.1.1. The system is set up in passive mode to track a set of retro-reflective spheres, and in order to track the head movement, we have four spheres attached to a swimming cap style hat, which the patient must wear. The Polaris system is therefore positioned at the back of the scanner, monitoring the spheres at the top of the patient's head.

The video head tracker, on the other hand, needs to monitor the patient's face, which is not visible from the back, and is therefore positioned at the front of the scanner. At this position, the back of the head and the spheres will not be visible to the video tracker. An additional step is therefore needed to calibrate the coordinates between Polaris and the video tracker. We used a tool which is included with the Polaris system, as shown in figure 5.1. By angling the tool to its side, the spheres will be visible to both Polaris and the video tracker. We took snapshots of the tool at several different positions, and manually identified the location of the spheres in each video frame. The 3D positions of the points were computed via triangulation. After matching these points with the corresponding points from Polaris, the transformation aligning the two coordinates was calculated via least-squares [87].

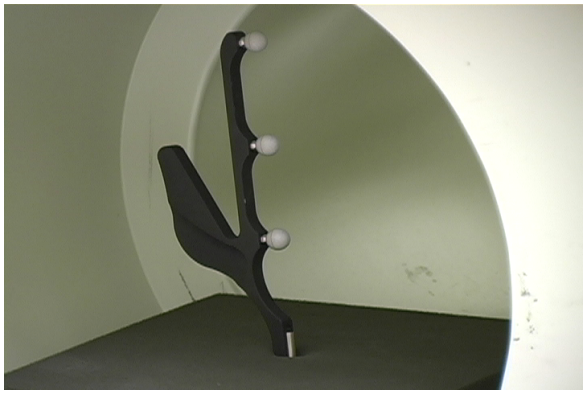
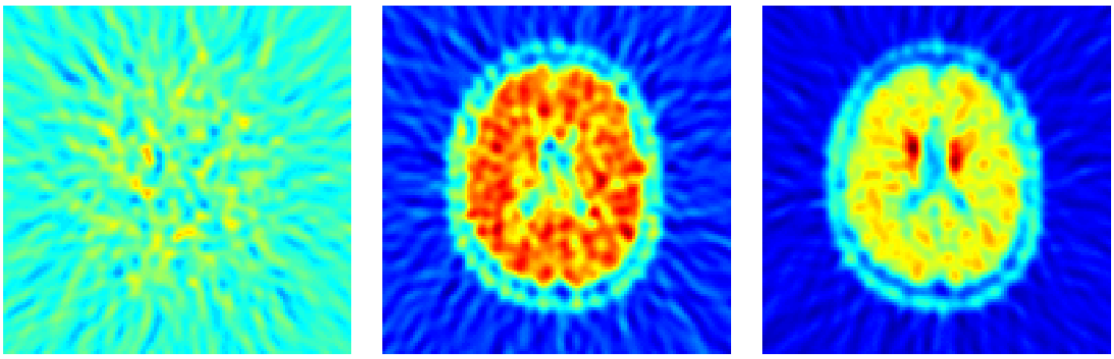


Figure 5.1: The tool used for calibrating between the Polaris system and the stereo-video head tracker. The points are manually located, and their 3D locations are matched with that of Polaris via least-squares.

We collected a video sequence of 3245 frames of a participant lying within a PET scanner, along with the corresponding motion measured by Polaris. The left vs. right vs. Polaris timing is matched by using both cameras to record a short sequence of the clock used for labeling the Polaris motion data. Since the video tracker records at 29.93 frames per second (fps), whereas Polaris records at roughly 20fps, we resampled the motions calculated from the video tracker to match the frame rate of Polaris. The collected sequence is divided into 15 sets, with 5 sets used for training and the remaining 10 used for testing the hybrid algorithm.

5.2 Datasets

Several datasets were used to either set up or test the hybrid approach. These datasets range from synthetic, simulated to real. A synthetic dataset is one which is made to appear like real data but does not necessary follow the steps used in creating the real data. A simulated dataset is created by simulating all the physical properties used in creating the real data, and therefore is usually very realistic. A real dataset is one reconstructed from scanning a real patient with a real tracer intake.



(a) Frame 1 (b) Frame 14 (c) Frame 27

Figure 5.2: 2D slice of frame 1, 14 and 27 of the PET-SORTEO [^{18}F]dopa sequence. The pre-applied noise limited the possibility of testing with this dataset, and is therefore primarily used for training purpose.

5.2.1 PET-SORTEO

The simulated data we used is part of the publicly available PET-SORTEO dataset [80]. We chose the [^{18}F]dopa PET volumes. Each [^{18}F]Dopa sequence consists of 27 volumes $128 \times 128 \times 63$ in size with voxel dimension $2.11168 \times 2.11168 \times 2.425\text{mm}^3$ (figure 5.2). The 27 time steps used are 6×30 seconds, 7×60 seconds, 5×120 seconds, 4×300 seconds, and 5×600 seconds. Applying transformation to these images will introduce voxels that were outside the volumes before the transformation. Since this dataset already includes all major sources of noise (such as from scattered or random events) in the final images, the new voxels are filled with a fixed intensity value which results in regions with no noise. These empty regions will alter the performance of any registration algorithm. Therefore, instead of using this dataset to test the performance of the hybrid algorithm, we used it as a training dataset for measuring the correlation between the registration metric value and registration performance, and also for setting any necessary parameters.

5.2.2 Synthetic Data

For the synthetic data, we generated a [^{18}F]FDG-PET sequence from a segmented MRI image. The original MRI image is $181 \times 217 \times 181$ in dimension with voxel size $1 \times 1 \times 1\text{mm}^3$

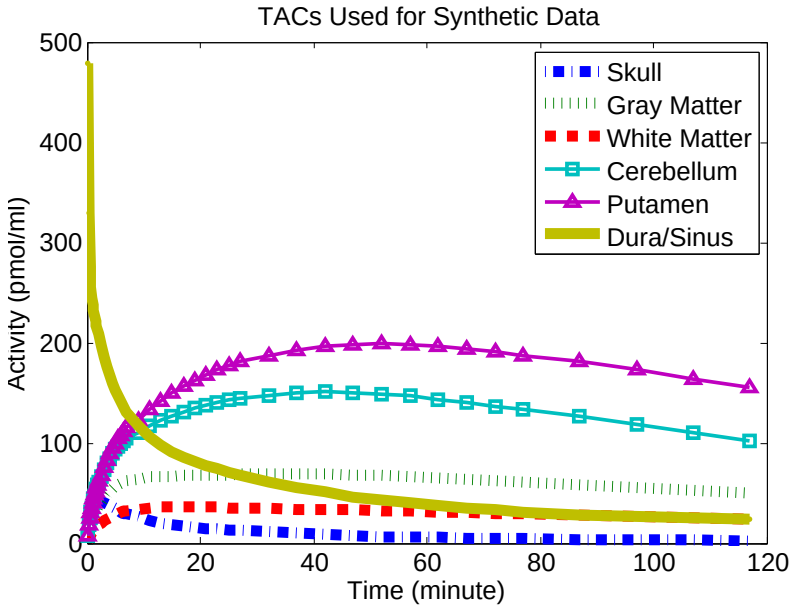


Figure 5.3: The TAC for different functional regions of the synthetic data. They are calculated using COMKAT to simulate a [^{18}F]FDG sequence.

and has 99 different segmented regions. We grouped the segments into 7 functional regions: background, skull, gray matter, white matter, cerebellum, putamen, and dura/sinus. The dynamic of each region is simulated with real kinetic parameters from the dPET clinical literature via the Compartment Model Kinetic Analysis Tool (COMKAT) [60] (figure 5.3). This results in a sequence with 46 frames with time steps: 12×10 seconds, 10×30 seconds, 10×120 seconds, 10×300 seconds, and 4×600 seconds (figure 5.4). To mimic the larger voxel size and the partial volume effect (PVE) of common PET images, each volume is down-sampled to $2 \times 2 \times 2\text{mm}^3$ voxel size and blurred with a Gaussian filter.

