SIMULTANEOUS RF/EO TRACKING AND CHARACTERIZATION OF DISMOUNTS

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Engineering

by

JASON M. BLACKABY
Department of Electrical Engineering
Wright State University

2008
Wright State University
I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Jason M. Blackaby ENTITLED Simultaneous RF/EO Tracking and Characterization of Dismounts BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science in Engineering.

Brian D. Rigling, Ph.D.
Thesis Director

Fred Garber, Ph.D.
Department Chair

Committee on Final Examination

Brian D. Rigling, Ph.D.

Fred Garber, Ph.D.

Xiaodong Zhang, Ph.D.

Joseph F. Thomas, Jr., Ph.D.
Dean, School of Graduate Studies
ABSTRACT


This thesis discusses the fusion of radar frequency (RF) data and electro-optical (EO) data for tracking and characterization of dismounts (i.e., humans). Each of these sensor modalities provides unique information about the location, structure, and movement of a dismount. The person’s location is tracked on the 2D ground plane using RF data for range measurements and EO data for angle measurements. Using this information, measurements are made on the structure and dynamic motion (gait) of the person. An imaging approach is used to create spatio-temporal activity maps along with a three-dimensional reconstruction of the dismount.
## CONTENTS

1 Introduction
   1.1 Previous Work .............................................. 1
   1.2 Outline .................................................. 2

2 Measured Data Set
   2.1 EO Data Model .............................................. 5
       2.1.1 RF Data Model ........................................ 8

3 Background Suppression .............................................
   3.1 EO Segmentation .............................................. 13
   3.2 RF Clutter Suppression ....................................... 18

4 Joint Position Tracking .............................................
   4.1 Detection .................................................. 21
   4.2 Extended Kalman Filter ..................................... 22

5 Spatio-Temporal Activity Maps .....................................
   5.1 EO Activity Maps ............................................ 27
   5.2 RF Activity Maps ............................................. 28
       5.2.1 RF Range Alignment .................................... 28
       5.2.2 RF Activity Map ........................................ 30

6 Three-Dimensional Reconstruction ...................................
   6.1 RF Back-projection ........................................... 33
6.2 EO Tomosynthesis ........................................ 35
6.3 RF and EO Fusion ........................................ 39

7 Conclusion .................................................. 41

References .................................................. 42
LIST OF FIGURES

2.1 Data Collection Setup ............................................ 5
2.2 Projective Camera Geometry .................................. 6
3.1 EO Segmentation Results ........................................ 18
3.2 Doppler Filtering .................................................. 20
4.1 Tracking Results .................................................... 26
5.1 RF Activity Maps .................................................... 32
6.1 RF Back-projection Results ...................................... 36
6.2 EO Tomosynthesis Results ....................................... 39
6.3 Fused 3-D image ..................................................... 40
LIST OF TABLES

3.1 Segmentation Refinement Algorithm ...................................... 16
ACKNOWLEDGEMENT

I would first like to thank my advisor, Brian Rigling, for all of his help and direction. I would like to thank the Air Force Research Laboratory’s Sensors Directorate, Automatic Target Recognition Division (AFRL/RYA) for funding this research. I would like to individually thank William Pierson, formerly AFRL; Kyle Erickson, AFRL; Philip Hanna, AFRL; Gregory Arnold, AFRL; and Erik Blasch, AFRL for their technical input as well as their help with administrative tasks. Finally, I would like to thank my family for all of their help and support.
CHAPTER 1
INTRODUCTION

Remote sensing and characterization of articulated humans undergoing movement is a fairly new and challenging task. While remote sensing of large rigid objects, such as aircraft and ground vehicles has existed for some time, the continuing advancement in digital computing in recent years has thrust remote sensing into new areas. Articulated human motion presents a number of new challenges that do not exist with large rigid objects.

Specifically, humans are smaller relative to the objects largely sensed in the past, thus requiring finer sensor resolution. Humans are made up of an articulated structure with a range of possible poses, while most recognition algorithms are heavily dependent on shape. Humans can perform a wide range of motions making them difficult to model. A vehicle generally moves forward or backward at a variable velocity, where a human can change directions quickly at any time. Human sensing often must be done in the presence of significant structural clutter (e.g. buildings, trees, etc.). Apart from dense urban environments, clutter is less of a problem in vehicle sensing.

1.1 Previous Work

Dismount tracking and characterization has received much attention in recent years. With the advancement in computing power, real-time video processing for human tracking has become possible. A number of studies have looked at the viability of using human gait as a biometric. Studies have shown that human observers can
recognize people and their gender by gait [1][2][3]. The problem of automating this process has been the subject of much research in the areas of signal processing and automatic target recognition. Visual information from EO sensors is the basis for most automated human characterization methods. In most cases, the background of a video sequence is removed leaving just pixels on the person. The resulting binary silhouette and its dynamics is input to a classification or recognition algorithm. The authors of [4] provide a good overview of the state of the art in human gait recognition using EO sensors.

While sensing humans with EO sensors is well developed, RF human sensing is relatively new. Time-frequency analysis of micro-Doppler signatures has been examined [5][6]. Independent component analysis of micro-Doppler signatures has been used to characterize the human gait [7]. Others have looked at the micro-Doppler phenomenon in general [8][9][10]. The authors of [11] used a human walking model to simulate radar signatures and estimate walking parameters from radar data. A simple detector for dismounts using a continuous wave radar was developed in [12]. Some have looked at challenges for ISAR imaging in the presence of micro-Doppler by separating micro-Doppler returns from those from the gross target motion [13][14].

1.2 Outline

This paper considers new methods to characterize humans as they move through a scene. Measured radio-frequency and passive electro-optical sensor dismount data is used as demonstration. A non-scanning, wide angle RF sensor provides good resolution along the range dimension, but little or no resolution in azimuth or elevation. Conversely, an EO sensor provides resolution in the azimuth and elevation dimensions, but little or no resolution in range. By fusing information from both sensors, the dismount can be resolved in azimuth, elevation, and range, which in turn can be mapped to the $x, y, z$ Cartesian space.
Some processing is required before a dismount can be imaged. The RF data contains a large amount of ground clutter that is removed using a standard *moving target indicator* (MTI) filter. The person is segmented in the video using background and shadow subtraction techniques. The quality of the segmentation is a key part of the process since much of the tracking and imaging is directly based upon it. With the dismount returns isolated in both sensor domains, position measurements are made and registered in time and space. The dismount position is tracked using an *extended kalman filter* (EKF). The tracking information is used for dismount motion compensation. With the gross motion of the dismount removed, some imaging techniques can be employed. RF activity maps capture the *micro-motion*, motion apart from the gross dismount motion (i.e., arms and legs) of the dismount in a single image. Tomography is used to form a 3-D reconstruction of the dismount. EO tomosynthesis and RF back-projection are performed on the motion-compensated sensor data to form a 3-D map of the dismount.

The remainder of this paper is outlined as follows. Chapter 2 explains the measured data set and the data models for each sensor modality. Chapter 3 discusses the suppression of background and non-dismount sensor returns. Chapter 4 describes a joint RF/EO dismount tracking system. Chapter 5 explains the creation of spatio-temporal activity maps to capture the nature of dismount motion in a single image. Chapter 6 discusses 3-D imaging techniques for each sensor modality and their fusion to form dismount images.
CHAPTER 2
MEASURED DATA SET

The data used in this study is unique in that it contains fine resolution radar and video measurements of a number of dismount scenarios. The radar is a coherent, pulse-Doppler radar operating in the X ($f_c = 10$ GHz) and Ku ($f_c = 15$ GHz) bands. The radar has a pulse repetition frequency (PRF) of 1000 Hz and a bandwidth of 4 GHz, giving a range resolution of 1.48 inches. This is important for dismount characterization as the motion of the arms and legs can be resolved from the torso. The radar has a beamwidth of roughly $45^\circ$, so it provides very little angle resolution.

