OVERCOMING POSE LIMITATIONS OF A SKIN-CUED HISTOGRAMS OF ORIENTED GRADIENTS DISMOUNT DETECTOR THROUGH CONTEXTUAL USE OF SKIN ISLANDS AND MULTIPLE SUPPORT VECTOR MACHINES

THESIS

Jonathon Climer, Second Lieutenant, USAF

AFIT/GE/ENG/11-05

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GE/ENG/11-05

OVERCOMING POSE LIMITATIONS OF A SKIN-CUED HISTOGRAMS OF ORIENTED GRADIENTS DISMOUNT DETECTOR THROUGH CONTEXTUAL USE OF SKIN ISLANDS AND MULTIPLE SUPPORT VECTOR MACHINES

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Electrical Engineering

Jonathon Climer, BSEE

Second Lieutenant, USAF

March 2011

AFIT/GE/ENG/11-05

OVERCOMING POSE LIMITATIONS OF A SKIN-CUED HISTOGRAMS OF ORIENTED GRADIENTS DISMOUNT DETECTOR THROUGH CONTEXTUAL USE OF SKIN ISLANDS AND MULTIPLE SUPPORT VECTOR MACHINES

Jonathon Climer, BSEE
Second Lieutenant, USAF

Approved:

Maj Michael J. Mendenhall, PhD (Chairman)	9 MAR 2011 Date
Dr Gilbert L. Peterson (Member)	8 MAR 2011 Date
LtCol (Ret) Juan R. Vasquez, PhD (Member)	9 Mar 2011 Date

AFIT/GE/ENG/11-05

Abstract

In light of an increased need for the U.S. military to identify and combat non-conventional adversaries, improved situational awareness with efforts focusing on individuals, small groups, and their intent is crucial. Dismount detection offers powerful abilities to automatically detect and potentially track individuals in imagery, fulfilling a key role in combat and non-combat scenarios, with applications for base defense and search and rescue.

This research examines a premier histograms of oriented gradients (HOG) dismount detector that cues based off of the presence of human skin (to limit false detections and to reduce the search space complexity). While this skin cued detector performs well for typical head on standing poses, there are anticipated limitations in the detection of dismounts over a full range of motion and camera angles. This thesis focuses on extending the robustness of this promising skin cued detector system while improving the suppression of false detections.

A novel visualization method is developed in this thesis to analyze the impact that articulations in dismount pose and camera aspect angle have on HOG features and eventual detections. Insights from these relationships are used to identify the breaking points in the current system's ability to detect non-upright poses from a variety of camera angles. Improvements to detector performance are made by further leveraging available skin information and anthropometry, overall reducing false detections by an additional order of magnitude. A framework is formed to augment the existing detector with supplemental support vector machines (SVMs) to detect additional pose groups. The "multi-SVM detector", while being trained on computer simulated data, detects a broad spectrum of dismount poses in live imagery, offering superior performance to the baseline skin-cued detector.

The multi-SVM detector showcases a 7-fold increase detection probability when applied to challenging crouching poses over extensive camera angle ranges (with elevation angles up to 50°). These dramatic improvements clearly demonstrate the viability of a multi-SVM approach, which can be extended to include increased pose configurations.

Acknowledgements

My success at AFIT has hardly been a product of my efforts alone. I owe a large debt of gratitude to the sustaining power and tender mercies of God. I could not have continued at AFIT without the patience and support of my brilliant wife and my son; their love, their encouragement, and their reminders for me to eat meals have gotten me through it all.

I'd like to thank Maj. Mendenhall for his vigilance, attention, and for his interest in my learning and success. I also owe a big thanks to Matt Maier and Brad Koch for their help with numerous data collections, rides, general advice, and for being great friends.

Jonathon Climer

Table of Contents

Page

List of Figures

Figure Page

List of Tables

OVERCOMING POSE LIMITATIONS OF A SKIN-CUED HISTOGRAMS OF ORIENTED GRADIENTS DISMOUNT DETECTOR THROUGH CONTEXTUAL USE OF SKIN ISLANDS AND MULTIPLE SUPPORT VECTOR MACHINES

1. Introduction

In light of an increased need to identify and combat non-conventional adversaries, the United States Air Force has listed one of its top priorities as "getting more intelligence, surveillance and reconnaissance to the war zone" [28]. The issue of identifying individuals (dismounts) and their intent is of particular interest and attracting attention from the Joint Improvised Explosive Device Defeat Organization (JIEDDO) as well as many other military organizations [33]. Dismount detection offers powerful abilities to automatically detect and potentially track individuals in imagery, fulfilling a key role in combat and non-combat scenarios, with applications for surveillance and search and rescue. The problem of detecting the presence of individuals is also approached by many different commercial groups. Some of these include the automotive industry, working to develop pedestrian avoidance systems [24] [34]. Most of these systems utilize standard cameras operating in the visible range of the electromagnetic spectrum from a vehicle's point of view.

One technique for dismount detection was demonstrated in [5] which incorporates several short wave infrared wavelengths in addition to the visible spectra in order to identify human skin [29] and selectively scan the image for the presence of dismounts. A histograms of oriented gradients (HOG) feature is then used to describe scaled image subsections, which are then compared against pedestrian training data from [13] to establish detections. This work is very promising as the selective skin cuing drastically reduces the necessary number of search windows and suppresses a wide range of false alarms.

The dismount detection system in [5] performs very well for imaged subjects who are standing upright with both feet on the ground. However, it is anticipated that the HOG

feature descriptors are highly dependent on the dismount pose and the angle of the acquisition camera. Consequently, it is expected that limitations exist on the humans' range of motion or stance that still yields positive detections. It is essential to understand the extent of these limitations and design around them to offer a more versatile dismount detector operating with the same skin cuing abilities.

1.1 Problem Statement

Human beings experience a wide range of physical movement as they perform daily activities. Even as someone walks down the street they may sway as they walk, wave to a friend, or crouch down and tie a shoe. However, a given dismount detector may fail in detecting any of these articulations in pose due to the limitations in their trainings data. In order to create an effective dismount detector, it is essential to understand the effects typical human movements have on the detection process. First, it is difficult to predict how changes in pose affect HOG features. It is quite challenging to even conceptualize a HOG feature even if presented with its values. In this thesis, several methods are explored to represent the HOG feature in a more understandable and intuitive fashion. Second, it is essential to understand the limitations on human subject motion in a given dismount detection system. These bounds may be found by capturing images displaying common ranges of motion and computing the corresponding prediction strength from the classification tools. These prediction strengths can then be compared against a detection threshold, to determine the span of acceptable motion.

In order to extend the use of the Brooks [5] dismount detection system, additional training data is accumulated and grouped according to pose and camera angle. The additional dismount configurations build a support base to detect a greater range of human motion in imagery and support detections from a variety of camera angles.

1.2 Scope

The scope of this thesis effort must be limited in order to accomplish the previously stated goals. Accordingly, the focus centers around:

- Providing initial improvements to the skin-cued dismount detector in [5] using available skin island information;
- Examining how small articulations in pose manifest in the HOG features and subsequent predictions;
- Identifying the limitation in support of detectors over several typical ranges of motion and camera angles; and
- Comparing the performance of the skin-cued dismount detector in [5] with modified versions developed in this thesis that extend detection coverage and suppress additional false detections.

Initial improvements to the skin-cued detector are presented in Section 3.1 and Section 4.3. These improvements leverage anthropometry and detected regions of skin to to compensate for imager imperfections, detect broader ranges of dismount poses, and suppress false alarms. The effect of articulations in pose and camera aspect angle on HOG features are visualized through a method presented in Section 3.2 to correlate spatial location and gradient orientation angle. This visualization method highlights significant impact on HOG features due to small changes in imagery with results further discussed in Section 3.3. The changes in HOG features and predictions are analyzed to identify "holes" in the detection coverage that are identified in Section 3.3.3 and Section 3.4. The skin cued dismount detector is augmented to detect dismounts in additional poses through the use of supplemental SVMs (training is discussed in Section 3.5 and Section 4.4). Performance comparison results are presented in Section 4.5 and Section 4.6.

1.3 Document Organization

Chapter II of this document provides useful background information and related work pertaining to dismount detection stages, skin detection, and various notations and conventions.

Chapter III details the methodology developed in this thesis and specifically focuses on techniques for HOG visualization, analyzing the invariance of HOG vectors and their relationship to SVM weights, identifying the limitations in the Brooks [5] dismount detector, and methods for overcoming detection limitations.

Chapter IV shows experimental results and their related analysis.

Chapter V emphasizes the impact and contributions of this thesis efforts and recommendations for future work.