Different levels of Gaussian noise can be added to the final images to simulate noise in PET images. The noise applied to each frame t is scaled by a time varying factor $\sigma(t)$ that depends on the ratio of the mean activity and the time step used (similar to [83]). This results in the highest noise level when the mean activity is high and the time step is short, and properly mimics the low SNR in the early PET frames. As the addition of this noise is under our full control, we can apply it after we have corrupted the volume with motion. This makes this dataset suitable for most of our major testing.

This synthetic dataset is simulating one of the worst case scenarios, where dura/sinus reacts to the tracer differently compared to the other areas, as shown by the different TAC in figure 5.3. This results in an image sequence where the first frame and the last frame are highly dissimilar.

5.2.3 Real Data

The real data was provided by UBC-TRIUMF PET Group. It is a [^{11}C]Raclopride 16 frames sequence of a real patient and is relatively motion-free (figure 5.5). This dataset differs from the other sets in that it has a much smaller voxel size of $1.21875 \times 1.21875 \times 1.21875\text{mm}^3$. The dimensions are $256 \times 256 \times 207$ and the time steps are 4×60 seconds, 3×120 seconds, 8×300 seconds, and 1×600 seconds. The volumes also differ in that they all have zero noise in the background, making the contour of the head easily visible.

5.3 Results

The following sections summarize the tests we have done to determine the accuracy of both the video tracker and the hybrid approach.

5.3.1 Tracker Performance

Table 5.1 summarizes the performance of the video tracker, showing the mean, standard deviation, median, minimum and maximum of the absolute error of the six transformation parameters when our video sequence is compared with the output from Polaris. Since translation is applied after rotation, its parameters' error and standard deviation would change depends on which coordinate we defined the transformation. Therefore, we decided to calculate the errors in PET coordinate space, to closely approximate the real performance expected when applied to PET images. One point to note here is that, while Polaris provides a good approximation of the ground truth, it is not the ground truth itself. As mentioned in section 2.1.1, the manufacturer reports an accuracy of 0.35mm RMS, and its real error would be slightly higher in real application. The hat the participant wore might also slip, as is common in this type of marker based tracking.

In terms of actual rotation angle in the x, y and z-axis, the three rotation parameters w_1, w_2 and w_3 together give the errors in degrees shown in table 5.2. From these values,

	w_1	w_2	w_3	T_x (mm)	T_y (mm)	T_z (mm)
Mean	0.01607	0.00989	0.01457	2.12285	1.31440	2.63992
STD	0.01307	0.00831	0.01307	1.93154	1.29183	2.87563
Median	0.01356	0.00772	0.01018	1.74218	0.95034	1.77787
Min	0.00000	0.00000	0.00001	0.00747	0.00085	0.00175
Max	0.07912	0.04557	0.07158	28.26456	10.80654	24.95331

Table 5.1: The absolute error of video tracker performance compared to Polaris.

	x-axis (degree)	y-axis (degree)	z-axis (degree)
Mean	0.84075	0.56190	0.91909
STD	0.75639	0.47229	0.74823
Median	0.59180	0.44189	0.44189
Min	0.00003	0.00013	0.00055
Max	4.03639	2.63193	4.51491

Table 5.2: The rotation errors expressed in degrees.

the performance of our video tracker is acceptable. However, there are some video frames where SIFT had trouble finding matching features, resulting in larger than average errors shown in the maximum row.

5.3.2 Hybrid Approach Performance

We examined the accuracy of our hybrid approach by looking at the Target Registration Error (TRE) and at its ability to retrieve the original TACs relative to using pure registration. TRE is simply the distance between corresponding points of the motionless volume and the motion corrected volume [34]. We generated our synthetic dPET data under 5 trials of 8 noise levels (0,1,3,5,6,7,8,10)$\sigma(t)$. Noise level 0's volumes are already shown in figure 5.4, and the rest are shown in figure 5.6. For each noise level, we transformed the volumes using motions taken from the 10 test sets from section 5.1. This means each trial of each noise level was tested under 10 different sequences of motions, giving a total of 50 sets of volumes for each noise level. At any given time, only one motion is associated with one particular PET image, and we used the motionless version of the last PET frame as the reference image for our registration algorithm.

We computed the mean TRE on voxels in the dura/sinus region in selected volumes, and average over all tests for each noise level. The dura/sinus region is a very thin layer of membrane and blood vessels in the outer region between the skull and the brain. Figure 5.7 shows the results calculated over the first 3 frames, and over all frames. As we expected, the video tracker's performance was not affected by noise in PET images. This test also showed that, by using the final metric value as a factor for determining the accuracy of the registration framework, we can accurately determine at which point registration is becoming less trustworthy and start preferring the video tracker's results. This is true both in terms of the low SNR in the early frames (figure 5.7a), and in terms of overall noise level in all images (figure 5.7b).

We computed the TACs for different functional regions of the brain after using the hybrid method to correct the motion corrupted images, and similarly for pure registration and pure video tracker's results. For each TAC, we calculated the difference between the corrected TAC and the ground truth motionless TAC using

$$|| \sum_t (h_t - g_t)^2 ||^{1/2} \tag{5.1}$$

where t is the frame/volume number, h_t is the activity level or image intensity calculated at frame t after correction, and g_t is the activity level measured from the ground truth motionless PET sequence. In other word, we treated each curve as a point in a multidimensional space, with the number of dimensions equals to the number of PET volumes. We calculated the difference between two TACs as the distance in this multidimensional space, with each dimension in the unit of pmol/ml.

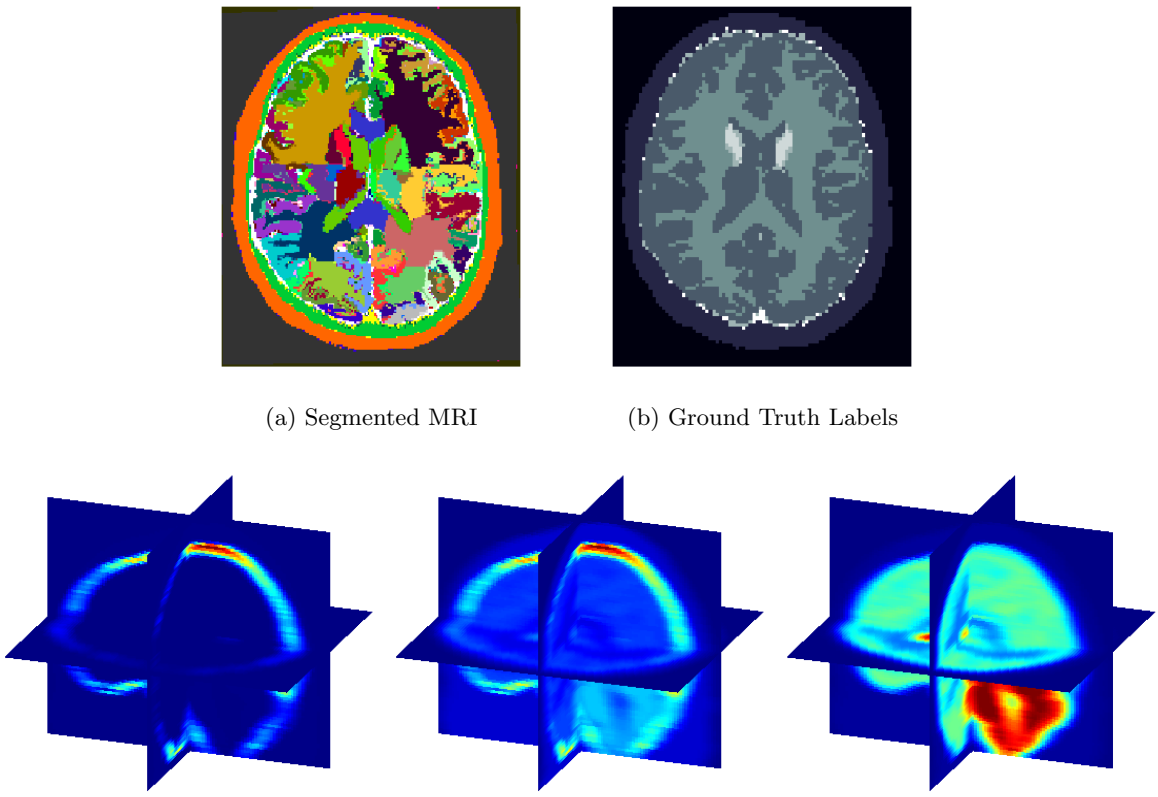
Figure 5.8 shows the advantage of our hybrid method for retrieving the TACs under different ROI and noise level. For most of the regions within the brain, the hybrid method performs better than pure registration. For regions with high activity and near the outside surface of the brain (such as the thin dura/sinus region), the performance of the hybrid method is not as good. One possible explanation is that since the dura/sinus is a very thin layer, slight errors in translation cause the TAC to include the neighboring low activity regions (*e.g.* background). Nonetheless, when the noise level is high, the hybrid method is able to outperform registration in all regions.