Figure 2.1 shows the geometry of the scene in the data collection. The radar was fourteen feet above the ground, which produces the positive effect of separating the dismount returns from head to feet in range.

The EO data was captured with a standard NTSC camera with 720x480 pixel resolution at 30 frames per second. The camera was situated directly below the radar antenna, about five feet above the ground. The camera parameters were unknown and were estimated based on known positions in the scene.

A number of scenarios are contained in the data, including dismounts walking, jogging, running, carrying, standing, and limping. These actions are performed along various paths relative to the sensor platform. RF and EO data were captured simultaneously and can be easily registered in time.
Figure 2.1: **Data Collection Setup.** This figure denotes the ground truth dimensions of the measured data set. The dismounts followed pre-defined paths on the ground plane. The RF and EO sensors were located in the same location, but at different heights. Note that the ground truth is not drawn to scale as the star pattern is not symmetric.

### 2.1 EO Data Model

A basic EO system projects points in 3-D space onto a 2-D image surface. Figure 2.2 illustrates the projective camera geometry and image formation. Let $\gamma$ represent a point in 3-D space and $\hat{\gamma}$ represent the corresponding projected point on the focal plane. The 3-D point is projected onto the focal plane by the projective camera matrix $P$,

$$\hat{\gamma} = P\gamma$$  \hspace{1cm} (2.1)
Figure 2.2: **Projective Camera Geometry.** This figure shows the projective EO system geometry with respect to camera center $c$. The line formed by the camera center and the point $\gamma$ intersects the image plane at the point $\hat{\gamma}$.

where $P$ is a $3 \times 4$ projection matrix determined by the camera characteristics, position, and orientation. A homogenous coordinate system is used such that

$$
\gamma = \begin{bmatrix}
X \\
Y \\
Z \\
1
\end{bmatrix} \quad \text{and} \quad \hat{\gamma} = \begin{bmatrix}
U \\
V \\
1
\end{bmatrix}
$$

The common *pinhole camera* assumption is made in which the camera is represented by a single point, $c$, in 3-D space. Any point, $\gamma$, and $c$ form a line containing the points [15]

$$
\gamma(\lambda) = \lambda \gamma + (1 - \lambda)c
$$

As $\lambda$ varies, $\gamma(\lambda)$ traverses the line formed by $\gamma$ and $c$. One point on this line will occur at the intersection of the image plane with the line. This point, $\hat{\gamma}$ is the projected point on the 2-D image plane of all other points on the line, $\gamma(\lambda)$, so (2.1) can
be extended to

\[ \hat{\gamma} = P\gamma(\lambda) \]  

(2.4)

This indicates that in a perspective EO system, each 2-D image point can be the projection of any point along a line in 3-D space. Consequently, the 3-D point corresponding to an image point cannot be determined completely by the image point. However, an image point does provide the azimuth(\(\phi\)) and elevation(\(\theta\)) angle of the 3-D ray on which the corresponding 3-D point lies if the camera parameters are known. We define

\[ \phi(U) = \arctan \frac{U}{f} \]  

(2.5a)

\[ \theta(V) = \arctan \frac{V}{f} \]  

(2.5b)

where \(U\) and \(V\) are the coordinates on the image plane with \((U, V) = (0, 0)\) at the center of the image. The variable \(f\) is the focal length of the camera lens. The focal length and the image size may not be known in many instances, but \(\phi\) and \(\theta\) can also be determined from the field of view of the camera as well. The field of view is given by the horizontal (\(\Phi\)) and vertical (\(\Theta\)) angles covered by the extents of the image. The field of view is defined completely by the focal length and image size, but the focal length and image size cannot necessarily be retrieved from the field of view. The field of view can be estimated much easier than the focal length. If the dimensions and range of an object in the scene are known, the field of view can be estimated. However, determining the focal length usually requires calibration based on a complex object such as a checkerboard cube. Assuming there is no skew or
distortion in the image, \( \phi \) and \( \theta \) linearly span \( \Phi \) and \( \Theta \).

\[
\phi(U) = \frac{U}{L_U} \Phi \tag{2.6a}
\]
\[
\theta(V) = \frac{V}{L_V} \Theta \tag{2.6b}
\]

where \( L_U \) and \( L_V \) are the horizontal and vertical lengths of the image plane. The value of this representation can be seen in the digital image domain. Let \( u \) and \( v \) represent the horizontal and vertical pixel indexes with \((u, v) = (0, 0)\) occurring at the center of the image. In this case, \( \phi \) and \( \theta \) are determined by

\[
\phi(u) = \frac{u}{N_u} \Phi \tag{2.7a}
\]
\[
\theta(v) = \frac{v}{N_v} \Theta \tag{2.7b}
\]

where \( N_u \) and \( N_v \) are the number of horizontal and vertical pixels respectively. This indicates that knowledge of the field of view of the camera allows us to determine the projection angles for each pixel in a digital image without additional information. Knowledge of the projection angles is key to the dismount tracking process in Chapter 4 and the back-projection of \( \hat{\gamma} \) to \( \gamma(\lambda) \) used in the tomosynthesis process discussed in Chapter 6.

### 2.1.1 RF Data Model

A pulse Doppler radar generally consists of a transmitter that radiates an electromagnetic wave and a receiver that captures the reflections from objects in a scene. In the monostatic case, the transmitter and receiver are in the same location and usually share the same antenna. Many transmission waveforms are used in radar systems, depending on the application.

For the measured data in this study, a stepped frequency system was employed,
resulting in measured returns well-modeled by

$$S(f_i, \tau_k) = \sum_m A_m \exp \left[ -j2\pi f_i \frac{2R_m(\tau_k)}{c} \right]$$ \hfill (2.8)

where $\tau_k$ represents the time of the $k^{th}$ pulse, $R_m(\tau_k)$ is the range to the $m^{th}$ scatterer at time $\tau_k$, $c$ is the speed of light, $f_i$ is the $i^{th}$ returned frequency sample, and $A_m$ is the reflection coefficient of the $m^{th}$ scatterer. The discrete frequency samples are evenly spaced throughout the bandwidth

$$f_i = f_0 + \frac{iB}{N_f}, i = 0, 1, ... N_f - 1 \hfill (2.9)$$

where $B$ is the bandwidth of the transmitted signal, $f_0$ is the starting frequency of the band, and $N_f$ is the number of frequency samples within the band.

**Range Compression**

After frequency sampling, energy returned from a single range is spread over multiple frequency samples in the form of a complex exponential. Range compression attempts to consolidate the energy from a single range into a distinct sample. Range compression is performed using a matched filter. The frequency sampled signal is simply matched with complex exponentials known to be the result of returns from each range.