2. Background

This chapter provides an overview of the dismount detection problem as well as relevant previous work and useful supplemental information to aid in the understanding of this thesis document. First, this chapter documents terminology and notational conventions included in this document. Second, a background of various dismount detector techniques is provided. Third, the five stages of the dismount detection methodology developed by Brooks [5] is introduced (herein referred to as the Brooks [5] detector). An overview of methods common to each detection stage are then detailed. Finally, additional processes are described, providing background for later discussion on dismount detection.

2.1 Notation and Terminology Conventions

2.1.1 Boldface and Uppercase Notation.

This document makes use of boldface and uppercase notation in order to distinguish between scalars, vectors, two-dimensional matrices (henceforth *matrices*), and three dimensional matrices (henceforth *cubes*). Lowercase variables appearing in normal text color are scalars (e.g., s is a scalar). Vectors are represented by a lowercase letter in bold (e.g., $\mathbf{v}$ is a vector), whereas matrices are uppercase letters in bold (e.g., $\mathbf{M}$ is a matrix). Finally, cubes are represented by an underlined uppercase variable in bold (e.g., $\underline{\mathbf{C}}$ is a cube).

2.1.2 Representing Multi-Dimensional Structures.

Standard linear algebra conventions are applied when discussing indices and dimensionality of matrices, images, and other multi-dimensional structures. Henceforth, the first dimension listed pertains to the structure's height, followed by the width, then depth. For example Fig. 1 shows an $l \times m \times n$ structure, with indexing beginning at the top, left, and front of the displayed cube.

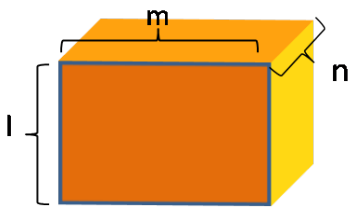


Figure 1. Standard linear algebra conventions refer to this object as an $l \times m \times n$ structure with indexing beginning at the top, left, and front.

2.1.3 Use of Spherical Coordinate System in Plots.

Hemispherical plots are useful in describing various camera angles and are referenced using a spherical coordinate system with values (r, θ, ϕ). In this thesis, all values of r remain constant, so this coordinate will frequently be omitted. The angle ϕ (as seen in Fig. 2) is the azimuth angle, and the angle θ is the angle of elevation. In plots such as Fig. 2, the center point P is considered to have an azimuth and angle of elevation equaling 0°, whereas a point on the right half of the hemisphere (such as Q) has positive ϕ and θ values.

2.2 Dismount Detection Techniques

Numerous approaches exist for detecting dismounts within imagery. A primary category of detection techniques adopt a "whole body" detection methodology, representing the entirety of a dismount by a feature descriptor. Common spatial features include Haar wavelets [27] [36], dense edge orientation encodings such as HOG [9] [10], [34], and sparser edge orientation encodings such as the scale-invariant feature transform (SIFT) [26]. The spatial features describing test image patches are then classified based off of a collection of prototype features. Other whole body detectors represent dismounts as a function of body geometry derived from spatial cues [15] [32]. Another detection technique stitches together expert body part detectors to create an ontological representation of dismounts [16] [37] [38]. While the fusion of parts based detectors has fewer dependencies on pose, correct cuing and association of body parts in a cluttered environment presents significant challenges for single frame dismount detection.

While there are a variety of systems in use for the the detection of dismounts in imagery,

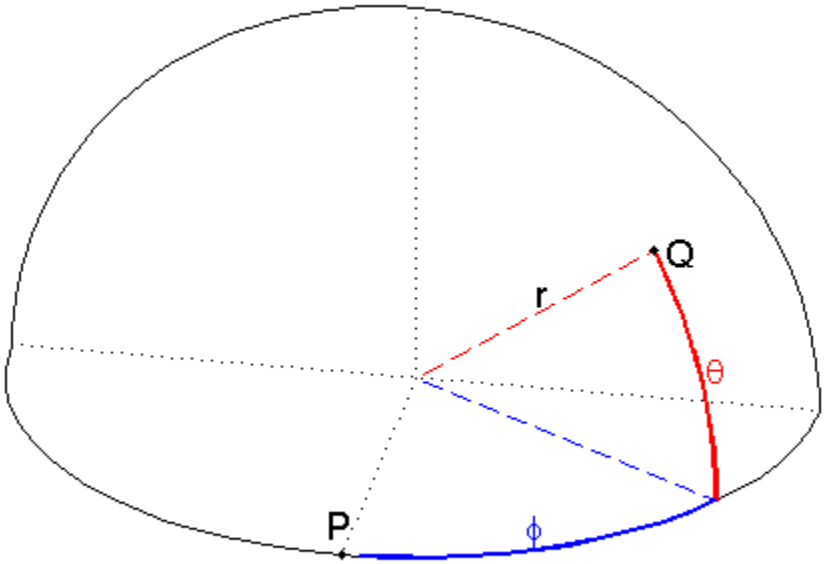


Figure 2. A spherical coordinate system is used in hemispherical plots presented throughout this thesis. As such r represents the radius from the origin to an arbitrary point on the sphere, Q with ϕ showing the azimuth angle and θ being the angle of elevation. The point P represents the front of the front of the hemispherical plot with ϕ and θ values of 0.

this thesis effort focuses on a promising design by Brooks [5] that employs multispectral skin detection to reduce the search space in a HOG-based dismount detector and is an extension to that described in [13].

2.3 Brooks Dismount Detection Structure

The Brooks [5] detector can be generalized into a five stage model. First, is the acquisition of raw image cube. Second, beneficial image pre-processing steps are applied. Third, regions of interest or sub-portions of the original image are identified for further analysis. Fourth, the image data from the region of interest (search window) is transformed into a HOG feature representation as in [13]. Last, once the image data is represented in HOG feature space, it is separated with a hyperplane using a support vector machine as defined in [13] for classification.

Figure 3. The five stage model shown represents the dismount detector structure described in [5].

2.3.1 Image Acquisition Methods.

Many different sources of image data are potentially useful in detecting dismounts. Cameras operating in the visible region of the electromagnetic (EM) spectrum (VIS) are frequently used because of their high image quality and low cost. Both monochrome (gray-scale) and red-green-blue (RGB) cameras are commonplace for the detection of dismounts [9][10][34] . Another common alternative is far infrared imagery which offers strong contrast between dismounts and backgrounds, as this type of remote sensing is strongly characterized by thermal emissions[14]. Near infrared (NIR) and short-wave infrared (SWIR) imagery present another potential option for dismount detection. While they do not offer the benefits of detecting thermal emissions, they have been proven useful in skin [21] and face detection applications [11]. Hyperspectral Imagery, often used in geological and biological surveys, was a source of inspiration for the Brooks [5] detector due to the richness of spectral data that it offers and its potential use in feature aided tracking.

The Brooks [5] detector adopts a multispectral approach from [29], fusing RGB imagery with several SWIR cameras. These selected portions of the EM spectra are used in concert to detect the presence of humans in a given environment based off of the spectral properties of their skin.

2.3.2 Pre-processing Steps.

While the resulting images from the RGB and SWIR cameras offer useful content, it is not a true representation of the reflectance of the objects in the image. Despite the diligence of the data collection, imperfections are preserved in the obtained image cubes due to atmospheric, lighting, and other distortions. There are complex atmospheric radiometric transfer models, such as MODTRAN [23], that model the environmental factors at a given

time and location. However, instead of using computationally intensive modeling to account for such distortions, the Brooks [5] detector system applies linear regression to compute estimated reflectance, $\hat{\rho}_\lambda$, based off of measured intensities of in scene objects with known reflectance. This is known as the empirical line method (ELM) [12]. With the assumptions of equal illumination of all image pixels and a linear relationship between image intensity and reflectance, ELM is applied to estimate reflectance as:

$$\hat{a}_\lambda = \frac{\mu_\lambda^w - \mu_\lambda^b}{\rho_\lambda^w - \rho_\lambda^b}, \tag{1}$$

$$\hat{b}_\lambda = \frac{\mu_\lambda^b \rho_\lambda^w - \mu_\lambda^w \rho_\lambda^b}{\rho_\lambda^w - \rho_\lambda^b}, \tag{2}$$

$$\hat{\rho}_\lambda = \frac{X_\lambda - \hat{b}_\lambda}{\hat{a}_\lambda}, \tag{3}$$

where X_λ is the raw intensity space representation of the input image at wavelength λ, $\hat{\rho}_\lambda^w$ and $\hat{\rho}_\lambda^b$ are the known reflectances of specific bright and dark in-scene objects, and μ_λ^w and μ_λ^b are the image intensity of the same bright and dark in-scene objects.