One interesting thing to note here is that, even by using a simple technique such as linearly combining two sources of information, we are not restricted to getting results that is only as good as either of the two originals. One example is the gray matter region, where

the hybrid approach is better than both pure registration and pure video tracker in all noise levels.

We also compared the recovered Kinetic Model's parameters by calculating the FDG glucose metabolic rate $K = K_1k_3/(k_2+k_3)$ [29] where K_1, k_2 and k_3 are recovered parameters returned by COMKAT. These parameters describe the relationship between the tracer FDG and the tissue in the body. For example, K_1 is the transport rate from blood to extra-vascular space. Given the TAC of a region, COMKAT uses the known characteristic of FDG (*e.g.* speed of decay) to solve for Ks. Figure 5.9 summarizes these results, and this figure is similar to figure 5.8 except there are cases where the hybrid method has the worst performance, even though the TAC is better. This is likely to be because, in order to find the different parameters, COMKAT begins by curve fitting to the retrieved TAC. Since registration looks at the intensity value within an image itself, its results tend to produce smoother TAC. The hybrid method on the other hand combines information from two independent sources and results in curve that is rougher. A possible solution for this is to smooth the TAC returned by the hybrid method.

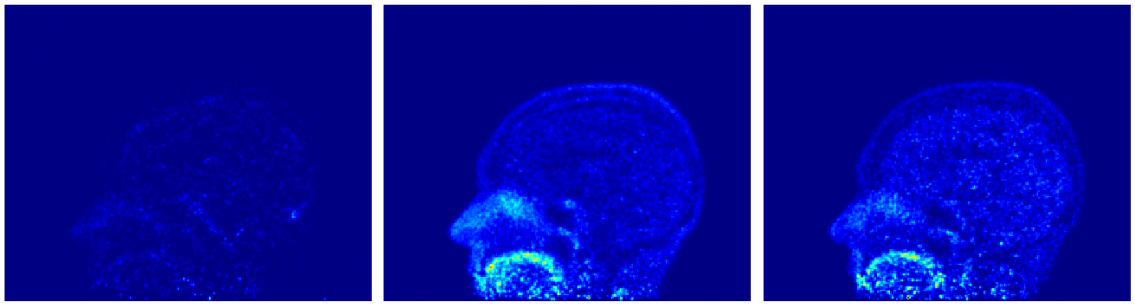
On the real dataset, since the ground truth labeling is not available, we tried to calculate the TRE on a set of evenly spaced points distributed all over the volume. However, since this dataset has zero noise in the background, registration is able to outperform our video tracker and hybrid algorithm even on the first frame. The reason is that with zero background noise, the contour of the head is easily visible even on the first frame with relatively low activity. The location of unique areas such as the nose are clearly identified, and most registration algorithms can correctly align the images by simply aligning the zero valued background, regardless of how much the activity level changes within the brain. In this type of images, a video tracker which matches the quality of marker-based tracker such as Polaris would be needed.



(a) Segmented MRI (b) Ground Truth Labels

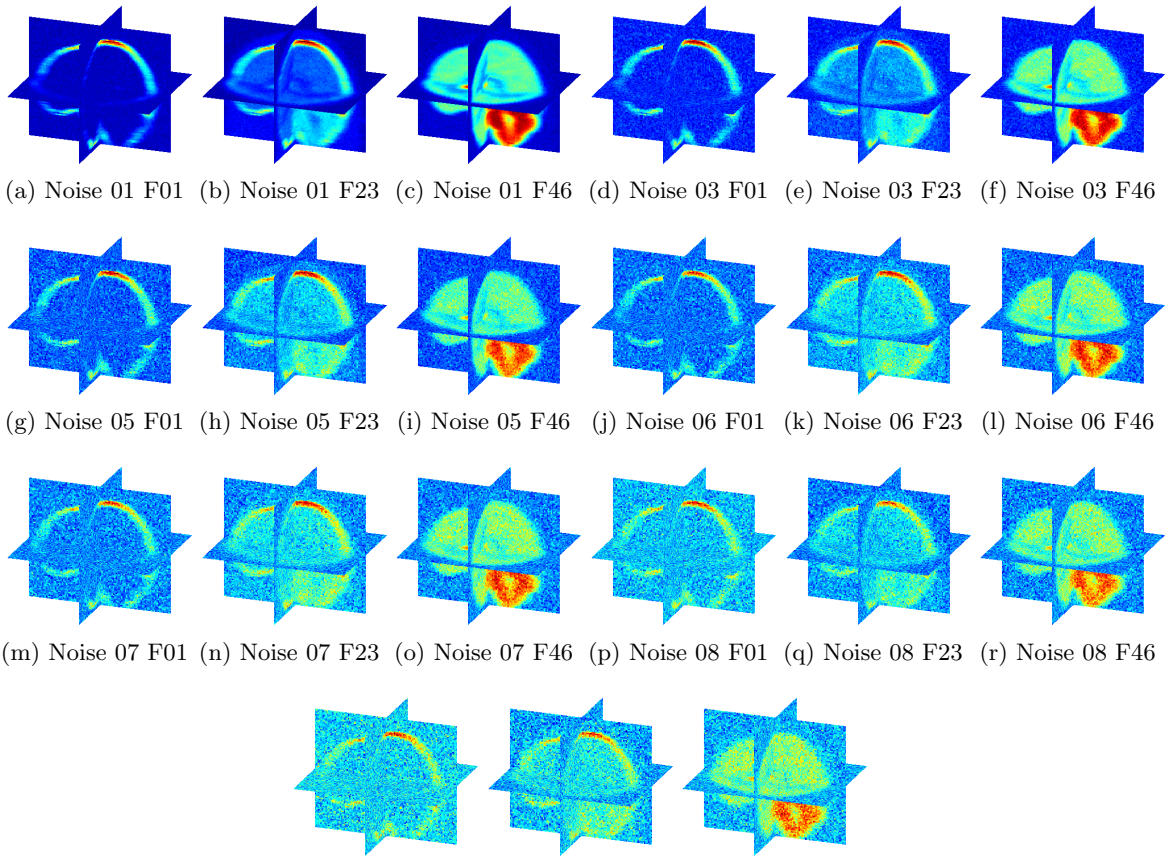
(c) Frame 01 (d) Frame 23 (e) Frame 46

Figure 5.4: (a) 2D slice of the segmented MRI image, (b) the 7 functional regions' ground truth, and (c)-(e) the generated noise-free [^{18}F]FDG-PET volumes. The colormap is scaled to match the minimum and maximum of each image individually to improve visibility.