Thus, matched filtering can be performed on the discrete frequency-sampled signal given in (2.8) by computing the sum

$$s(r, \tau_k) = \sum_{i=0}^{N_f-1} S(f_i, \tau_k) \exp \left[ j2\pi f_i \frac{2r}{c} \right]$$ \hfill (2.10)
Fast-time, \( t = \frac{2r}{c} \) is sampled according to

\[
\Delta t = \frac{2\Delta R}{c} = \frac{1}{B}, \tag{2.11}
\]
\[
t = \frac{2r}{c} = n\Delta t = \frac{n}{B}, \tag{2.12}
\]

where \( \Delta R \) is the range resolution determined by the bandwidth. The frequency samples are replaced with the values given by (2.9) giving the discrete-time and frequency matched filter,

\[
s(n, \tau_k) = \sum_{i=0}^{N_f-1} S(f_i, \tau_k) \exp \left[ j2\pi \left( f_0 + \frac{iB}{N_f} \right) \left( \frac{n}{B} \right) \right] \tag{2.13a}
\]
\[
= \sum_{i=0}^{N_f-1} S(f_i, \tau_k) \exp \left[ j2\pi \frac{f_0n}{B} \right] \exp \left[ j2\pi \frac{in}{N_f} \right] \tag{2.13b}
\]
\[
= \exp \left[ j2\pi \frac{f_0n}{B} \right] \sum_{i=0}^{N_f-1} S(f_i, \tau_k) \exp \left[ j2\pi \frac{in}{N_f} \right] \tag{2.13c}
\]

The term, \( \exp \left[ j2\pi \frac{f_0n}{B} \right] \), is a focusing phase with a magnitude of one, so it is often discarded when considering the magnitude response. The inner sum is the inverse DFT (IDFT) that can be computed efficiently via Fast Fourier Transform (FFT) algorithm. The \( n \)-indexed samples of \( s(n, \tau_k) \) are representative of range bins. The range of each sample is given by

\[
\tau_n = \frac{cn}{2B} \tag{2.14}
\]

Because the IDFT is performed over a finite number of samples, sidelobes emerge in the output. A windowing function can be applied to the phase history before the DFT to reduce the sidelobes [16].

So, a range-time image is now achievable by performing range compression on
each pulse.

\[ s(r_n, \tau_k) = \sum_{i=0}^{N_f-1} S(f_i, \tau_k) \exp \left[ j2\pi \frac{in}{N_f} \right] \tag{2.15} \]

**Doppler Processing**

The Doppler effect refers to the change in frequency that occurs when an electromagnetic wave comes in contact with a moving object. With a coherent radar, these frequency changes can be determined, giving a range-rate measurement in addition to the range measurement. Most pulsed Doppler radars do not directly measure the Doppler frequency shift since it is insignificant in comparison to the carrier frequency. Doppler frequency shift measurements are made by calculating the rate of change of the phase of the returned signal

\[ f_d = \frac{1}{2\pi} \frac{d\psi}{dt} \tag{2.16} \]

Consider the phase history of a single scatterer over a series of pulses:

\[ S_m(f_i, \tau_k) = A_m \exp \left[ -j4\pi f_i \frac{R_m(\tau_k)}{c} \right] \tag{2.17} \]

Using (2.16), the Doppler frequency shift is given by

\[ f_d(f_i, \tau_k) = \frac{-2f_i dR_m(\tau_k)}{c} \tag{2.18} \]

which shows that the Doppler frequency shift is directly dependent on the radial velocity with respect to time, \( \frac{dR(\tau_k)}{dt} \).

Doppler processing is often performed on range compressed data in order to produce a range-Doppler image. Doppler processing simply involves extracting the rate of change of the phase with respect to slow-time. Since \( \frac{d\psi}{dt} \) represents the frequency, the
DFT can be applied to transform the range compressed data into the range-Doppler image, \( S(r_n, f_q) \).

\[
S(r_n, f_q) = \sum_{k=0}^{N_d-1} s(r_n, \tau_k) \exp \left[ -j2\pi \frac{k q}{N_d} \right]
\]  

(2.19)

Here, \( f_q \) are the samples in the Doppler frequency domain, and \( N_d \) is the number of slow-time pulses over which the DFT is performed and therefore equals the number of Doppler frequency samples obtained. The Doppler frequency samples are dependent on the PRF \( (f_p) \)

\[
f_q = \frac{q f_p}{N_d}, \quad q = -\frac{N_d}{2}, ..., 0, ..., \frac{N_d}{2} - 1
\]

(2.20)

This indicates that the quality of a range-Doppler image is directly related to the number of pulses used in the slow-time DFT. The Doppler resolution is given by

\[
\Delta f_q = \frac{f_p}{N_d}
\]

(2.21)

so as \( N_d \) is increased with a fixed PRF, a finer resolution is achieved. Note that regardless of the number of Doppler frequency samples, \( f_q \) covers the same bandwidth; \( f_q \in \left[-\frac{f_p}{2}, \frac{f_p}{2}\right] \).
CHAPTER 3
BACKGROUND SUPPRESSION

Here we seek to separate the sensor returns that originate from the dismount from those that pertain to noise, background, or other sources. All subsequent processing hinges on the quality of the segmentation in each sensor domain. In the RF domain, segmentation involves clutter removal; while in the EO domain, segmentation involves background removal.

3.1 EO Segmentation

Segmenting a moving object from a static background video frame is a common task in computer vision. Since the background is static, background modeling techniques are used. Each frame is compared to the background model, and regions that are not similar to the model are marked as foreground. Let the sequence of EO frames be represented as

\[
i_k(u, v) = \begin{bmatrix} r_k(u, v) \\ g_k(u, v) \\ b_k(u, v) \end{bmatrix}
\]  

(3.1)

where \( k \) is the frame index and \( r_k, g_k, \) and \( b_k \) are the red, green, and blue color components.

Many approaches have been proposed in the literature for background modeling. The simplest form of background modeling is frame differencing. In this method, the
previous frame is considered as the background. Simple frame differencing provides poor results when the object is moving slowly enough to occupy some of the same pixel region in sequential frames. In this case, only the leading edges of the object are sensed as different from the background. A better background model uses the median pixel value over a sequence of frames as the background model [17] [18]. Let $\mathbf{b}$ represent the background RGB image. The median background is given by

$$\mathbf{b}(u,v) = \mathcal{M}_k(i_k(u,v))$$

(3.2)

where $\mathcal{M}_k$ represents the median function over $k$. The median model assumes that any given pixel takes on the background value for a longer period of time than the foreground value over the sequence of frames. The median value is used as opposed to the mean value since it provides a value that actually occurs in the image sequence. The mean pixel value over a sequence of frames is affected not only by the background, but the foreground as well, resulting in a slightly skewed background estimation for pixels that contain foreground values for some of the frames in the sequence.

Some other background modeling techniques include mixture of Gaussians [19] and Eigenbackgrounds [20]. These methods provide better performance in the presence of noise and non-stationary backgrounds; however, they increase complexity. Median background modeling produces similar results to the mixture of Gaussian and Eigenbackground techniques in the data used in this research; however, a dynamic scene might require one of these advanced methods.

After the background image is calculated, it is compared with each frame of the image sequence to determine foreground and background regions. To reduce the effect of noise, the neighborhood around each pixel is also factored into the difference. The luminance (intensity) difference, $L_k(u,v)$ between each frame and the background is
computed as

\[
L_k(u, v) = \frac{\sum_{m=u-M}^{u+M} \sum_{n=v-M}^{v+M} w(m, n) |B(m, n) - I_k(m, n)|}{(2M + 1)^2}
\] (3.3)

where \( I \) and \( B \) represent the intensity values of the foreground and background, respectively. The intensity value is a scalar that can be calculated from the color vector

\[
I_k(u, v) = [0.299\ 0.587\ 0.114] \hat{i}_k(u, v)
\] (3.4)

The size of the neighborhood around each pixel to be processed is denoted by \( M \) and \( w(m, n) \) is a weighting factor for each pixel within the neighborhood. The weighting should have the highest weight for the center pixel \((m = n)\) and should reduce as the distance from the center pixel increases. A two-dimensional Gaussian function is used for \( w(m, n) \). While not necessary, \( \sum_{mn} w(m, n) = 1 \) will maintain the same scaling as the original intensity image. Equation (3.3) can also be calculated using

\[
L_k(u, v) = [B(u, v) - I_k(u, v)] * w(m, n)
\] (3.5)

where \(*\) denotes two-dimensional convolution. This is essentially a blurring of the difference between the foreground and background intensities.