2.3.3 Generate Search Windows.

It is common in the literature to generate constant sized image patches from resized image subsections. The image patches are frequently obtained by applying a sliding window of set proportions at every possible scale and position supported by the picture [13] [38]. While such a method is relatively easy to implement, it requires a great deal of computational time, much of which is wasted in areas of the image that do not contain the object of interest. Many alternative methods only generate search windows around objects of interest. The Brooks detector uses indicators of human skin to trigger the formation of search windows.

2.3.3.1 Skin Detection Algorithms.

Two supporting methods are used to indicate the presence of human skin in an image: the normalized difference skin index (NDSI) and the normalized difference green red index

(NDGRI). Both of these methods leverage spectral properties of skin in order to calculate values useful in skin detection.

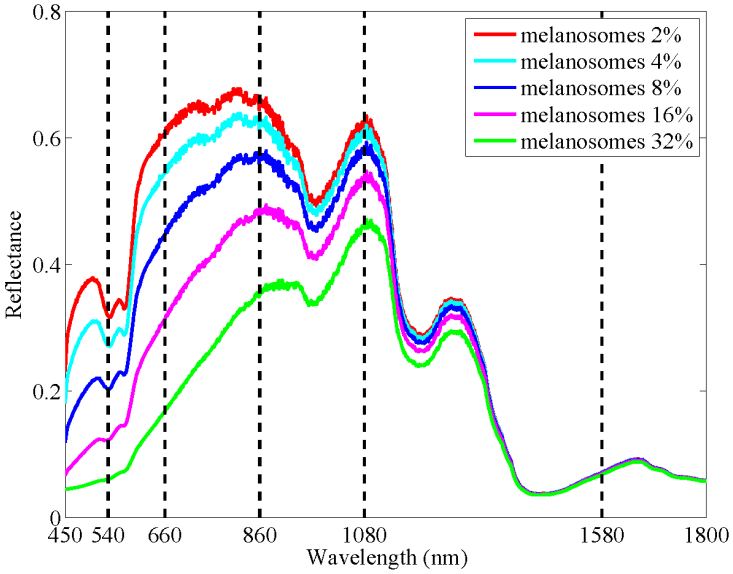


Figure 4. Model Generated skin data from [29] [30] shows the reflectance of five different skin tones with melanosomes ranging from 2% to 32%. A distinctive drop off in absorption between the 1080nm and 1580nm range can be observed regardless of skin tone. A noticeable jump can also be observed between the 540nm and 860nm bands, corresponding to the green and red components of skin.

NDSI takes advantage of a large drop off in skin reflectance in the range of 1080 and 1580nm due to water absorption, as seen in Fig. 4. The distinctive regional drop off holds true regardless of skin tone. The NDSI value is calculated as

$$\gamma = \frac{\rho_{\lambda_1} - \rho_{\lambda_2}}{\rho_{\lambda_1} + \rho_{\lambda_2}} \tag{4}$$

where ρ_λ is the reflectance measurement at wavelength λ ($\lambda_1 = 1080nm$ and $\lambda_2 = 1580nm$). However, the use of NDSI values alone may be insufficient to accurately detect human skin as there are several other materials (including certain kinds of vegetation) that possess similar water-absorption characteristics. In order to mitigate the number of false alarms from vegetation, NDGRI is introduced, which compares the amount of red versus green in each pixel (from an RGB camera) against a threshold. Similar to Eqn. 5, the NDSI is

computed as

$$\beta = \frac{\rho_{\lambda_1} - \rho_{\lambda_2}}{\rho_{\lambda_1} + \rho_{\lambda_2}} \tag{5}$$

where λ_1 and λ_2 are 540nm (green) and 660nm (red) respectively. Since human skin is more red than green, the addition of NDGRI effectively limits the influence of vegetation.

Given thresholding parameters $[\gamma_{min}, \gamma_{max}, \beta_{min}, \beta_{max}]$, the NDSI and NDGRI values are transformed into intermediate detection masks. Examples of such detection masks are seen in Fig. 5. The intersection of the intermediate detection masks forms the final skin detection mask (also seen in Fig. 5) that is used to trigger the dismount detection system. Contiguous groupings of pixels resulting from the skin detection mask are henceforth referred to as "skin islands", whether or not they actually originate from a dismount. To specify a skin island originating from the head of a dismount, the term "head skin island" is used.

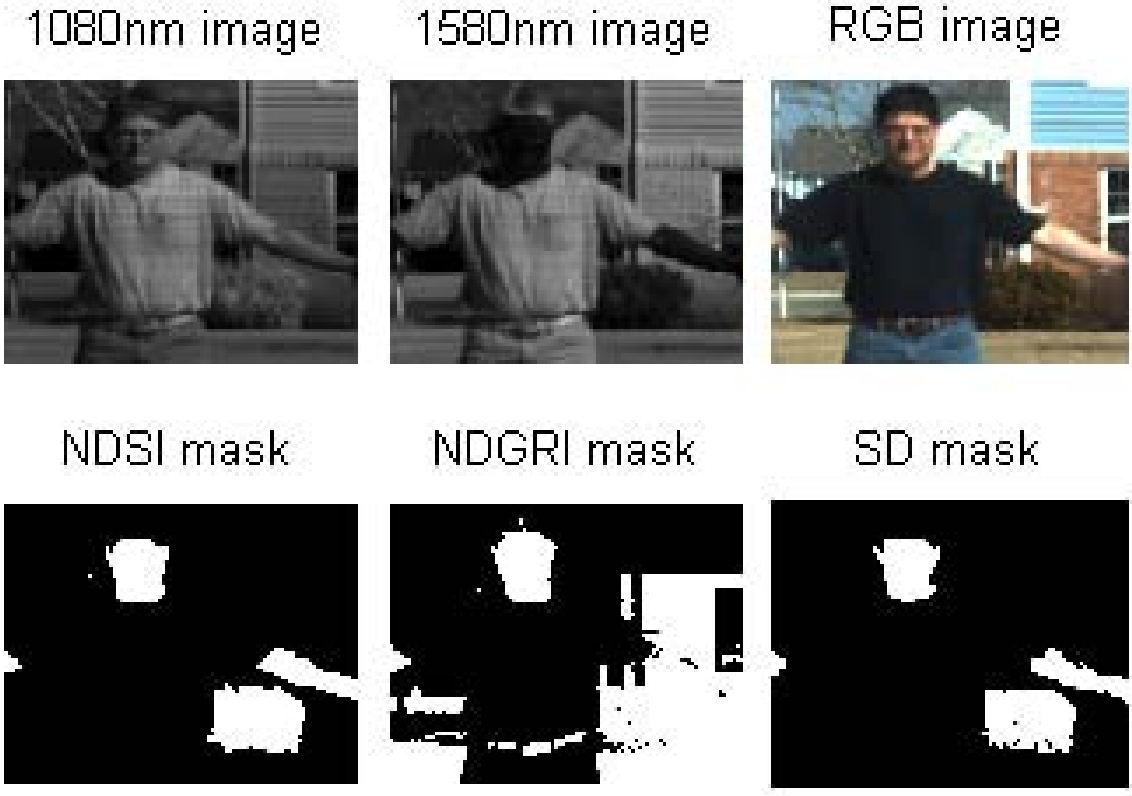


Figure 5. Top row: original imagery from the 1080nm, 1580nm, and RGB cameras. Bottom row: NDSI mask, NDGRI mask, and composite skin detection mask.

2.3.3.2 Extracting Image Chips.

Since applying a sliding window to an input image at every possible scale is computationally demanding [34], the previously mentioned detection masks are used to intelligently select promising image patches and greatly reduce the number of patches to search through. First, skin islands with fewer pixels than a threshold, η_i, are discarded from the detection mask in order to mitigate the effect of stray pixels registering as human skin, allowing only large skin patches to pass through. Each remaining "skin island" is subsequently given a unique label.

The most likely source of skin in an image corresponds to the frontal or side portions of a dismount's head and face. Due to the relatively constant position of the head with respect to the body, the corresponding skin island provides the best intuition as to the location of the dismount's body in the image and is used to form image patches. However, since the detection mask has no way of indicating which part of the body each skin island comes from, they are all considered to derive from the head. Extraneous image patches are later filtered out and discarded (as seen in section 2.3.5).

One method [5] employs to generate search windows (with a height to width ratio of 2 to 1) around dismounts is to center the window horizontally around the centroid of the skin island. The top of this window is then shifted so that it is located a specified number of pixels, $\Delta v \times s$ above the top of the skin island, where Δv is a parameter offset and s is a scale factor. (Experimental work from [5] on a pedestrian dataset recommends a $\Delta v = 15$ wtih a starting value of $s = 0.75$ and thereafter increasing by a factor of 1.1). The image region contained within the search window is then resized to be 96×48 pixels image patch. This procedure produces consistent image patches with potential dismounts that can be compared against a training set of data in later steps.