(a) Frame 1 (b) Frame 8 (c) Frame 16

Figure 5.5: 2D slice of frame 1, 8 and 16 of the [^{11}C]Raclopride sequence. These were taken with a real PET scanner with real patient.



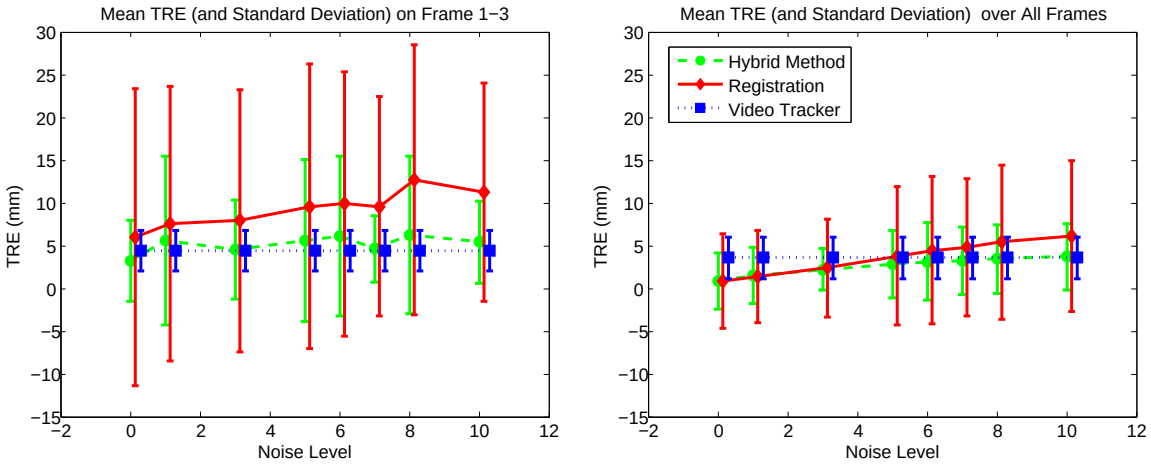
(a) Noise 01 F01 (b) Noise 01 F23 (c) Noise 01 F46 (d) Noise 03 F01 (e) Noise 03 F23 (f) Noise 03 F46

(g) Noise 05 F01 (h) Noise 05 F23 (i) Noise 05 F46 (j) Noise 06 F01 (k) Noise 06 F23 (l) Noise 06 F46

(m) Noise 07 F01 (n) Noise 07 F23 (o) Noise 07 F46 (p) Noise 08 F01 (q) Noise 08 F23 (r) Noise 08 F46

(s) Noise 10 F01 (t) Noise 10 F23 (u) Noise 10 F46

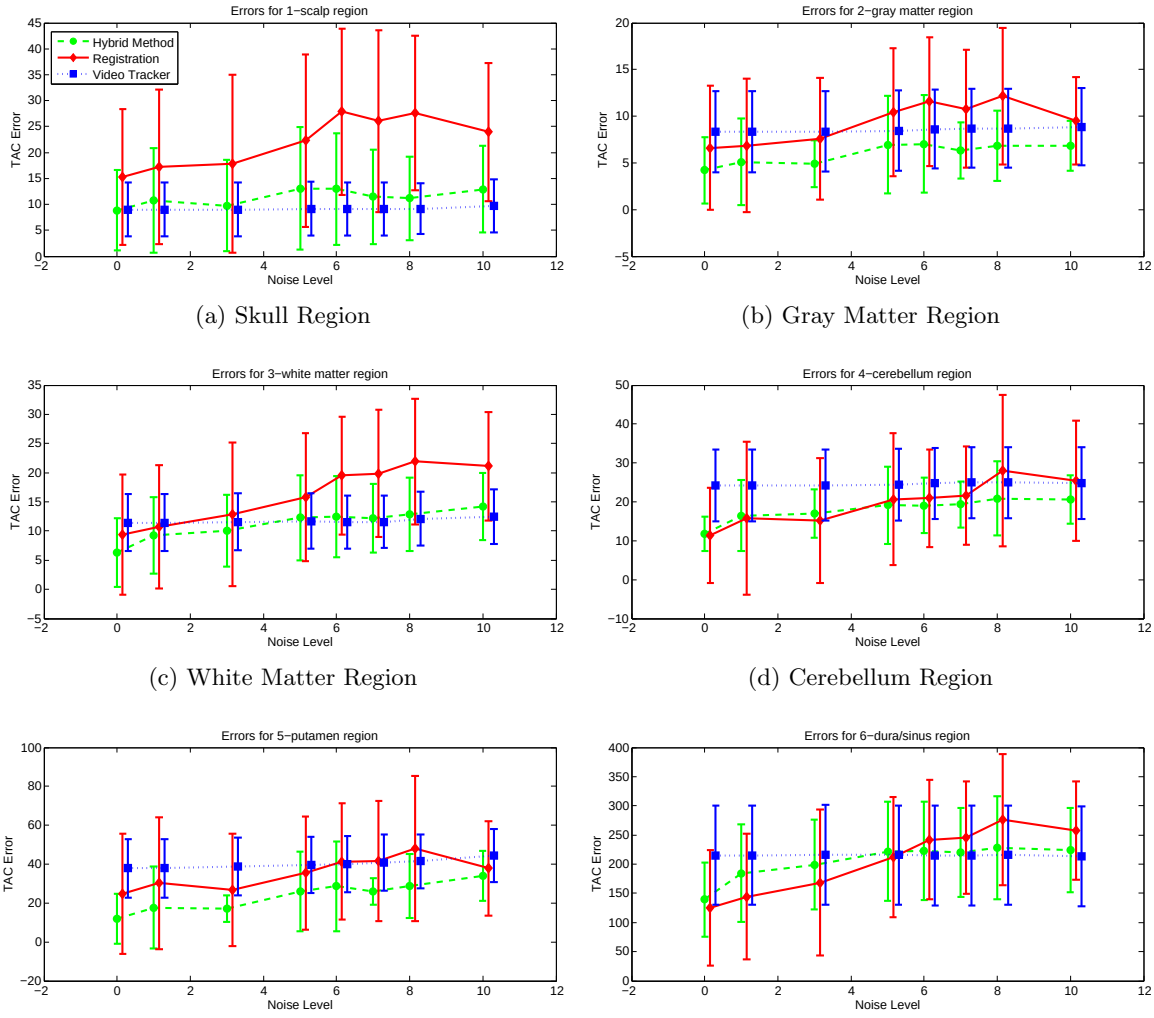
Figure 5.6: The motionless version of volume 1, 23, and 46 of noise level 1, 3, 5, 6, 7, 8, 10. Each set of volumes for each noise level is tested under 10 different motion sequences, and the test was repeated 5 times, each time with the noise regenerated. The colormap is scaled to each individual image to improve visibility.



(a) Frame 1-3 Result

(b) All Frames Result

Figure 5.7: The mean TRE (and standard deviation) over dura/sinus' voxels under different noise levels. The hybrid method is able to handle the low SNR in the early frames, and the overall changes in noise level.