A binary mask, \( D_k(u, v) \), is created that contains 1’s for foreground pixels and 0’s for background pixels. A threshold, \( \epsilon \), is applied to \( L_k(u, v) \) to create the mask. The threshold is set empirically, and depends on the scene and the noise present. In this case, \( \epsilon = 0.04 \) was chosen.

\[
D_k(u, v) = \begin{cases} 
1 & |L_k(u, v)| \geq \epsilon \\
0 & \text{otherwise}
\end{cases}
\] (3.6)
Table 3.1: Segmentation Refinement Algorithm

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Binary Clean; 2 times</td>
</tr>
<tr>
<td>2</td>
<td>Binary Majority; 2 times</td>
</tr>
<tr>
<td>3</td>
<td>Binary Close; 5x5</td>
</tr>
<tr>
<td>4</td>
<td>Label Regions</td>
</tr>
<tr>
<td>5</td>
<td>Keep Largest Region</td>
</tr>
<tr>
<td>6</td>
<td>Fill Holes</td>
</tr>
<tr>
<td>7</td>
<td>Binary Dilation; 3x3, 3 times</td>
</tr>
<tr>
<td>8</td>
<td>Binary Erosion; 3x3, 2 times</td>
</tr>
</tbody>
</table>

The foreground mask, $D_k(u,v)$, will typically contain a number of false foreground classifications and missed foreground detections, so additional processing is needed to segment out a single contiguous region representing the dismount. The assumption is made that there is a single dismount present in the scene, so the largest group of pixels classified as foreground is assumed to represent the dismount. A morphological region merging and segmentation algorithm is used to form a single contiguous region. The refinement algorithm is shown in Table 3.1. Let $\tilde{D}_k(u,v)$ represent the refined foreground mask.

The primary shortcoming of this background subtraction technique is that shadows cast by the dismount are also classified as foreground pixels. A separate process is needed to remove the shadow from the frame. Shadow removal exploits the fact that shadows are similar in color to the background, but have a lower intensity value. To test for color similarity, a colorspace invariant to intensity changes is used. A number of intensity invariant colorspaces exist. A proposed [21] colorspace invariant to shading and intensity changes in matte surfaces is given by

$$c_1 = \arctan\left( \frac{r}{\max(g,b)} \right)$$  \hspace{1cm} (3.7a)

$$c_2 = \arctan\left( \frac{g}{\max(r,b)} \right)$$  \hspace{1cm} (3.7b)

$$c_3 = \arctan\left( \frac{b}{\max(r,g)} \right)$$  \hspace{1cm} (3.7c)
Let $\tilde{i}_k$ be the $k$th frame in the invariant colorspace, and let $\tilde{b}$ denote the background image in the invariant colorspace. A pixel is classified as a shadow if it has a similar value to the background in the $c_1, c_2, c_3$-colorspace and has a lower intensity value than the background. Difference in the $c_1, c_2, c_3$-colorspace is measured as the sum of squared differences:

$$C_k(u, v) = \|\tilde{b}(u, v) - \tilde{i}_k(u, v)\| * w(m, n)$$  \hspace{1cm} (3.8)

Again, the difference is convolved with a weighting function to allow adjacent pixels to affect the difference value of each pixel.

A shadow mask is created with ones representing shadow pixels and zeros representing foreground pixels with the following boolean function

$$S_k(u, v) = [C_k(u, v) < \epsilon_c] \cap [\epsilon_1 < L_k(u, v) < \epsilon_2] \cap [D_k(u, v)]$$  \hspace{1cm} (3.9)

where $\cap$ denotes the intersection of binary masks. The thresholds, $\epsilon_1$ and $\epsilon_2$, are lower and upper bounds on the intensity difference between the frame and the background. The intensity difference is upper and lower thresholded since shadow areas usually lower the background intensity by a constant value. If the intensity is lowered by a larger amount than the upper threshold, the pixel is most likely in the foreground. The color difference between the foreground and background is thresholded by $\epsilon_c$. Equation (3.9) indicates that a shadow is declared if the color difference between frame and background is lower than a threshold, and the frame has a lower intensity than the background within some bounds. Shadow pixels must also have been declared foreground pixels in $D_k(u, v)$.

Using the knowledge that the shadow cast by a dismount will be a single region,
Figure 3.1: **EO Segmentation Results.** Pixels classified as foreground are highlighted and pixels classified as shadow are shaded. The edges of the foreground and shadow sections are marked. Note that the image is zoomed and does not represent the full extent of the view

the region refinement algorithm can be used to select a single contiguous shadow region. Let $\tilde{S}_k(u, v)$ denote the shadow mask after region refinement.

A single binary mask, $G_k(u, v)$ is formed that contains only foreground pixels that are not classified as shadow:

$$G_k(u, v) = \tilde{D}_k(u, v) \cap \tilde{S}_k(u, v)' \quad (3.10)$$

where $(...)'$ denotes binary negation. The result $G_k$ ideally contains only pixels that lie on the dismount. Figure 3.1 Shows the EO segmentation results for a single frame.

### 3.2 RF Clutter Suppression

Segmentation in the RF domain is better described as clutter suppression. The aim is to reject all radar returns that do not originate from the dismount. MTI filtering is widely used in the Doppler domain to reject returns from objects with zero radial velocity with respect to the radar. Coherent subtraction is also useful for
clutter suppression, however, it requires a stable radar platform and effective motion-compensation. Coherent subtraction has the advantage that it preserves radar returns with zero velocity. In many moving target applications, zero radial velocity returns do not offer much useful information, but when extracting features from dismount motion, zero radial velocity returns are important. The feet have zero velocity when in contact with the ground, and the arms usually show negative velocities on their back swing.

MTI filtering is accomplished by simply applying a notched filter in the Doppler domain. The notch occurs at \( f_d = 0 \), which represents zero radial velocity. The most common way to apply the filter is a sinusoid (single delay-line canceller [22]) with the magnitude response,

\[
H_{\text{MTI}}(r_n, f_q) = 2 \sin(2\pi f_q T_p) 
\] (3.11)

This gives zeros at frequency multiples of \( \pm \frac{f_q}{T_p} = if_p \) for \( i = 0, 1, 2, ... \). Note that the MTI filter does not depend on \( r_n \) so the same filter is applied for each range cell. MTI filtering is performed by

\[
\tilde{s}_m(r_n, \tau_k) = s_m(r_n, \tau_k) \circ h_{\text{MTI}}(r_n, \tau_k) 
\] (3.12)

where \( \circ \) denotes 1-D convolution across slow-time and \( h_{\text{MTI}}(r_n, \tau_k) \) is the slow time domain representation of \( H_{\text{MTI}}(r_n, f_q) \). For faster performance, MTI filtering is performed in the discrete Doppler frequency domain where convolution is replaced by multiplication.

\[
\tilde{S}_m(r_n, f_q) = S_m(r_n, f_q)H_{\text{MTI}}(r_n, f_q) 
\] (3.13)