2.3.4 Generate Histogram of Oriented Gradients (HOG) Features.

Each of the 96×48 pixel image chips identified in the previous step can be represented by a HOG spatial feature. This method of feature generation highlights the directional

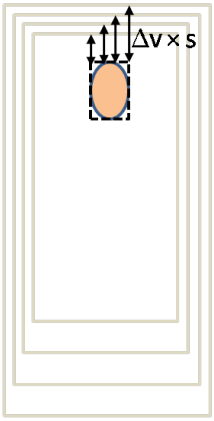


Figure 6. Four potential search windows are outlined by gray rectangles that are spaced $\Delta v \times s$ pixels from the top of the skin island (shown as a tan ellipse) and centered horizontally. Four scale values of s are represented for a constant Δv.

changes in contrast and is very popular in the current literature [2][9][13][25][38][39] and is employed by the Brooks [5] detector. The first step in the formation of a HOG vector is to convolve the image patches by a [-1 0 1] mask in both the x and y direction to calculate the image gradient, emphasizing contrast changes in the image. The magnitude r and direction ($0^\circ \leq \phi \leq 180^\circ$) of the x and y gradients ∇x, ∇y are found for each pixel by

$$r = \sqrt{(\nabla x)^2 + (\nabla y)^2}, \tag{6}$$

$$\phi = arctan \frac{\nabla y}{\nabla x}, \tag{7}$$

where ϕ may be rotated 180° to fall within $[0^\circ, 180^\circ]$ [9]. In order to examine trends in the gradient orientation, the patch is then subdivided into non-overlapping 8 x 8 pixel "cells" (forming a 12 cell structure as seen in Fig. 8(a)). A histogram with nine evenly spaced orientation angle bins centered at 10°, 30°,50°,...,170° is then computed for each cell, assigning "votes" according to the orientation and magnitude of the gradient at each pixel location. Each pixel receives a number of votes equal to its gradient magnitude. These cell votes are divided between the two closest orientation bins, proportional to distance between the gradient's orientation angle and the angle of the bin centers. Fig. 7 demonstrates the voting principle as a generic pixel is shown with a magnitude of 100 and a 25° orientation angle (lying between the 10° and the 30° bin centers). Since the pixel gradient angle is 75%

closer to the 30° bin center than the 10° bin center, 75 of the 100 votes are tallied in the 30° bin, while the 10° bin receives 25 votes. The example pixel votes are then be accumulated with votes from other pixels within the same cell, to form a cell histogram, containing vote tallies for the nine orientation bins. A compilation of all the cell histograms is handled as a $12 \times 6 \times 9$ structure as seen in Fig. 8(b).

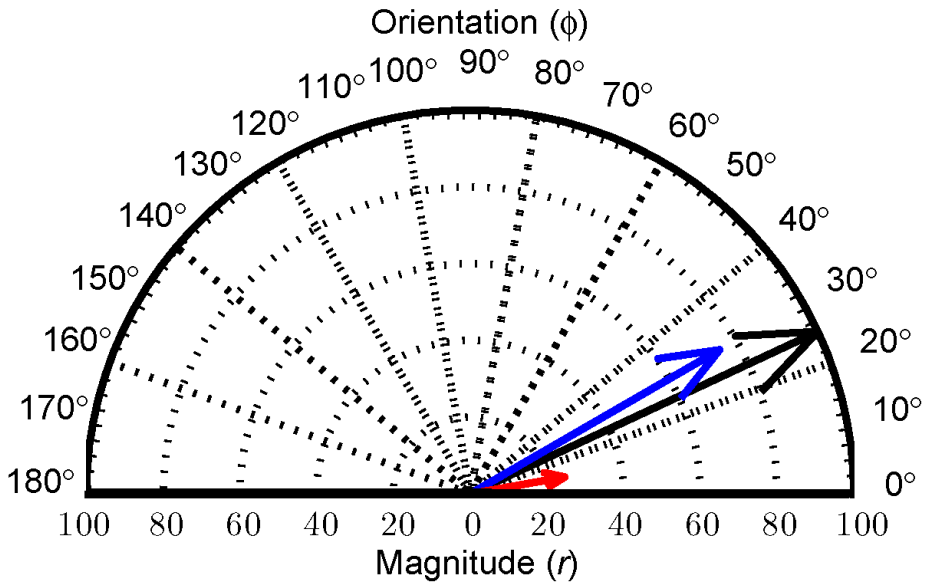


Figure 7. An example pixel gradient is represented by a black arrow with orientation angle ϕ and magnitude r. The contributions of this pixel's votes into the 10° and 30° are shown by the blue and red arrows with magnitudes corresponding to the number of votes received (figure from [5]).

After the histograms are computed for each cell, a new structure is formed from overlapping blocks of 2×2 cells. The cells contained inside each block concatenate their 9 histogram values together. These 36 values are then normalized according to the ℓ_2 norm. This normalization with neighboring cells performs a type of equalization, minimizing the effects of illumination issues. Fig. 8(c) provides an example of this step as four neighboring cells are selected for combination within a block. Fig. 8(d) demonstrates how the nine dimensional cell content is concatenated. As this process is extended for all 55 possible blocks in the image patch a $11 \times 5 \times 36$ structure is formed, which is then restructured to form a 1980 element long HOG feature. Brooks [5] shows the general formula for determining the length of a HOG feature as:

$$length = (\#\text{bins}) \times (\#\text{cells per block}) \times (\#\text{blocks}), \tag{8}$$

$$\#\text{blocks} = \left(\frac{W_x}{\#\text{pixels per cell}} - 1\right) \times \left(\frac{W_y}{\#\text{pixels per cell}} - 1\right). \tag{9}$$

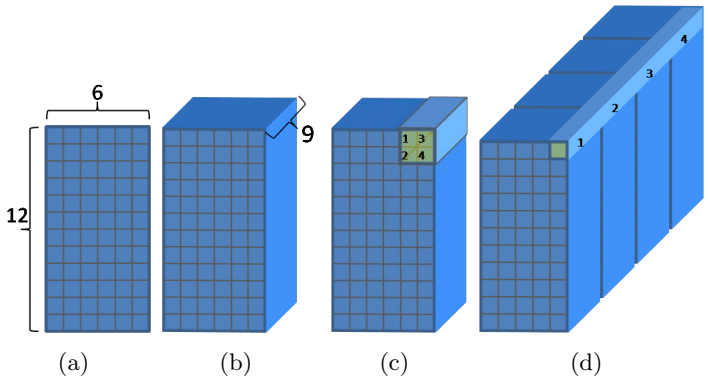


Figure 8. (a) A 9 bin histogram (of angles, weighted by magnitudes) are computed for each 8x8 pixel non-overlapping region or cell. (b) This resulting structure is visualized in three dimensions with the two positional dimensions as well magnitudes corresponding to nine different directions forming a matrix of size $12 \times 6 \times 9$. (c) Overlapping 2×2 cell structures are then used to form blocks. (d) Their third dimensions are concatenated forming a Matrix of size $11 \times 5 \times 36$. Note: most cells are used four times in different blocks. The third dimension is then scaled according to its ℓ_1 norm.

2.3.5 Support Vector Machines.

Support Vector Machines (SVMs) are a kernel-based method for binary classification that calculates the best separating hyperplane. The hyperplane then serves as a decision surface to determine which class a sample fits into [9]. Fig. 9 shows two linearly separable classes of data (one in red and one in blue) as well as two possible hyperplanes. A line is drawn perpendicularly from each decision boundary hyperplane to the nearest sample point. The Euclidean distance spanned by this line is referred to as the margin, ϵ, and is more formally defined as:

$$\epsilon = \min_m \frac{y_m \mathbf{w}^T \mathbf{x_m}}{||\mathbf{w}||}. \tag{10}$$

SVMs belong to a class of large margin classifiers meaning that the hyperplane generated