(a) Skull Region

(b) Gray Matter Region

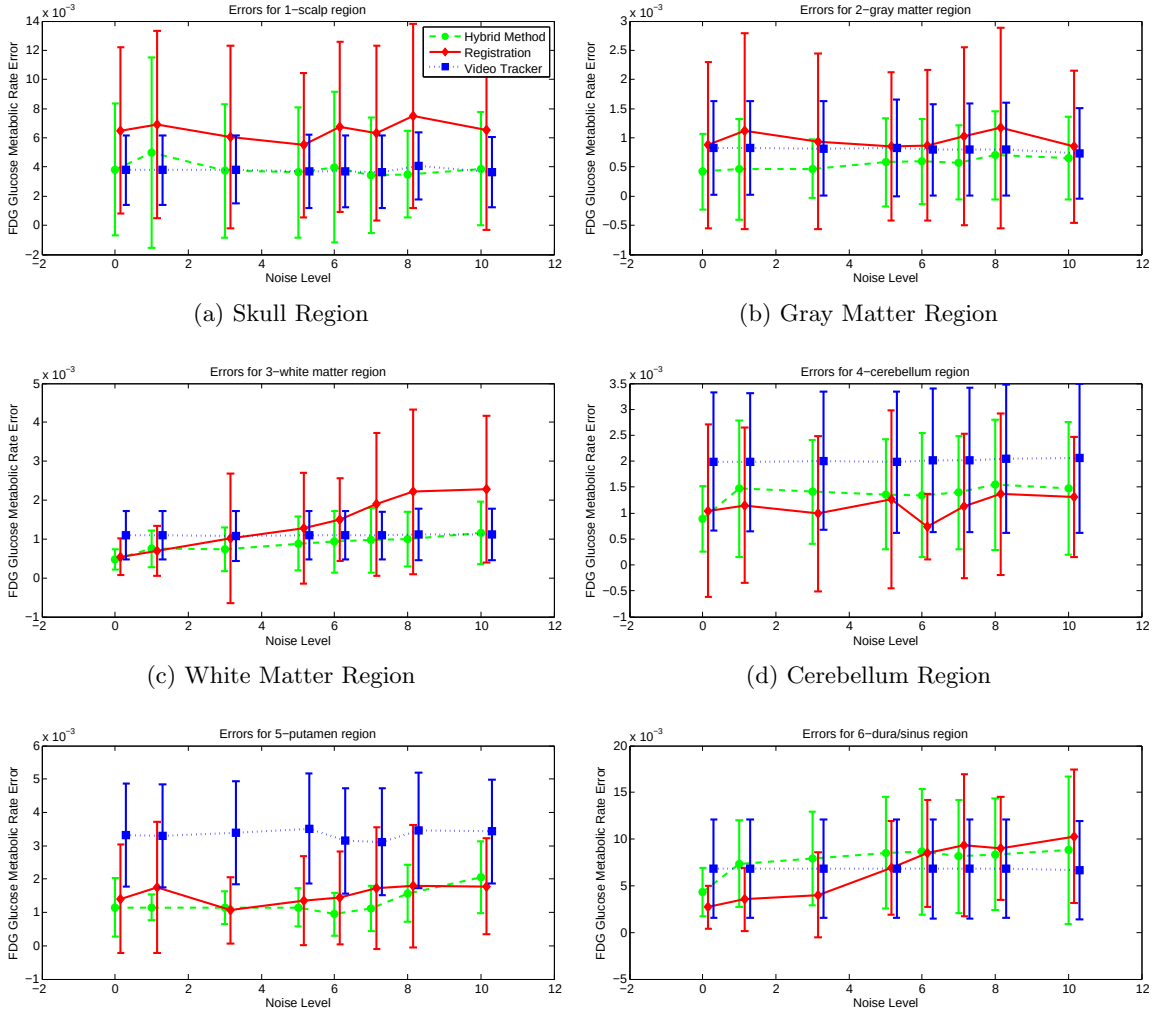
(c) White Matter Region

(d) Cerebellum Region

(e) Putamen Region

(f) Dura/Sinus Region

Figure 5.8: Comparison of the mean (and standard deviation) measured TAC error over different noise levels in different region, by treating each TAC as a point in a multidimensional space. The hybrid method is able to find a better TAC than registration on most regions, especially on the higher noise level. 50 sets of 46 volumes were tested in each noise level. The dura/sinus region is a very thin layer of membrane and blood vessels between the brain and the skull. Its high activity at the beginning of the scan, and having a very low activity neighbor, caused the region to be more susceptible to small errors in translation.



(a) Skull Region

(b) Gray Matter Region

(c) White Matter Region

(d) Cerebellum Region

(e) Putamen Region

(f) Dura/Sinus Region

Figure 5.9: Comparison of errors in the calculated FDG glucose metabolic rate. The hybrid method is performing reasonably well in most cases, although the roughness of the TAC calculated after corrected by hybrid method might have increased the error in K.

Chapter 6

Conclusions

The availability of dPET allows physicians to assess brain function, usually by measuring time-varying tissue or blood radioactivity level at target ROI. Due to the extended period a usual scan requires, patient motion can severely corrupt the final results. PET images become blurred with activities from different regions overlapping one another. The usual approaches could involve attaching markers on patients and tracking their movement, restricting the patient's movement via a specially constructed mask, or image registration. However, all of these methods have shortcomings, such as patient discomfort, or inability to perform motion correction when the SNR is low.

In this work, we have developed a markerless head pose tracking system for estimating head motion of subjects while undergoing a functional medical imaging scan. The work combines the usage of stereo vision with the SIFT features detector for tracking. Features are first detected in each stereo image pair and their 3D locations extracted after satisfying some predefined constraints. These 3D points are then passed as measurement to the UKF, which returns the final transformation with respect to a set of base frames. The primary focus here was to introduce the markerless head tracking approach rather than developing a state of the art face tracker. Algorithms which produce even higher accuracy do exist in computer vision, and are projected to continue to improve. An advantage of using a markerless tracking algorithm, besides not requiring markers, is that it allows alignment of brain images of the same patient under a long study, where the patient can be scanned multiple times over the course of several months or years. Feature positions are less likely to change over time, compared to the manual attachment of markers.

We also showed a possibility of bridging external tracker based PET motion correction

with frame-to-frame based registration. We approached this by using a hybrid framework which uses the video tracker's information when the registration performance is poor, such as the beginning frames in a PET sequence. These early frames have low SNR due to the short sampling time required to capture the rapid uptake of tracer activity. In contrast, an external tracker is not affected by these noises, and can therefore be used to improve the registration result. The hybrid method works by assigning a probability that determines whether the registration result is accurate. This probability is based upon the final value of the metric being minimized by the registration framework. When registration is determined to be inaccurate, its result is interpolated with the video tracker's result, using a time dependent weight.

Experimentation on a synthetic [^{18}F]FDG-PET sequence generated from a segmented MRI image shows that video tracker's results can be used to improve upon registration, especially in the high noise cases. When the image SNR is low, the algorithm is able to detect the lowered performance of the registration framework, and use the time dependent weight to include the video data. When the image noise is high, results show that the hybrid method has lower TRE and TAC errors. On most PET data, roughly 20% of the volumes suffer from the low SNR problem, and the hybrid method is therefore aimed to improve alignment on these images. Analysis on the FDG glucose metabolic rate error shows mixed results, possibly due to the limitation of the curve fitting needed for this calculation. A possible solution would be to smooth the TAC prior to calculate the Kinetic parameters. On the other hand, experiments on high resolution PET images with no background noise show pure registration will suffice in this type of images. This does not, however, remove the advantage of an even higher performance video tracker. When video tracking matches the quality of registration algorithms, the hybrid method will still be useful given the fact that resolution of PET scanners will always be increasing to the point where no human can perceive a difference with higher resolution.