Here, \( H_{\text{MTI}} \) is constant with respect to \( r_n \). Beyond clutter removal, Doppler filtering
Figure 3.2: **Doppler Filtering.** Here, all returns but those from the hands and feet are suppressed based on their Doppler shifts. The dismount is performing a standard walking motion while moving toward the sensor platform. The forward motion of the hands and feet can be seen.

is used to separate dismount parts based on radial velocity. By filtering out all slow moving returns, the forward motion of the hands and feet are isolated as shown in Figure 3.2. This can obviously be used for very accurate gait measurements such as step frequency and stride length. Note that since the radar is 14 feet above the ground, returns from the feet occur at a range farther than those of the hands relative to the radar. Doppler frequency filters are designed using common FIR digital filter design techniques[23].
CHAPTER 4
JOINT POSITION TRACKING

Tracking the dismount on the two-dimensional ground plane is accomplished using the radar for a range measurement and the video for an azimuth angle measurement. It is assumed that there is no variation in elevation since the dismount is walking along the ground plane. Since the radar is stationary, has a wide beamwidth, and does not scan, it provides almost no azimuth resolution. The video along with the camera parameters give fairly accurate azimuth information. Using both of the sensors allows for accurate position tracking of a dismount. To join the RF and EO data and track the dismount, an extended Kalman filter is used.

4.1 Detection

Range and range rate measurements are detected from the RF data and an angle measurement is detected from the EO data. Since the data is received at two differing rates (30 Hz for EO and 1000-1200 Hz for RF), the measurement period is set to the slowest rate, $T = \frac{1}{30}$. Range and range rate are detected in the range-Doppler domain. Doppler processing is performed on a window centered at the pulse that corresponds in time with the current EO frame. Range is detected as the farthest range with a radar return above a threshold. Since the radar is above the ground, the farthest range on the dismount corresponds with the foot in contact with the ground plane. Detecting range to the point on the ground plane allows the range vector to be
accurately projected onto the ground.

\[ r = \sqrt{r_d^2 - h_r^2} \]  \hspace{1cm} (4.1)

where \( r_d \) is the detected range from radar to foot, \( r \) is the length of the projection of \( r_d \) onto the ground plane, and \( h_r \) is the height of the radar.

Dismount Doppler frequency is detected from the range-Doppler data as the Doppler frequency of the intensity centroid. Radial velocity is then calculated from the Doppler shift:

\[ \dot{r}_d = \frac{f_d c}{2 f_c} \]  \hspace{1cm} (4.2)

Range rate is also projected onto the ground plane with the same ratio as the range measurement,

\[ \dot{r} = \frac{\dot{r}_d}{r_d} \]  \hspace{1cm} (4.3)

The angle, \( \phi \) is detected as the azimuth angle of the ray formed by the camera center and the centroid pixel of the EO mask, \( G_k \). Equation (2.7a) relates the centroid pixel position \((u, v)\) to azimuth angle \( \phi \).

### 4.2 Extended Kalman Filter

The Kalman filter estimates the state of a dynamic system when the measurements are corrupted by noise. The Kalman filter is recursive, making it computationally efficient for real-time tracking [24]. The extended Kalman filter (EKF) is an extension of the Kalman filter used in the case of a non-linear relationship between the measurement space and the state space or a non-linear state update equation. The EKF is used in this case because of the non-linear conversion from the polar measurement
space to the Cartesian state space.

Let the state vector \( \hat{x} \) represent the state estimate vector. In this case, a state estimate is made for each EO frame, so the index \( k \) is used for the discrete time instance. The state vector, \( x \), and the measurement vector, \( z \), are structured as follows

\[
\begin{align*}
x &= \begin{bmatrix} x \\ \dot{x} \\ y \\ \dot{y} \end{bmatrix} \\
z &= \begin{bmatrix} r \\ \dot{r} \\ \phi \end{bmatrix}
\end{align*}
\tag{4.4}
\]

where \( x \) and \( y \) are the Cartesian coordinates of the dismount on the ground plane.

The Kalman filter can be broken into two parts: prediction and innovation. In the prediction stage, the previous estimate and covariance information is used along with the dynamic model to predict the next state and state covariance. Prediction gives \( \hat{X}_{k|k-1} \) and \( P_{k|k-1} \), which are the state estimates based on all of the previous measurements, but not the current measurement. Innovation involves using the current measurement to update the estimate and estimate covariance. The subscript notation, \( \hat{X}_{p|q} \), denotes the state estimate at discrete time instant \( p \) based on measurements 1, 2, ..., \( q \). The extended Kalman filter equations are as follows.

Prediction:

\[
\begin{align*}
\hat{X}_{k|k-1} &= F\hat{X}_{k-1|k-1} \\
P_{k|k-1} &= FP_{k-1|k-1}F^T + Q \\
\hat{z}_k &= h(\hat{X}_{k|k-1}) \\
H_k &= \frac{\partial h}{\partial x} \bigg|_{x=\hat{X}_{k|k-1}}
\end{align*}
\tag{4.5}
\]
Innovation:

\[
K_{k} = P_{k|k-1}H_{k}^{T}(H_{k}P_{k|k-1}H_{k}^{T} + R)^{-1} \quad (4.5e)
\]

\[
\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_{k}(z_{k} - \hat{z}_{k}) \quad (4.5f)
\]

\[
P_{k|k} = (I - K_{k}H_{k})P_{k|k-1} \quad (4.5g)
\]

Above, \( F \) is the dynamic state model, \( P \) is the state error covariance, \( Q \) is the covariance of the system noise, \( \hat{z} \) is the measurement prediction, \( H \) is the measurement matrix, \( R \) is the measurement noise covariance, \( K \) is the Kalman gain, and \( I \) is the identity matrix.

Matrix, \( F \), is the dynamic model for state estimation. This model describes the motion of the dismount based on the previous estimate. A constant velocity model is used:

\[
F = \begin{bmatrix}
1 & T & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & T \\
0 & 0 & 0 & 1 \\
\end{bmatrix} \quad (4.6)
\]

where \( T \) is the sampling period of the system (\( T = \frac{1}{30} \) seconds per frame), and \( Q \) is the covariance matrix of the noise in the system. For the constant velocity model,

\[
Q = \begin{bmatrix}
\frac{T^4}{4} & \frac{T^3}{2} & 0 & 0 \\
\frac{T^3}{2} & T^2 & 0 & 0 \\
0 & 0 & \frac{T^4}{4} & \frac{T^3}{2} \\
0 & 0 & \frac{T^3}{2} & T^2 \\
\end{bmatrix} \sigma^2 \quad (4.7)
\]

where \( \sigma^2 \) is the variance of the system. A lower system variance results in more reliance on the model than the measurement and a high model variance results in more reliance on the measurement to make the state estimate. In this case, an analytical
solution for $\sigma^2$ is not possible since the nature of the dismount motion cannot be predicted, therefore $Q$ is chosen empirically, based on the type of movement being tracked.