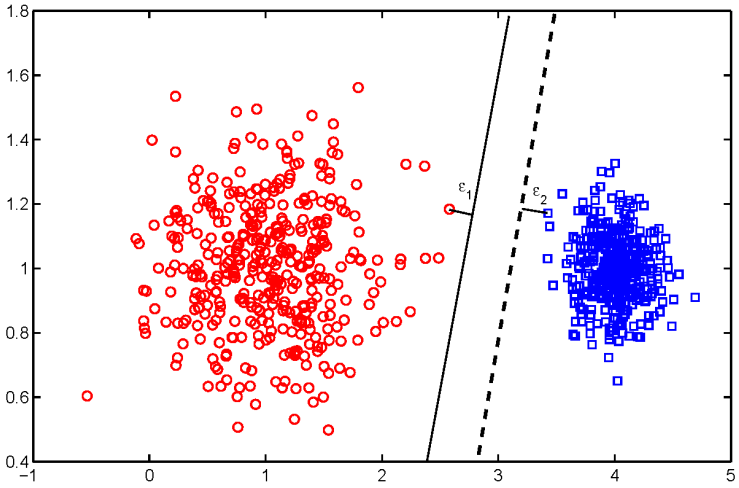


Figure 9. An example of two possible separating hyperplane decision boundaries for binary classification are shown with their corresponding margins [5]

provides the maximum margin possible and consequently the best separation between these two classes [7]. Given a data set $\{\mathbf{x}_1, ..., \mathbf{x}_m\}$ with $y_i \in \{\pm 1\}$ being the class label of $\mathbf{x}_i$, the process for calculating the hyperplane is defined as:

$$\mathbf{w}^T \mathbf{x}_i + b = 0 \tag{11}$$

where b is an offset value and $\mathbf{w}$ is a weight vector describing the decision boundary. Consequently, the input samples that fall into each class can be identified and assigned a label:

$$\mathbf{w}^T \mathbf{x}_i + b \geq 0 \qquad \text{for } y_i = +1 \tag{12}$$

$$\mathbf{w}^T \mathbf{x}_i + b < 0 \qquad \text{for } y_i = -1. \tag{13}$$

In the case that the data set is not linearly separable, a "soft margin" can be formed from a hyperplane that minimizes the number of mis-classifications and attributes a cost to each misclassified sample [3].

To train the dismount detector, a concatenated stack of m training samples, each of length n, are fed into the SVM trainer as well as a vector of training labels. For the SVM

described in [5], approximately 25,000 "true" image patches containing dismounts from an urban environment and a similar number of "false" patches were used in the training process. After their resulting HOG vectors are mapped and separated by a hyperplane, the data points found to be most useful in the division of the two classes are known as the support vectors. Each of these l support vectors are assigned a weight concatenated in vector $\mathbf{a}$. The stack of support vectors ($\mathbf{S}$) when multiplied by the respective weights contained in $\mathbf{a}$, yields a vector, $\mathbf{w}$, indicating the relative importance of each element in a given HOG vector as:

$$\mathbf{w} = \mathbf{S^T} \cdot \mathbf{a}. \tag{14}$$

When $\mathbf{w}$ is multiplied by a HOG vector and adjusted by a bias (b), a scalar prediction value results, indicating how strongly it judges the patch to be centered on a human. The prediction values are then constrained by a threshold (η_t) to complete the binary classification process.

During the formation of a SVM, it can be challenging to define an appropriate negative set of images that adequately defines the decision boundary. In order to produce iteratively stronger SVMs that do a better job of classifying results, the present SVM can be used to classify an additional large set of known negative data samples, and identify any false alarms that result. The data samples that yield false alarms quickly show themselves to be the set of common confusers that fall on the wrong side of the hyperplane. These samples are then chosen to define the set of negative samples in the next iteration of re-training the SVM. Fig. 10 shows a red decision boundary line that has been previously formed through the use of a linear SVM. When the prior positive training data (blue circles) as well as a new set of negative training data (red circles) are applied to this decision boundary, several false alarms (black plus signs) result. This process of accumulating false alarms to build a stronger negative set of samples produces a better defined decision boundary for each subsequent round of SVM training. However, accumulating a number of false alarms equal to the number of positive samples to train the next SVM iteration can be a time consuming

process as noted by [5].

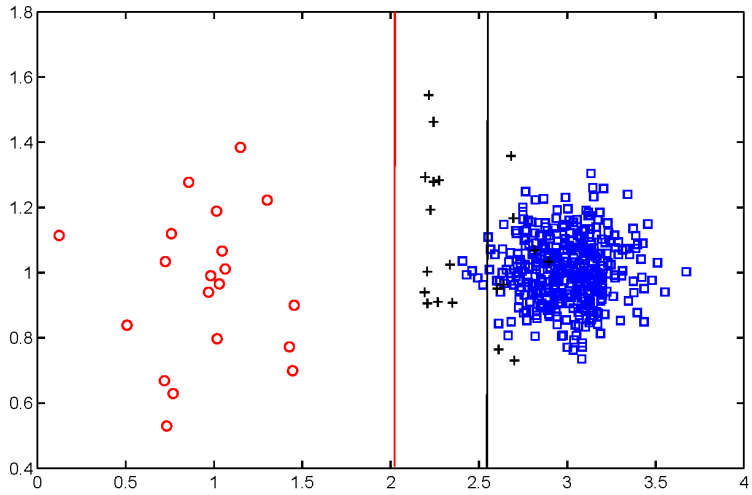


Figure 10. The false alarms resulting from the application of the red decision boundary are shown as black plus signs [5]

2.4 Classifying and Manipulating Unlabeled High Dimensional Data

In efforts to improve the detection of dismounts, additional training data is accumulated (as documented in later chapters of this thesis). This section provides greater detail on the mechanics of two tools to manipulate and classify unlabeled data: ISOMAP embedding and K-means clustering.

2.4.1 ISOMAP Embedding.

ISOMAP [35] provides a method of embedding high dimensional data into a lower dimensional manifold, while preserving inherant relationships within the data. First, the Euclidean distance is computed between the high dimensional samples to form a difference matrix as defined by:

$$D(i,j) = \sum_{x} |\mathbf{p_i}(x) - \mathbf{p_j}(x)|, \tag{15}$$

where $\mathbf{p_i}$ is the ith sample and $\mathbf{p_i}(x)$ is the xth element of the ith sample. ISOMAP utilizes the distance matrix to configure a neighborhood structure within the data samples. A lower dimensional surface is then warped to best fit the high dimensionality, closely

maintaining the relationship between neighbors. Consequently, ISOMAP preserves the geodesic or curved distances between the data samples while allowing for a reduction in dimensionality.

2.4.2 K-means Clustering.

K-means clustering is an iterative method that partitions an unlabeled set of n data points into K clusters [17]. This process begins by initializing the K mean vectors, $\mathbf{m}_i$ for $i = 1 : K$. Next, each sample is assigned to the mean vector ($\mathbf{m}_i$) that minimizes the Euclidean distance. After all n data points are assigned to a mean vector, the centroid of each of the new K clusters is computed, becoming the new mean. This process is repeated until convergence is reached.

3. Methodology

This thesis expands the work in [5] and seeks to provide a more in-depth understanding of HOG-based dismount detection, ultimately allowing for the improved detection of dismounts. This chapter begins by discussing several improvements to the baseline Brooks [5] detector that are implemented prior to the work explained in the remainder of this thesis. Next, a novel method for visualizing HOG vectors is outlined. The HOG visualization section is followed by the details of a process to explore the relationship between HOG features, particular SVM weights, and individual detections. A discussion of the limitations of the Brooks [5] detector follows. Finally, methods for extending the Brooks [5] detector to recognize additional poses are discussed, most notable is the inclusion of the layout for a dismount detector incorporating multiple SVMs.

3.1 Improving the Brooks Detector System

3.1.1 Use of the Peskosky Camera System.

An initial change from the methods used in the Brooks [5] detector is shown in the tools used to capture imagery. The Brooks system relied on the HyperSpecTIR version 3 (HST3) imager [19], a rather large hyperspectral imager capable of capturing approximately 1 frame every 10 seconds in a vertical scanning pattern (with a resolution of 250×1023 pixels). While this commercial system takes high quality images with all spectral bands properly registered, it is large and unwieldy, and requires significant post processing. Since this time, an experimental camera system designed by Peskosky [31] has become available. The Peskosky [31] imager provides a streamlined tool set, producing precisely the needed RGB and 1080nm and 1580nm wavelength frames at a resolution of 512×640 pixels with a camera frame rate of 30 frames per second.

3.1.2 Further Constraining Skin Islands.

False alarms often trigger off of either tall pole-like structures whose reflective properties register as skin, or tiny background pixels mimicking skin reflective properties. In order to suppress such common sources of false alarms, additional refinements are added to the detection process to constrain the size of the skin islands so they are consistent with expected proportions of a human head to body. Accordingly, thresholds are assigned to the minimum and maximum ratio (R_{min}, R_{max}) of the expected standing height of a dismount to their facial height. The facial height is measured by the vertical distance between the bounding box surrounding the assumed head skin island, h_i. The expected standing height, h_s, is measured as the distance from the top of the same skin island to the bottom of the image patch minus the approximate size of the bottom border, h_b. Fig. 11 displays four different scenarios regarding these height measurements. Fig. 11(a) shows the case when the target skin island height extends the total range of the dismount's head, allowing for optimal comparison of the dismount's standing and facial height. Often, however, contiguous skin islands are only achieved for a sub-portion of a dismount's head. Fig. 11(b) demonstrates the use of a range of allowable thresholds. Fig. 11(c) shows a likely case for rejection as a tall vertical object has comparable values of h_i and h_s. Conversely, the small value of h_i relative to h_s in Fig. 11(d) indicates a skin island that is likely too small to represent the head skin island of a dismount for a given window. This refinement on the number of valid skin islands both decreases search time and limits the number of false detections.