In developing the hybrid framework, we have assumed that motions only occur from frame-to-frame, and there is no motion corruption within a single frame. To correct motions within a frame, techniques such as deconvolution, LOR-rebinning, or changes to the EM reconstruction algorithm are needed. Some of these techniques are already proven to produce the correct image, but are not widely adopted due to the complexity involved. The markerless tracker can alleviate some of these difficulties, but a more straightforward approach which does not require changing the reconstruction step might still be desirable.

One possibility is to build upon our hybrid framework. Instead of combining the two sources after registration, the motion information could be included in the registration algorithm itself. This could be accomplished by, for example, changing the minimization algorithm such that it simultaneously minimizes the image intensity metric, and the difference between the two sequences of motions affecting the fixed and moving image. Depending on how the difference is calculated, this might favor matching the mean or the mode of the transformations. Alternatively, one can develop a new deconvolution algorithm, for which recent publications are showing signs of renewed interest.

Bibliography

[1] Jesper L. R. Andersson. How to obtain high-accuracy image registration: application to movement correction of dynamic positron emission tomography data. *European Journal of Nuclear Medicine and Molecular Imaging*, 25(6):575–586, 1998.

[2] J.L.R. Andersson. A rapid and accurate method to realign pet scans utilizing image edge information. *J Nucl Med*, 36:657669, 1995.

[3] Sergey Anishchenko, Vladislav Osinov, Dmitry Shaposhnikov, Lubov Podlachikova, Richard Comley, and Xiaohong W. Gao. Toward a robust system to monitor head motions during pet based on facial landmark detection: A new approach. In *CBMS '08: Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems*, pages 50–52, Washington, DC, USA, 2008. IEEE Computer Society.

[4] S. Anishenko, V. Osinov, D. Shaposhnikov, L. Podladchikova, R. Comley, K. Sukholentsev, and X.W. Gao. A motion correction system for brain tomography based on biologically motivated models. In *7th IEEE International Conference on Cybernetic Intelligent Systems*, pages 1–5, Sept. 2008.

[5] M.S. Atkins and M. Menke. Effects of head movements measured during pet scans. *IEEE Nuclear Science Symposium and Medical Imaging Conference Record*, 3:1776–1780, 1995.

[6] A. Azarbayejani, B. Horowitz, and A. Pentland. Recursive estimation of structure and motion using relative orientation constraints. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 294–299, 1993.

[7] Peter R. Bannister. *Motion Correction for Functional Magnetic Resonance Images*. PhD thesis, University of Oxford, 2004.

[8] S. Basu, I. Essa, and A. Pentland. Motion regularization for model-based head tracking. In *Proceedings of the International Conference on Pattern Recognition*, page 611, Washington, DC, USA, 1996. IEEE Computer Society.

[9] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proceedings of the Fifth International Conference on Computer Vision*, page 374, Washington, DC, USA, 1995. IEEE Computer Society.

[10] P.M. Bloomfield, T.J. Spinks, J. Reed, L. Schnorr, A.M. Westrip, L. Livieratos, R. Fulton, and T. Jones. The design and implementation of a motion correction scheme for neurological pet. *Physics in Medicine and Biology*, 48:959–978(20), 2003.

[11] Jean-Yves Bouguet. Camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/.

[12] P. Buhler, U. Just, E. Will, J. Kotzerke, and J. van den Hoff. An accurate method for correction of head movement in pet. *IEEE Transactions on Medical Imaging*, 23(9):1176–1185, 2004.

[13] G. E. Christensen, A. A. Kane, J. L. Marsh, and M. W. Vannier. Synthesis of an individualized cranial atlas with dysmorphic shape. In *Proceedings of the 1996 Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA '96)*, page 309, Washington, DC, USA, 1996. IEEE Computer Society.

[14] Douglas Decarlo and Dimitris Metaxas. Optical flow constraints on deformable models with applications to face tracking. *International Journal of Computer Vision*, 38(2):99–127, 2000.

[15] K. Dinelle, S. Blinder, Ju-Chieh Cheng, S. Lidstone, K. Buckley, T.J. Ruth, and V. Sossi. Investigation of subject motion encountered during a typical positron emission tomography scan. *IEEE Nuclear Science Symposium Conference Record*, 6:3283–3287, 2006.

[16] S Eberl, I. Kanno, R.R. Fulton, A. Ryan, B.F. Hutton, and M.J. Fulham. Automated interstudy image registration technique for spect and pet. *J Nucl Med*, 37:137145, 1996.

[17] L. Eriksson, K. Wienhard, M. Eriksson, M.E. Casey, C. Knoess, T. Bruckbauer, J. Hamill, M. Schmand, T. Gremillion, M. Lenox, M. Conti, B. Bendriem, W.D. Heiss, and R. Nutt. The ecat hrrt: Nema nec evaluation of the hrrt system, the new high-resolution research tomograph. *IEEE Transactions on Nuclear Science*, 49(5):2085–2088, 2002.

[18] T. L. Faber, N. Raghunath, D. Tudorascu, and J. R. Votaw. Motion correction of pet brain images through deconvolution: I. theoretical development and analysis in software simulations. *Physics in Medicine and Biology*, 54:797–811, February 2009.

[19] H. Foroosh, J.B. Zerubia, and M. Berthod. Extension of phase correlation to subpixel registration. *IEEE Transactions on Image Processing*, 11(3):188–200, Mar 2002.

[20] David A. Forsyth and Jean Ponce. *Computer vision: a modern approach*, chapter 10, pages 215–218. Prentice Hall, 2003.

[21] David A. Forsyth and Jean Ponce. *Computer vision: a modern approach*, chapter 15, pages 346–348. Prentice Hall, 2003.

[22] David A. Forsyth and Jean Ponce. *Computer vision: a modern approach*, chapter 17, pages 383–388. Prentice Hall, 2003.

[23] R. Fulton, I. Nickel, L. Tellmann, S. Meikle, U. Pietrzyk, and H. Herzog. Event-by-event motion compensation in 3d pet. *IEEE Nuclear Science Symposium Conference Record*, 5:3286–3289, 2003.

[24] R. Fulton, L. Tellmann, U. Pietrzyk, O. Winz, I. Stangier, I. Nickel, A. Schmid, S. Meikle, and H. Herzog. Accuracy of motion correction methods for pet brain imaging. *IEEE Nuclear Science Symposium Conference Record*, 7:4226–4230, Oct. 2004.

[25] R.R. Fulton, S.R. Meikle, S. Eberl, J. Pfeiffer, and C.J. Constable. Correction for head movements in positron emission tomography using an optical motion-tracking system. *IEEE Transactions on Nuclear Science*, 49(1):116–123, 2002.

[26] X.W. Gao, S. Anishenko, D. Shaposhnikov, L. Podladchikova, S. Batty, and J. Clark. High-precision detection of facial landmarks to estimate head motions based on vision models. *Journal of Computer Science*, 3(7):528–532, 2007.

[27] S.R. Goldstein, M.E. Daube-Witherspoon, M.V. Green, and A. Eidsath. A head motion measurement system suitable for emission computed tomography. *IEEE Transactions on Medical Imaging*, 16(1):17–27, 1997.

[28] Michael V. Green, Jrgen Seidel, Stacey D. Stein, Thomas E. Tedder, Kenneth M. Kempner, Caroline Kertzman, and Tom A. Zeffiro. Head movement in normal subjects during simulated pet brain imaging with and without head restraint. *The Journal of Nuclear Medicine*, 35(9):1538–1546, 1994.

[29] H. Guo, R. Renaut, K. Chen, and E. Reiman. Clustering huge data sets for parametric pet imaging. *Journal of Biosystems*, 71:8192, 2001.

[30] Yanling Hao, Zhilan Xiong, Feng Sun, and Xiaogang Wang. Comparison of unscented kalman filters. *International Conference on Mechatronics and Automation*, pages 895–899, 2007.