Equations (4.5c) and (4.5d) present the extension of the Kalman filter for the non-linear relationship between $x$ and $z$. In this case, $h(x)$ is a function that converts an estimate in state coordinates to measurement coordinates.

\[
h(x) = h\left(\begin{bmatrix} x \\ \dot{x} \\ y \\ \dot{y} \end{bmatrix}\right) = \begin{bmatrix} \sqrt{x^2 + y^2} \\ \frac{x\dot{y} + y\dot{x}}{\sqrt{x^2 + y^2}} \\ \arctan(y/x) \end{bmatrix}
\] (4.8)

Since calculation of the Kalman gain, $K_k$, and the estimate error covariance, $P_{k|k}$, require a matrix as a linear operator, $H_k$ must be estimated by a Jacobian matrix. Thus, (4.5d) is evaluated as

\[
\frac{\partial h(x)}{\partial x} = \begin{bmatrix} \frac{\partial x}{\partial x} & \frac{\partial x}{\partial y} & \frac{\partial y}{\partial x} & \frac{\partial y}{\partial y} \\ \frac{\partial \dot{x}}{\partial x} & \frac{\partial \dot{x}}{\partial y} & \frac{\partial \dot{y}}{\partial x} & \frac{\partial \dot{y}}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} & \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{x}{r} & 0 & \frac{y}{r} & 0 \\ 0 & \frac{x}{r} & 0 & \frac{y}{r} \\ \frac{y}{r^2} & 0 & \frac{-x}{r^2} & 0 \end{bmatrix}
\] (4.9)

Here, $\frac{\partial \dot{x}}{\partial x} = \frac{\partial \dot{y}}{\partial y} = 0$ according to the formulation of the alternative EKF in [25] where the authors show that this linearization is less susceptible to numerical error. The measurement noise covariance $R$ is difficult to quantify analytically since the quality of the measurements depend on the sensor noise, the quality of the segmentation, and accuracy of the sensors. In this case, $R$ is determined by empirical results. Figure 4.1 shows the filtered track as well as the noisy measurements.
Figure 4.1: **Tracking Results.** The track on the ground plane is represented by the bold line while the measurements are shown by the thin line. The dashed star pattern represents the ground truth markings that can be seen in Figures 2.1 and 3.1.
CHAPTER 5
SPATIO-TEMPORAL ACTIVITY MAPS

Spatio-temporal activity maps capture dynamic motion in a single image. Activity maps are used as a first step in measuring the dynamic movement of the dismount since the method is fairly simple to implement and once a single image is obtained, standard image pattern recognition algorithms can be employed. Spatio-temporal images formed using silhouette sequences have been considered in recent literature. The same technique can be applied in the RF domain. An RF activity map is formed by performing range-doppler processing over a CPI that contains an entire gait cycle. A CPI of this length results in some phase errors, so the short time Fourier transform (STFT) is used to integrate smaller windows non-coherently.

5.1 EO Activity Maps

Several spatio-temporal representations of dismount masks have been introduced in recent literature. A motion history image (MHI) was presented in [26] where MHI pixel intensity corresponds with how recent motion in that pixel occurred. The gait energy image (GEI) proposed in [27] is simply the time average of the binary silhouette. The authors of [28] combine the MHI and GEI to form the gait history image (GHI). The gait moment image (GMI) presented in [29] uses a number of moments as features in addition to the GEI. The results from these studies will not be reproduced in this thesis.
5.2 RF Activity Maps

A range alignment process must be completed before the activity map can be formed. Without alignment, integration over time results in smearing since the dismount is moving through the scene.

5.2.1 RF Range Alignment

Range alignment is performed in two steps. A coarse adjustment is performed using the tracking information. Sub-pixel range errors are then corrected using an entropy-based alignment.

The coarse range alignment is performed on the phase history data using the tracking information. The phase correction is constant over a single pulse.

\[
S_{ra}(f_i, \tau_k) = S(f_i, \tau_k) \exp \left[ -j2\pi f_i \left( \frac{2R_c(\tau_k)}{c} \right) \right]
\]  
(5.1)

where

\[
R_c(\tau_k) = \sqrt{\hat{x}(\tau_k)^2 + \hat{y}(\tau_k)^2 + \hat{z}^2 - r_0}
\]  
(5.2)

is the range correction for the pulse at \( \tau_k \). Above, \( \hat{x} \) and \( \hat{y} \) are provided by the position track (\( \hat{\mathbf{x}} \)) and \( \hat{z} \) is simply the constant radar height. The variable \( r_0 \) is the range where the dismount returns are to be centered after alignment.

Entropy-based range alignment [30] involves reducing the entropy of the range profile through an unconstrained optimization procedure. A polynomial range error function is optimized to produce the lowest entropy, and thereby concentrates energy into a minimal number of range bins. The range error is considered constant within each pulse, and is therefore a function of slow-time, \( \tau_k \). The range error polynomial
is given by

\[ r_e(k) = r^T k \]  \hspace{1cm} (5.3)

where

\[ r = [r_n, r_{n-1}, \ldots, r_1, r_0]^T \]  \hspace{1cm} (5.4)

\[ k = [k^n, k^{n-1}, \ldots, k, 1]^T \]  \hspace{1cm} (5.5)

such that \( r \) is the coefficient vector of the \( n \)-degree range error polynomial. The phase shift is a function of \( r \) given by

\[ \Psi(r) = -j4\pi f_i \frac{r^T k}{c} \]  \hspace{1cm} (5.6)

The coefficient vector that minimizes the entropy is calculated by

\[ r_c = \arg \min_r H \left( \sum_{k=0}^{N_r-1} \left| \text{IDFT}_i \{ S_{ra}(f_i, \tau_k) e^{\Psi(r)} \} \right|^2 \right) \]  \hspace{1cm} (5.7)

where \( H \) is the entropy function is given by

\[ H(X) = \sum_{i=1}^{N_x} p_i \log p_i \quad , \quad p_i = \frac{|x_i|}{||X||} \]  \hspace{1cm} (5.8)

and \( \text{IDFT}_i \) is the IDFT over the frequency dimension (range compression). After range compression, the data is summed along the slow-time dimension, producing a one-dimensional range profile. Radar returns in fewer range bins result in higher peaks, less spread, and therefore lower entropy in this range profile. The minimization is performed using the MATLAB\textsuperscript{TM} Optimization Toolbox for unconstrained minimization. It is difficult to anticipate the nature of the range error, and therefore is difficult to choose the optimal order of the range error polynomial over which to min-
imize the entropy. One solution is to iterate over polynomials of different order, each
iteration refining the range-time image through the entropy-based range alignment.
The entropy-based range aligned data is given by

\[ s_{\text{era}}(r_n, \tau_k) = \text{IDFT}_i \{ S_{\text{ra}}(f_i, \tau_k) e^{i\psi_{\mathcal{E}}} \} \]  \hspace{1cm} (5.9)

### 5.2.2 RF Activity Map

The RF activity map is produced by Doppler processing with the STFT. The
STFT provides Doppler frequency information localized in time. A version of the
discrete STFT is given by

\[ S_{\text{TF}}(n, q, k) = \sum_{i=0}^{N_w-1} s_{\text{era}} \left( n, k + i - \frac{N_w}{2} \right) g(i) \exp \left[ -j2\pi \frac{iq}{N_w} \right] \]  \hspace{1cm} (5.10)

where \( n \) denotes the range index \( (r_n) \), \( q \) denotes the Doppler frequency index \( (f_q) \), and
\( k \) denotes the pulse index \( (\tau_k) \). The function \( g(i) \) is a discrete windowing function
(such as Hamming or Gaussian) of length \( N_w \) (assumed to be even). The STFT
amounts to sliding a window across the slow-time dimension of range-time data and
performing the DFT within the window. Here, \( k \) indicates the center pulse of the
window, so the range-Doppler image obtained at a given pulse, \( k \), is the integration
of all Doppler returns within the window surrounding pulse \( k \). It is important to note
that clutter suppression is performed by coherent subtraction in the activity map
formation process. MTI filtering has the consequence of suppressing arm and leg
motions near zero-velocity, therefore destroying much of the important information
provided by the activity map.