3.1.3 Allowing Horizontal Shifts.

The Brooks [5] detector noted the limitations in its testing set that primarily featured dismounts facing the imager in standing poses. For these image patches, the detected head skin islands generally lined up with the body of the dismount (as in Fig. 12(a)), corresponding well to the training data from the Daimler Benchmark set in [13], eliminating the need for horizontal offsets.

However, as the range of test data is expanded to include additional poses, camera angles,

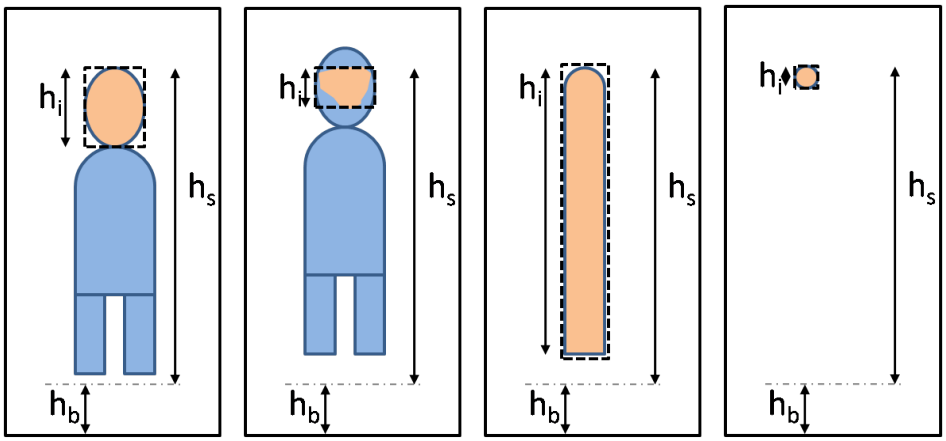


Figure 11. In the four subplots registered skin islands are shown in tan and are surrounded by a dotted bounding box. The height of these skin islands are h_i. Additionally shown are the estimated border height h_b and the standing height h_s. (a) A standard comparison of these heights is shown in an optimal case of skin detection. (b) A sub-optimal instance of skin detection indicates the need for flexibility in the span of R_{min} and R_{max}. (c) Image patches with tall skin islands can be effectively rejected by raising the threshold R_{min}. (d) Image patches with short skin islands can be rejected by lowering the threshold R_{max}.

and different lighting conditions, it is common for detection windows perfectly centered around the head skin island to yield sub-optimal results. As seen later in this chapter, proper alignment of the trunk of the body accounts for a large portion of an image patch's detection strength when using the SVM trained by [5]. Poor detections resulting from misalignment of the body trunk can occur in a variety of instances: when the dismount is viewed from a side perspective, the face is partially occluded, the dismount is standing in a configuration where their head is not centered over their body (as in Fig. 12(c)), or if the skin detection mask is faulty due to over-saturation or shadowing (as in Fig. 12(b)). In order to compensate for this potential difficulty in cuing, a limited range of horizontal offsets are added to each skin island centroid as possible cuing locations, using the head skin island as a contextual unit of measurement. For a skin island, I with a centroid at I_x,I_y, and width I_dx, search windows are formed around the set of points $\mathbf{P}$ such that:

$$\mathbf{P} = I_x + I_{dx} \cdot \mathbf{o_x}, I_y \tag{16}$$

where $\mathbf{o_x} = [\ -1.5 \quad -1.25 \quad -1.0 \quad -0.75 \quad -0.5 \quad -0.25 \quad 0 \quad 0.25 \quad 0.5 \quad 0.75 \quad 1.0 \quad 1.25 \quad 1.5\ ]$.

Fig. 12(d) demonstrates a new offset position (red diamond) that allows for better align-

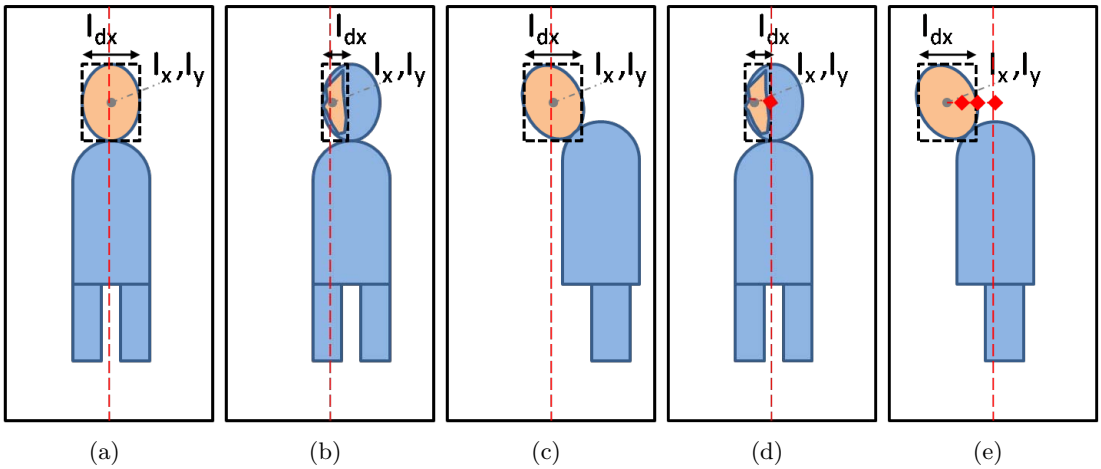


Figure 12. Five examples of image patches are shown above with registered skin islands highlighted in tan, surrounded by a dotted bounding box. The width of each skin island is I_{dx} with centroid, (I_x,I_y). Red dashed lines indicate the horizontal center of the image patch. (a) An image patch horizontally centered on the skin island yields good results when the dismount is facing forward with their entire face detected as skin. (b) A sub-optimal detection occurs when only a portion of the face is detected as skin, as the patch does not properly center on the bulk of the body. (c) Poor results occur when the entire face is properly detected as skin yet the head is not properly centered above the body. (d) The red diamond indicates a new offset positions (based on the width of the skin island) that allows for better alignment in the case of partial skin detection. (e) The red diamonds indicate several new offset positions (based upon the width of the skin island) that allow for better alignment in the case the head is not centered over the body.

ment in the case of a partial skin detection. Similarly, Fig. 12(e) shows several new offset positions (based upon the width of the skin island) that allow for better alignment in the case the head is not centered over the body.

The use of I_{dx} as a unit of measurement leverages available intuition as to the approximate size of a dismount's head necessitating fewer arbitrary pixel shifts (as in [5]) to achieve the same result, allowing for savings in computational time.

3.1.4 Limiting Multiple Detection Windows.

A problem experienced with the Brooks [5] detector is the issue of multiple detections for a single dismount (generally 10 per skin island). Brooks compensated for this issue by using a coverage statistic to examine overlapping detection windows in order to identify the best detection window for a dismount (see [5] [13] for a more detailed explanation).

While this process mitigates many of the detection windows generated from one skin island, still, multiple detection windows remain for each skin island. Additionally, this is a time consuming process as it requires the pairwise comparison of all detection windows. Fortunately, due to the nature of the skin cued detector, a more elegant solution is available by maintaining a running list, identifying which skin islands are used in the generation of each detection window. By choosing to only maintain the maximal positive detection (according to prediction value) for each skin island, virtually all multiple detections are avoided with computational ease. The only possibility for multiple detections is the case that multiple skin islands are in close proximity and are located directly around a dismount's head.

3.2 HOG Visualization

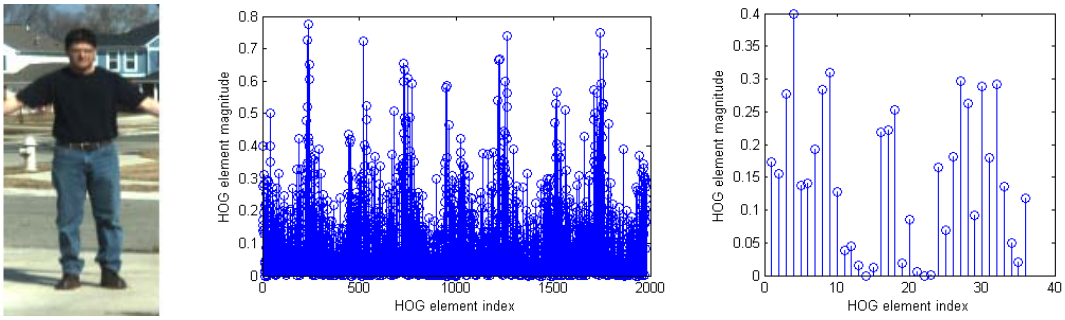
A primary focus of this thesis effort concentrates on understanding the impact that small changes in images have on their resultant HOG features. HOG features are useful in describing the direction of contrast in an image, however, the raw HOG feature on its own is quite abstract and it is difficult to gain an intuition from a linear string of numbers.