[31] T. Hasegawa, Y. Fukushima, H. Muraishi, T. Nakano, T. Kuribayashi, Y. Shiba, K. Maruyama, T. Yamaya, E. Yoshida, H. Murayama, N. Hagiwara, and T. Obi. Motion correction for jpet-d4: improvement of measurement accuracy with a solid marker. *IEEE Nuclear Science Symposium Conference Record*, 3:4, 2005.

[32] Hans Herzog, Lutz Tellmann, Roger Fulton, Isabelle Stangier, Elena Rota Kops, Kay Bente, Christian Boy, Rene Hurlemann, and Uwe Pietrzyk. Motion artifact reduction on parametric pet images of neuroreceptor binding. *Journal of Nuclear Medicine*, 46(6):1059–1065, 2005.

[33] Derek L. Hill, Colin Studholme, and David J. Hawkes. Voxel similarity measures for automated image registration. *Visualization in Biomedical Computing*, 2359(1):205–216, 1994.

[34] Derek L G Hill, Philipp G Batchelor, Mark Holden, and David J Hawkes. Medical image registration. *Physics in Medicine and Biology*, 46(3):R1–R45, 2001.

[35] C.K. Hoh, M. Dahlbom, G Harris, Y. Choi, R.A. Hawkins, M.E. Phelps, and J. Maddahi. Automated iterative three-dimensional registration of positron emission tomography images. *Journal of Nuclear Medicine*, 34:311321, 1993.

[36] T. Horprasert, Y. Yacoob, and L. S. Davis. Computing 3-d head orientation from a monocular image sequence. In *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, page 242, Washington, DC, USA, 1996. IEEE Computer Society.

[37] Dongming Hu, C. Hayden, M. Casey, and Z. Burbar. Stereo computer vision system for measuring movement of patient's head in pet scanning. *IEEE Nuclear Science Symposium Conference Record*, 5:2864–2867, 2004.

[38] Anil K. Jain, Yu Zhong, and Sridhar Lakshmanan. Object matching using deformable templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(3):267–278, 1996.

[39] S.J. Julier and J.K. Uhlmann. Unscented filtering and nonlinear estimation. In *Proceedings of the IEEE*, volume 92, pages 401–422, 2004.

[40] A.Z. Kyme, V.W. Zhou, S.R. Meikle, and R.R. Fulton. An optical tracking system for motion correction in small animal pet. *IEEE Nuclear Science Symposium Conference Record*, 5:3555–3559, 26 2007-Nov. 3 2007.

[41] Tali Lerner. Motion correction in fmri images:. Master's thesis, Israel Institute of Technology, 2006.

[42] Kang-Ping Lin, Sung-Cheng Huang, Dan-Chu Yu, William Melega, Jorge R Barrio, and Michael E. Phelps. Automated image registration for fdopa pet studies. *Phys. Med. Biol.*, 41:2775–2788, 1996.

[43] B.J. Lopresti, A. Russo, W.F. Jones, T. Fisher, D.G. Crouch, D.E. Altenburger, and D.W. Townsend. Implementation and performance of an optical motion tracking system for high resolution brain pet imaging. *IEEE Transactions on Nuclear Science*, 46(6):2059–2067, 1999.

[44] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[45] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An invitation to 3-d vision: from images to geometric models*. Springer-Verlag New York, Inc., 2004.

[46] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An invitation to 3-d vision: from images to geometric models*, chapter 2, pages 15–43. Springer-Verlag New York, Inc., 2004.

[47] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, April 1997.

[48] J. B. A. Maintz, P. A. van den Elsen, and M. A. Viergever. Comparison of edge-based and ridge-based registration of ct and mr brain images. *Medical image analysis*, 1(2):151–161, 1996.

[49] J. B. Antoine Maintz and Max A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.

[50] G. Malandain, S. Fernandez-Vidal, and J.-M. Rocchisani. Rigid registration of 3-d objects by motion analysis. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, volume 1, pages 579–581, Oct 1994.

[51] Y. Matsumoto and A. Zelinsky. Real-time stereo face tracking system for visual human interfaces. In *Proceedings of International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pages 77–82, 1999.

[52] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank. Non-rigid multimodality image registration. *Medical Imaging 2001: Image Processing*, page 16091620, 2001.

[53] M. Menke, M.S. Atkins, and K.R. Buckley. Compensation methods for head motion detected during pet imaging. *IEEE Transactions on Nuclear Science*, 43(1):310–317, 1996.

[54] Satoshi Minoshima, Kevin L. Berger, Kien S. Lee, and Mark A. Mintun. An automated method for rotational correction and centering of three-dimensional functional brain images. *The Journal of Nuclear Medicine*, 33(8):1579–1585, 1992.

[55] M.H. Moghari and P. Abolmaesumi. Point-based rigid-body registration using an unscented kalman filter. *IEEE Transactions on Medical Imaging*, 26(12):1708–1728, 2007.

[56] L.-P. Morency, A. Rahimi, N. Checka, and T. Darrell. Fast stereo-based head tracking for interactive environments. In *Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 375–380, 2002.

[57] L.-P. Morency, A. Rahimi, and T. Darrell. Adaptive view-based appearance models. In *Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–803–I–810, June 2003.

[58] H. Muraishi, T. Hasegawa, K. Yoda, A. Takeuchi, Y. Shiba, K. Maruyama, K. Kitamura, T. Yamaya, E. Yoshida, and H. Murayania. Head motion correction for jpet-d4. *IEEE Nuclear Science Symposium Conference Record*, 4:2352–2355, 2004.

[59] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626, April 2009.

[60] Raymond F. Muzic and Shawn Cornelius. Comkat: compartment model kinetic analysis tool. *Journal of Nuclear Medicine*, 42:636–645, 2001.

[61] R. Newman, Y. Matsumoto, S. Rougeaux, and A. Zelinsky. Real-time stereo tracking for head pose and gaze estimation. In *Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 122–128, 2000.

[62] Robert Niese, Ayoub Al-Hamadi, and Bernd Michaelis. A stereo and color-based method for face pose estimation and facial feature extraction. In *Proceedings of the 18th International Conference on Pattern Recognition*, pages 299–302, Washington, DC, USA, 2006. IEEE Computer Society.

[63] Shay Ohayon and Ehud Rivlin. Robust 3d head tracking using camera pose estimation. In *Proceedings of the 18th International Conference on Pattern Recognition*, pages 1063–1066, Washington, DC, USA, 2006. IEEE Computer Society.

[64] J. Orchard. Globally optimal multimodal rigid registration: An analytic solution using edge information. *IEEE International Conference on Image Processing*, 1:485–488, 2007.

[65] J. Pascau, J.D. Gispert, M. Soto-Montenegro, A. Rodriguez-Ruano, V. Garcia-Vazquez, A. Udias, J.J. Vaquero, and M. Desco. Small-animal pet registration method with intrinsic validation designed for large datasets. *IEEE Nuclear Science Symposium Conference Record*, 5:3751 – 3753, 2007.

[66] Xavier Pennec, Charles R. G. Guttmann, and Jean-Philippe Thirion. Feature-based registration of medical images: Estimation and validation of the pose accuracy. In *MICCAI '98: Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 1107–1114, London, UK, 1998. Springer-Verlag.

[67] F. Perruchot, A. Reilhac, C. Grova, A.C. Evans, and A. Dagher. Motion correction of multi-frame pet data. *IEEE Nuclear Science Symposium Conference Record*, 5:3186–3190, Oct. 2004.

[68] Y. Picard and C.J. Thompson. Digitized video subject positioning and surveillance system for pet. *IEEE Transactions on Nuclear Science*, 42(4):1024–1029, 1995.