A tradeoff between time and frequency resolution exists when choosing the window
length. A shorter window length provides better time resolution, but only a small
number of frequency samples to cover the Doppler bandwidth. Increasing the window
length results in more frequency samples and therefore finer frequency resolution, but the frequency samples cannot be localized in time within the window, so time resolution suffers. In the case of a dismount, the arms and legs produce highly dynamic Doppler shifts. A long window (∼1 sec) may result in a range-Doppler map with Doppler shifts from an entire gait cycle. The time within the window that each of the Doppler shifts occurred cannot be obtained, so time resolution is poor for a long window length.

The goal of the activity map is to capture all of the Doppler information for a single dismount gait cycle. Here, we will sacrifice time resolution to produce a single image that captures dynamic movement information for an entire gait cycle. The ideal method to produce this map is to perform Doppler processing over a single window covering the gait cycle. The results of this direct method are not desirable because of small phase errors in the data or arising from the range alignment process. The coherent integration of an entire gait cycle results in a heavily distorted image. The solution is to non-coherently integrate Doppler returns from shorter window lengths. In this way, the activity map is given by

\[ A(r_n, f_q) = \sum_{k \in \kappa} |S_{TF}(r_n, f_q, \tau_k)|^2 \]  

(5.11)

where \( \kappa \) is a set of evenly spaced pulse indexes that cover a single gait cycle.

Figure 5.1 shows activity maps of two different dismounts undergoing a standard walking motion toward the radar. The activity maps were formed by integrating over a single gait cycle with window lengths of 256 pulses and an overlap of 200 pulses. The motion of the arms and legs and their radial velocities can clearly be seen. The second dismount appears to walk with higher velocity arm and leg motions relative to the torso than the first dismount. Note that the arms swing past actual zero-velocity, indicating they move away from the radar on their back-swing. Generally, the feet come in contact with the ground and do not move away from the radar when
Figure 5.1: **RF Activity Maps.** Spatio-temporal activity maps for two different dismounts. Range and Doppler have been normalized in space to the center of the dismount. Actual zero-velocity in relation to the dismounts is shown. Note that some clutter returns exist on the actual zero-velocity line since coherent subtraction was used in place of MTI filtering. Because of the height of the radar, feet motion occurs at a farther range than the arm motion.

The dismount is walking toward the radar; however, the second dismount has some negative velocity leg motion that may be attributed to micro-motion on a smaller scale than the leg as a whole, such as the pivot between the ankle and foot. Activity maps like these may be useful in the characterization and classification of dismounts.
CHAPTER 6
THREE-DIMENSIONAL RECONSTRUCTION

Here, a three-dimensional map is created using the fusion of RF and EO data. Tomographic algorithms exist in RF and EO domains; however, they each require a large sweep angle of measurements. By combining the output from the two sensors, a 3-D map of a walking dismount is formed; however, the micro-motion of the dismount will cause smearing in certain areas of the map. EO tomosynthesis and RF back-projection are similar algorithms for each of their sensor domains. Each uses measurements from multiple look angles to create a three-dimensional estimate of the object of interest.

6.1 RF Back-projection

The ideal image formation process for ISAR imaging involves matched filtering of the phase history data. A volume in Cartesian space is chosen over which to image. Continue to let \( \hat{x}(\tau_k), \hat{y}(\tau_k), \) and \( \hat{z}(\tau_k) \) represent the estimated dismount track in world coordinates (origin at EO camera center). Let \( \tilde{x}, \tilde{y}, \) and \( \tilde{z} \) represent the coordinates with the origin at the center of the 3-D image volume. Assume the coordinate axes of the image volume are oriented in the same way as the tracking (world) axes such that conversion between the two coordinate systems is purely translational. The origin of the image volume is maintained as the center of the dismount returns, so it varies with respect to slow-time. At a point \((\tilde{x}, \tilde{y}, \tilde{z})\) in the volume, the matched filter response
is

\[
P(\tilde{x}, \tilde{y}, \tilde{z}) = \frac{1}{N_f N_r} \sum_{i=0}^{N_f-1} \sum_{k=0}^{N_r-1} S(f_i, \tau_k) e^{j4\pi f_i \frac{R_{xyz}(\tau_k)}{c}}
\] (6.1)

where \( R_{xyz}(\tau_k) \) is the range to the imaging point, \((x, y, z)\) at pulse \( \tau_k \). The motion of the dismount must be compensated for when determining the range to the image center. The range from the radar to the image center is simply the tracked range of the dismount given by

\[
R_{000}(\tau_k) = \sqrt{\tilde{x}(\tau_k)^2 + \tilde{y}(\tau_k)^2 + \tilde{z}(\tau_k)^2}
\] (6.2)

Adjusting (6.2) with the imaging point in the image coordinates gives the range from the radar to the image point, \((\tilde{x}, \tilde{y}, \tilde{z})\),

\[
R_{xyz}(\tau_k) = \sqrt{\tilde{x}(\tau_k)^2 + \tilde{y}(\tau_k)^2 + \tilde{z}(\tau_k)^2}
\] (6.3)

Substituting (2.9) into (6.1) and substituting the time delay \( T_{xyx} = 2R_{xyz}/c \), the matched filter output is given by

\[
P(\tilde{x}, \tilde{y}, \tilde{z}) = \frac{1}{N_f N_r} \sum_{i=0}^{N_f-1} \sum_{k=0}^{N_r-1} S(f_i, \tau_k) e^{j2\pi \left(f_0 + \frac{iB}{N_f}\right) T_{xyz}(\tau_k)}
\]

\[
= \frac{1}{N_f N_r} \sum_{k=0}^{N_r-1} e^{j2\pi f_0 T_{xyz}(\tau_k)} \sum_{i=0}^{N_f-1} S(f_i, \tau_k) e^{j2\pi \frac{iB}{N_f} T_{xyz}(\tau_k)}
\]

\[
= \frac{1}{N_r} \sum_{k=0}^{N_r-1} e^{j2\pi f_0 T_{xyz}(\tau_k)} U\left(T_{xyz}(\tau_k)\right)
\] (6.4)

where

\[
U(T_{xyz}(\tau_k)) = \frac{1}{N_f} \sum_{i=0}^{N_f-1} S(f_i, \tau_k) \exp\left[j2\pi \frac{iB}{N_f} T_{xyz}(\tau_k)\right]
\] (6.5)
To allow for digital processing, the time delay, $T_{xyz}(\tau_k)$, can be sampled into discrete values according to

$$T_{xyz}(\tau_k) = \frac{n N_f}{B \mathcal{N}_R} \quad (6.6)$$

where $\mathcal{N}_R$ is the number of samples in range. Then, $U$ becomes a function of $n$ and is given by

$$U(n, \tau_k) = \frac{1}{N_f} \sum_{i=0}^{N_f-1} S(f_i, \tau_k) \exp \left[ j2\pi \frac{in}{\mathcal{N}_R} \right], \quad (6.7)$$

which is the IDFT of the phase history data in the fast-time dimension. The back-projection algorithm can therefore be performed by (6.4) using the IFFT of the phase history in fast-time multiplied by the focusing phase, $\exp[j2\pi f_0T_{xyz}(\tau_k)]$.

The image $P(\tilde{x}, \tilde{y}, \tilde{z})$ can be computed at any point in 3-D space; however, a volume surrounding the dismount is sampled with discrete coordinates, $\tilde{x}_i, \tilde{y}_j, \tilde{z}_l$. $U(n, \tau_k)$ provides data sampled along the range dimension; however, $P(\tilde{x}_i, \tilde{y}_j, \tilde{z}_l)$ is sampled in Cartesian coordinates. Linear interpolation is used to obtain the samples for $U(T_{xyz}(\tau_k))$ from $U(n, \tau_k)$. Because of the use of simple linear interpolation, range must be oversampled to minimize errors introduced by interpolation and to maintain phase coherence. In this case, range is oversampled by a factor of 64, giving $\mathcal{N}_R = 64 \mathcal{N}_f$. Oversampling of $U(n)$ is implemented as zero-padding when performing the FFT of the phase history data.