While many utilize HOG features to detect objects in imagery [2][5][13][25][38][39], few discuss methods for visualizing these structures. In many cases, HOG is primarily used as an intermediate step in the detection process. The work of [25] provides a good visualization of the edges of the un-normalized orientation histogram (the cell structure shown in Fig. 8(b)), however after these cells are normalized with their neighbors and concatenated together to form a HOG feature as described in Section 2.3.4, the same visualization technique provides limited intuition.

In order to provide a clearer medium to track changes in HOG vectors over successive image frames (and analysis in general), several different methods for visualizing HOG vectors are examined including linear inspection, an arrow representation on a gradient plot, a spatial representation of the block structure, and a novel restructuring to a cell representation.

Perhaps the most basic representation of the HOG vector is a linear stem plot of weight values. Such a representation is shown in Fig. 13(b). In this figure, the first 36 elements

are derived from the top left block of the image patch, with 9 consecutive elements from each of the four cells that compose the block (in a top-bottom left-right (t-b,l-r) ordering as in Fig. 8(c)). Subsequent chunks of 36 elements are obtained from following blocks according to the same (t-b,l-r) ordering. While the origin of each of the HOG elements may be explained, Fig. 13 demonstrates the difficulty in comparing a sample image with a stem plot of its HOG feature. Clearly, spatial context is necessary to visually relate HOG features to image patches.



(a) (b) (c)

Figure 13. (a) A sample image patch is shown for comparison with its HOG feature. (b) A linear representation of the full length HOG feature, containing 1980 elements, is challenging to visually compare against the sample image patch (c) An expanded view of the first 72 elements of the HOG similarly indicates the need for spatial context to match the HOG feature to the image patch.

An improvement to viewing the HOG feature as a stem plot is to draw a quiver plot by re-introducing the direction associated with each bin and scaling it according to the magnitude. In order to reduce overlaps in the plot, only one cell from each block is plotted at a time. The first quiver plot in Fig. 14 is from each of the top left cells from the image patch blocks, with successive plots from the remaining cells in the (t-b,l-r) ordering. This plot provides additional intuition by representing spatial location as well as direction and magnitude of gradients on the axes of the original image patch.

A HOG visualization method in [13] focuses on the total magnitude from each block. The 36 elements that correspond to each block (as seen in Fig. 8(d)) can be averaged in order to present a condensed form of the gradient magnitudes of the original image patch. The main advantage of this block magnitude method is that it is visually less complicated,

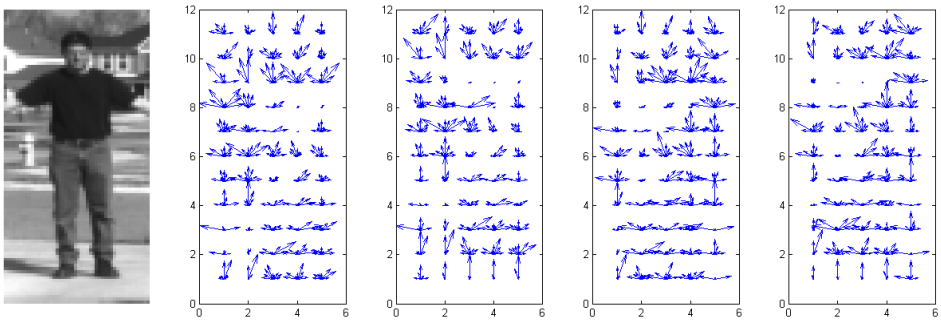


Figure 14. A quiver plot representation of the HOG feature vector. Blue arrows correspond to the magnitude and orientation of contrast changes for each cell with horizontal and vertical axis identifying cell position within the 12×6 cell structure.

representing only 55 distinct values (instead of all 1980 elements as in the previous two examples). Fig. 15 shows the simplicity of the block magnitude method indicating a clearer comparison between areas of contrast and block brightness.

(a) (b)

Figure 15. A representation of contrast magnitudes from the image patch in (a) is seen in (b), which is the combined magnitude representation. Low contrast areas are represented as black patches and areas of high contrast are represented as white patches.

The block magnitude representation, however, has several limitations as it includes a considerable amount of overlap and fails to offer details as it only tracks 55 different values.

A novel method for visualizing the HOG vector is to restructure the 1980 element long HOG feature to recover more of the original image patch's spatial context. As was mentioned in the background (and seen in Fig. 8), most cells from the patch are represented in four separate blocks after being normalized with different sets of neighbors. Consequently, when the HOG vector is mapped back to a spatial depiction (according to originating cells),

there are four representations for each orientation bin (one for each position a cell could hold within a block). Fig. 16 illustrates the four locations within the HOG feature wherein content from a particular cell is stored (when the HOG feature is viewed as a $11 \times 5 \times 36$ structure).

The top row of Fig. 17 contains nine frames featuring the normalized orientation bin votes (of cells in the first block position) mapped back to their contributing cell location. The tenth frame at the end of the first row is an average of these frames indicating the overall magnitude of the change in contrast for each cell in the top left block position. The following three rows similarly reflect the contributions of cells in other block positions. For simplicity in viewing, the 36 frames can be averaged across the different block positions as to display only the nine orientation frames for each sample chip. This method for viewing HOG features retains considerable detail regarding the content in each orientation bin and its spatial origination. (Note: due to the fact that cells on the edge of the chip are only) represented two times instead of four, the bottom row displays doubled weights for these border cells.) While this scaling serves as a valid equalization method, the horizontal edges of the combined 10° and 170° bins and the vertical edges of the combined 90° bin can become distorted due to edge effects during the gradient calculation. The distortion is especially apparent for benign backgrounds.

Mapping the complete HOG feature back into spatial cell context provides increased definition than that offered by the block magnitude representation. The 12×6 cell display presents a more refined scale with each square representing an 8×8 pixel cell instead of an overlapped 16×16 pixel block. While the new HOG visualization displays more information, it is less cluttered and more readable than a quiver or contour plot. Changes between two HOG features are also more apparent by glancing at comparative brightness than searching for changes in vector length in a quiver or contour plot visualization method. By glancing at the averaged 10° and 170° frames on the bottom row of Fig. 17, the dismount can be seen to exhibit strong horizontal contrast due to the vertical structure of their legs and arms. In all nine of the averaged orientation frames, a stark change in contrast is observed around the pixels associated with the head. This observation makes sense as the circular shape of