[69] Y. Picard and C.J. Thompson. Motion correction of pet images using multiple acquisition frames. *IEEE Transactions on Medical Imaging*, 16(2):137–144, 1997.

[70] U. Pietrzyk, K. Herholz, G. Fink, A. Jacobs, R. Mielke, I. Slansky, M. Wurker, and W. Heis. An interactive technique for three-dimensional image registration: validation for pet, spect, mri and ct brain studies. *Journal of nuclear medicine*, 35:20112018, 1994.

[71] Jinyi Qi and R.H. Huesman. List mode reconstruction for pet with motion compensation: a simulation study. In *Proceedings of 2002 IEEE International Symposium on Biomedical Imaging*, pages 413–416, 2002.

[72] Feng Qiao, Tinsu Pan, John W Clark Jr, and Osama R Mawlawi. A motion-incorporated reconstruction method for gated pet studies. *Phys. Med. Biol.*, 51:3769–3783, 2006.

[73] N. Raghunath, T. L. Faber, S. Suryanarayanan, and J. R. Votaw. Motion correction of pet brain images through deconvolution: Ii. practical implementation and algorithm optimization. *Physics in Medicine and Biology*, 54:813–829, February 2009.

[74] A. Rahimi, L.-P. Morency, and T. Darrell. Reducing drift in parametric motion tracking. In *Proceedings of Eighth IEEE International Conference on Computer Vision*, volume 1, pages 315–322, 2001.

[75] A. Rahimi, L-P Morency, and T. Darrell. Reducing drift in differential tracking. In *Computer Vision and Image Understanding*, 2006.

[76] A. Rahmim. Advanced motion correction methods in pet (review article). *Iran J. Nucl. Med.*, 13(241):1–17, 2005.

[77] A. Rahmim, P. Bloomfield, S. Houle, M. Lenox, C. Michel, K.R. Buckley, T.J. Ruth, and V. Sossi. Motion compensation in histogram-mode and list-mode em reconstructions: beyond the event-driven approach. *IEEE Transactions on Nuclear Science*, 51(5):2588–2596, 2004.

[78] A. Rahmim, K. Dinelle, J.-C. Cheng, M. A. Shilov, W. P. Segars, S. C. Lidstone, S. Blinder, O. G. Rousset, H. Vajihollahi, B. M. W. Tsui, D. F. Wong, and V. Sossi. Accurate event-driven motion compensation in high-resolution pet incorporating scattered and random events. *IEEE Transactions on Medical Imaging*, 27(8):1018–1033, Aug. 2008.

[79] A. Rahmim, O.G. Rousset, D.F. Wong, J.C. Cheng, K. Dinelle, V. Sossi, M. Shilov, W.P. Segars, and B.M.W. Tsui. System matrix modeling of externally tracked motion.

Nuclear Science Symposium Conference Record, 2006. IEEE, 4:2137–2141, 29 2006-Nov. 1 2006.

[80] A. Reilhac, G. Batan, C. Michel, C. Grova, J. Tohka, N. Costes, and A.C. Evans. Validation of pet sorteo: a platform for simulating realistic pet studies and development of a database of simulated pet volumes. *IEEE Trans. Nucl. Sci.*, 52:13211328, 2004.

[81] D. Rueckert, L.I. Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, and D.J. Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[82] Urs E. Ruttimann, Paul J. Andreason, and Daniel Rio. Head motion during positron emission tomography: is it significant? *Psychiatry Research: Neuroimaging*, 61(1):43–51, 1995.

[83] Ahmed Saad, Ghassan Hamarneh, Torsten Möller, and Ben Smith. Kinetic modeling based probabilistic segmentation for molecular images. In *MICCAI '08: Proceedings of the 11th international conference on Medical Image Computing and Computer-Assisted Intervention - Part I*, pages 244–252, Berlin, Heidelberg, 2008. Springer-Verlag.

[84] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[85] Richard Szeliski and Stéphane Lavallée. Matching 3-d anatomical surfaces with non-rigid deformations using octree-splines. *Int. J. Comput. Vision*, 18(2):171–186, 1996.

[86] Jean-Philippe Thirion. New feature points based on geometric invariants for 3d image registration. *Int. J. Comput. Vision*, 18(2):121–137, 1996.

[87] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

[88] Rudolph van der Merwe, Nando de Freitas, Arnaud Doucet, and Eric Wan. The unscented particle filter. Technical Report CUED/F-INFENG/TR380, Cambridge University Engineering Department, August 2000.

[89] Andrea Vedaldi. An open implementation of sift. http://vision.ucla.edu/ vedaldi/code/sift/sift.html.

[90] E. Wan and R. van der Merwe. The unscented Kalman filter. In S. Haykin, editor, *Kalman Filtering and Neural Networks*. Wiley Publishing, 2001.

[91] Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical report, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 1995.

[92] Jay B. West, J. Michael Fitzpatrick, Matthew Y. Wang, Benoit M. Dawant, Jr. Calvin R. Maurer, Robert M. Kessler, Robert J. Maciunas, Christian Barillot, Didier Lemoine, Andre M. F. Collignon, Frederik Maes, Paul Suetens, Dirk Vandermeulen, Petra A. van den Elsen, Paul F. Hemler, Sandy Napel, Thilaka S. Sumanaweera, Beth A. Harkness, Derek L. Hill, Colin Studholme, Gregoire Malandain, Xavier Pennec, Marilyn E. Noz, Jr. Gerald Q. Maguire, Michael Pollack, Charles A. Pelizzari, Richard A. Robb, Dennis P. Hanson, and Roger P. Woods. Comparison and evaluation of retrospective intermodality image registration techniques. *Medical Imaging 1996: Image Processing*, 2710(1):332–347, 1996.

[93] Sang-Keun Woo, H. Watabe, Yong Choi, Kyeong Min Kim, Chang Choon Park, P.M. Bloomfield, and H. Iida. Sinogram-based motion correction of pet images using optical motion tracking system and list-mode data acquisition. *IEEE Transactions on Nuclear Science*, 51(3):782–788, June 2004.

[94] R. P. Woods, S. R. Cherry, and J. C. Mazziotta. Rapid automated algorithm for aligning and reslicing pet images. *Journal of computer assisted tomography*, 16(4):620–33, 1992.

[95] Roger P. Woods, Scott T. Grafton, Colin J. Holmes, Simon R. Cherry, and John C. Mazziotta. Automated image registration: I. general methods and intrasubject, intramodality validation. *Journal of Computer Assisted Tomography*, 22(1):139–152, 1998.

[96] Roger P. Woods, Scott T. Grafton, John D. G. Watson, Nancy L. Sicotte, and John C. Mazziotta. Automated image registration: Ii. intersubject validation of linear and nonlinear models. *Journal of Computer Assisted Tomography*, 22(1):153–165, 1998.

[97] Ruigang Yang and Zhengyou Zhang. Model-based head pose tracking with stereovision. In *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, page 255, Washington, DC, USA, 2002. IEEE Computer Society.

[98] T.S. Yoo, M.J. Ackerman, W.E. Lorensen, W. Schroeder, V. Chalana, S. Aylward, D. Metaxas, and R. Whitaker. Engineering and algorithm design for an image processing api: a technical report on itk–the insight toolkit. *Stud Health Technol Inform*, 85:586–92, 2002.

[99] V.W. Zhou, A.Z. Kyme, S.R. Meikle, and R.R. Fulton. Event-by-event motion compensation for small animal pet. *IEEE Nuclear Science Symposium Conference Record*, 4:3109–3114, 26 2007-Nov. 3 2007.

[100] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000, 2003.

[101] T. Zuk, S. Atkins, and K. Booth. Approaches to registration using 3d surfaces. *Medical Imaging: Image Processing*, 2167:176187, 1994.