### 6.2 EO Tomosynthesis

Tomosynthesis[31] is analogous to RF back-projection in the EO domain. Tomosynthesis differs from traditional computed tomography (CT) in a few ways [32]. CT methods usually reconstruct a volume with a large number of projections, where
Figure 6.1: **RF Back-projection Results.** Radar returns for 1024 pulses are back-projected into 3-D space. Since the dismount sweeps through a small angular range, azimuth resolution is not noticeably improved. Results are shown on a dB scale with darker areas representing stronger returns.

tomosynthesis uses a relatively small number of projections. CT methods also use a wide range of aspect angles (sometimes full 360°), while tomosynthesis uses projections over a small range of aspect angles. For these reasons, CT is well suited for cooperative subjects and controlled sensor configuration applications, such as medical imaging, while tomosynthesis is well suited for non-cooperative surveillance applications where there is insufficient information for CT methods.

Tomosynthesis amounts to back-projecting the image point, \( \hat{\gamma} \), to the line, \( \gamma(\lambda) \) in 3-space. Back-projection is performed for each pixel in a frame, giving a set of rays that represent possible locations for the 3-D point that was projected into each image point. EO back-projection is theoretically defined by

\[
\gamma(\lambda) = \mathbf{P}^+ \hat{\gamma} + \lambda \mathbf{c}
\]  

(6.8)
where \( P^+ \) is the pseudo-inverse of the camera matrix, \( P \). \( P^+ \hat{\gamma} \) converts the image point, \( \hat{\gamma} \), from image coordinates to the 3-D world coordinate system while \( \lambda c \) forms the line between \( \gamma \) and \( \hat{\gamma} \) as \( \lambda \) varies.

Tomosynthesis involves a camera moving around a static object and capturing views from different aspect angles. In this case, inverse tomosynthesis is used, where the camera is static and the image volume is moving through the scene. Similar to RF back-projection imaging, the 3-D image coordinate center, \((\tilde{x}, \tilde{y}, \tilde{z}) = (0, 0, 0)\), is held at the center of the dismount over time. Therefore, the camera coordinate system remains static, but the 3-D image coordinate system moves along with the dismount. The movement of the 3-D image coordinate system is purely translational so as to preserve the back-projection angles. The 3-D image coordinate axis maintains the same orientation as the world coordinate system in the RF case. The camera orientation or position never changes, but as the dismount moves through the frame, the azimuth angles of the rays formed by the back-projection of the pixels on the dismount change. Since the camera is static and the dismount is moving along the ground plane, it is assumed that the projected rays are changing only in azimuth angle as the dismount moves through the scene. This assumption simplifies the back-projection process.

Equation (2.7) shows that the azimuth and elevation angles of back-projected rays can be determined with knowledge of the camera field of view. For each EO frame, the dismount mask, \( G_k(u, v) \), given by (3.10), is back-projected to form a set of rays in 3-D space. These rays are completely defined by their azimuth (\( \phi \)) and elevation (\( \theta \)) angles and the camera center, through which each ray passes. The camera center occurs at the origin of the world coordinate system, and the angles are determined by equation (2.7). The dismount track is used to locate the center of the 3-D image
volume to be sampled. Let $\mathbf{w}$ represent the 3-D image origin in world space given by

$$
\mathbf{w} = \begin{bmatrix}
\hat{x}(k) \\
\hat{y}(k) \\
0
\end{bmatrix} 
$$

(6.9)

Within the 3-D image volume, the back-projected rays are sampled on the same discrete Cartesian grid used in RF back-projection, $(\tilde{x}_i, \tilde{y}_j, \tilde{z}_l)$. Each of the rays is first sampled along the ray at the grid values for $\tilde{y}$. The sample locations are given by.

$$
\tilde{x} = (\hat{y}_k + \tilde{y}) \arctan[\phi(u)] - \hat{x}_k 
$$

(6.10a)

$$
\tilde{z} = \sqrt{(\hat{x}_k + \tilde{y})^2 + (\hat{y}_k + \tilde{y})^2} \arctan[\theta(v)] 
$$

(6.10b)

This gives a set of $(\tilde{x}, \tilde{y}, \tilde{z})$ samples in the 3-D image volume. Each of the samples corresponds to a binary value that indicates whether or not the ray on which the sample lies intersected the dismount mask. The sampled rays are interpolated onto the 3-D image grid using linear interpolation.

Back-projection and interpolation is performed for a sequence of frames, over which the aspect angle changes, providing different views of the dismount. The 3-D maps obtained from each frame are on the same coordinate system in relation to the dismount, so they can simply be integrated. The integration has the effect of enforcing voxels that lie on the dismount. Let $T(\tilde{x}_i, \tilde{y}_j, \tilde{z}_l)$ represent the sampled tomosynthesis output given by

$$
T(\tilde{x}_i, \tilde{y}_j, \tilde{z}_l) = \sum_k T_k(\tilde{x}_i, \tilde{y}_j, \tilde{z}_l) 
$$

(6.11)

where $T_k$ is the interpolated back-projection for a single frame.

As in RF back-projection, dismount micro-motion causes blurring in certain areas
Figure 6.2: **EO Tomosynthesis Results.** Five EO frames containing the same dismount pose are back-projected into 3-D space. Note that some range resolution is achieved. This can be seen as the right and left legs appear to be separated in range. Results are shown on an intensity scale.

since the arms and legs are not integrated into the same location from frame to frame. By using only EO frames from a single dismount pose, this can be minimized at the expense of some cross-range resolution. Figure 6.2 shows the tomosynthesis output using one frame per gait cycle. It is clear that the dismount only sweeps through enough look angles to remove a fraction of the range ambiguity.

### 6.3 RF and EO Fusion

The RF and EO 3-D images are fused to form a single 3-D map of the dismount. A common coordinate system was used in the formation of the RF and EO images, so the images are already registered in 3-D space. Voxels on the dismount are those in the intersection of the 3-D maps obtained from the RF and EO imaging. The RF intensity and EO images are normalized to values from 0 to 1. The final 3-D image is
Figure 6.3: **Fused 3-D image.** EO and RF back-projected points are fused to form a single 3-D image. The image clearly shows the shape of the dismount and the extension of the carried object. Results are shown in dB with higher values corresponding to darker points.

formed by the voxel-wise multiplication of the normalized RF and EO images. Much of the ambiguity that exists in the RF and EO images alone is removed in the fused image.
CHAPTER 7
CONCLUSION

The simultaneous use of RF and EO sensors for dismount characterization and imaging provides some advantages over the use of each sensor individually. By exploiting the range information from an RF sensor and the azimuth and elevation information from an EO sensor, the ambiguities of each of the sensors are diminished. RF and EO processing techniques were presented as well as a fused tracking algorithm. 3-D images of dismounts were formed using the fusion of tomographic techniques in the RF and EO domains. Imaging dismounts presents a number of problems because of their articulated motion. Standard tomographic techniques perform relatively well, but there is still a need to better compensate for dismount micro-motion when imaging. Future research in this area will involve refining the 3-D imaging techniques to better handle the human micro-motion and form more accurate images for each time instant. An accurate 3-D image of a dismount will provide information on dismount structure and motion.
REFERENCES