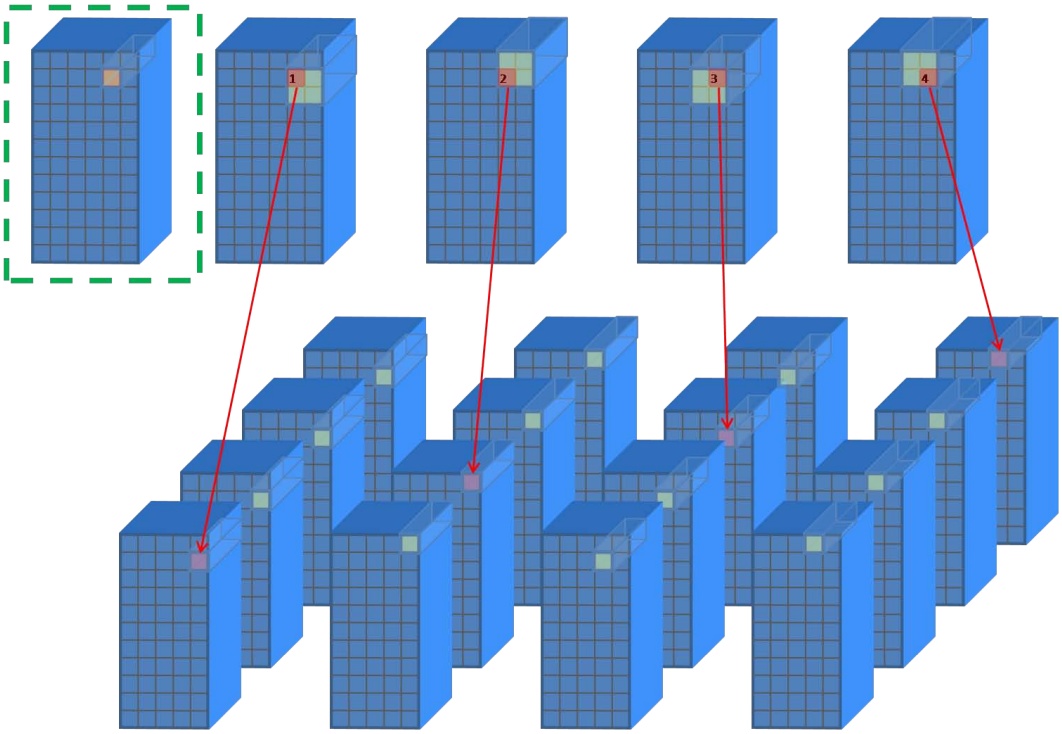


Figure 16. Five identical instances of the same $12 \times 6 \times 9$ cell structure are shown in the top row. The bottom row shows four identical instances of the same $11 \times 5 \times 36$ HOG feature that have been segmented depth-wise (also seen in Fig. 8(d)). An image patch's cell structure outlined by a green dashed line is represented in multiple location's throughout the HOG feature. The red highlighted cell belongs to four different blocks as represented in the top row by the white 2×2 boxes. The bottom row indicates the four locations (within the same HOG feature) that the cell content from each block is located.

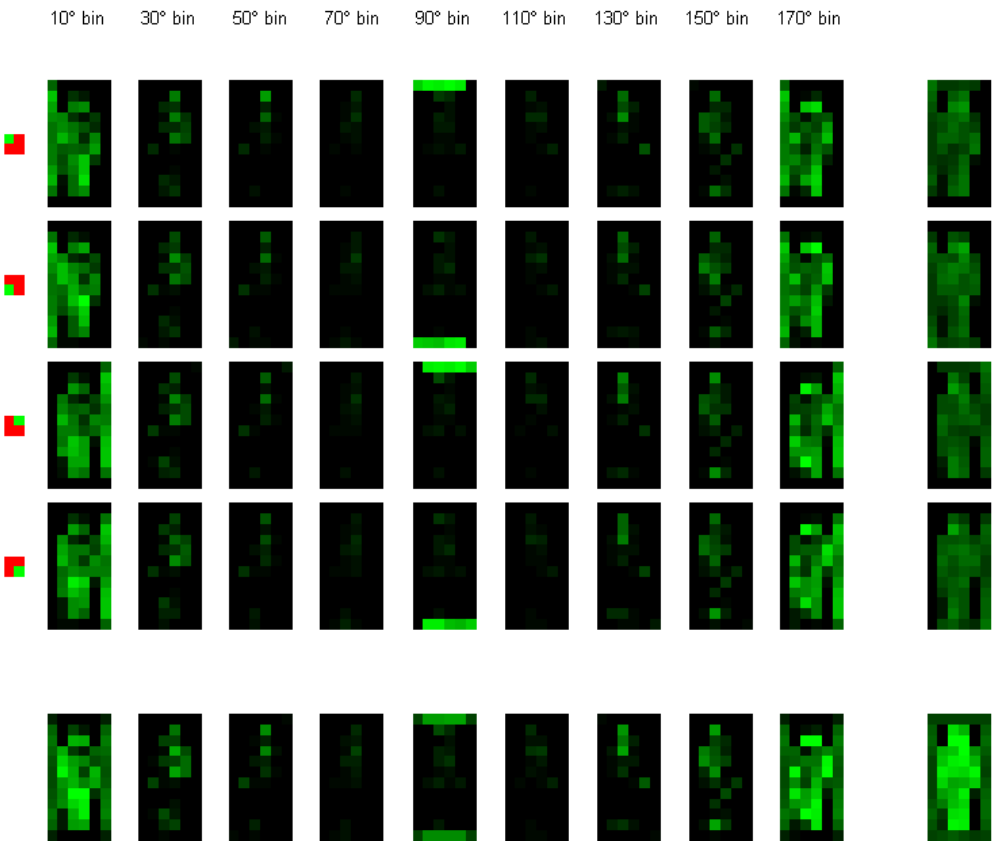


Figure 17. Spatial mapping of a HOG feature divided into orientation bins across the top and originating block position down the left. The block position for each row is indicated on the left as a green square occupying one of four positions on a red background. In the remainder of the figure, brighter shades of green corresponds to larger values. For simplicity in viewing, the 36 frames can be averaged across orientation angle (far right column) or originating block position (bottom row). The bottom right frame displays an averaged cell magnitude representation of the dismount.

the head would produce gradient vectors at all angles. Additionally, the right arm and side receive votes in the 30° bin as they are not completely vertical (a similar statement may be made about the left side of the body). Finally, the bottom right hand frame displays an averaged cell magnitude representation of the dismount, clearly showing the outlines of the body.

The display of the entire unaltered HOG feature in segmented frames according to orientation bin facilitates the inspection of individual body parts. The ability to analyze attributes of body parts in specific orientation bins is especially useful when examining a series of image patches displaying a dismount with slight movement. This method for HOG visualization allows for more intuitive understanding of how HOG features change in relation to articulations in pose and camera aspect angle.

3.3 Impact of Image Changes on HOG features

It is critical to asses the dependencies that HOG features have on changes in pose and aspect angle. Multiple ranges of motion are generated in imagery (from 3D models) to see how small changes in pose affect the HOG feature.

3.3.1 Use of 3D Data and Location of Simulated Skin Islands.

In order to analyze the effect small changes have on resulting HOG features, a sample individual is generated using the DAZ Studio^{TM}3D modeling program [18]. This software offers lifelike representations of people which can reflect realistic joint motion, body proportions, and gives the opportunity for varying camera views. This is particularly useful to manipulate the 3D-model's pose to reflect a desired range of motion or change in pose over a multiple frames. 3D-modeling allows for extremely tight control of all aspects of the image scene and makes the analysis of isolated changes more significant.

While the 3D-modeling program does not provide any spectral data in the near-infrared region that would normally be used for skin-aided cuing, it is still possible to exact locations of the head skin islands. Before the 2-D renders are made, the skin on the face and neck can be accented (in a duplicate copy of the image) with a specific marker color that is easy to

detect. In this thesis, blue is used as the base color of head skin regions. After the images are rendered, a color threshold is applied to find the image pixels such that:

$$\rho_B > \rho_R + \rho_G \tag{17}$$

where ρ_R, ρ_G, and ρ_B represents the red, green, and blue values over the image. Fig. 18 shows the successful application of Eqn. 17 and effectively identifies the simulated dismount's head skin island. The ability to achieve an effective skin mask (in the absence of near infrared data) allows the use of 3D models to simulate a variety of poses to test the dismount detector.

(a) (b) (c)

Figure 18. (a) An original image patch. (b) A "blue masked" version of the same dismount and pose as in the image patch. (c) The application of Eqn. 17 effectively identifies the head skin island.

3.3.2 Changes in HOG Due to Slight Changes in Imagery.

Sequences of images demonstrating different ranges of motion such as sitting, crouching, arm extensions, and bending to the side are generated to examine the Brooks [5] detector. Variations in HOG features are also analyzed as camera aspect angle changes for a dismount in a constant pose. For each range of motion, a 3D model is generated in 6-10 incrementally progressing poses maintaining a constant window size. As the primary purpose of this section is to illustrate the effect of controlled changes, cuing based off the location of the head is only used to correctly select the first image patch in each sequence. The resulting

HOG features are then visualized utilizing the same methods demonstrated in Fig. 17 to show spatial and orientation angle changes for various cell positions. Selected examples are discussed below.

3.3.2.1 Arm Extension Example.

Fig. 19 displays dismounts standing in the same location with arms held at varying angles. As the arms and shoulders shift angle, strong gradient magnitudes are seen to reflect these changes. In Fig. 19(a), the arms and shoulder appear close to the torso and are largely reflected in the 10° and 170° orientation bins. In Fig. 19(b), larger magnitudes are registered in the 50° and 130° bins mirroring the position the right and left arms (respectively) are located as the arms are held at a 45° below the horizontal. Similarly, strong magnitudes are shown from 70° to 110° due to arms being near parallel in Fig. 19(c) and 19(d). Clear changes in gradient magnitudes are seen in the 50° and 130° orientation bins of Fig. 19(e) corresponding to the left and the right arms and shoulders.

3.3.2.2 Limited Rotation Example.

A limited rotation is applied to a dismount in Fig. 20 as the camera pans through a 0° to 90° azimuth angle. As the dismount shifts from a front to side perspective view, their width decreases and the area previously occupied by their arms receives fewer gradient votes across all orientation bins.

3.3.2.3 Side Bend Example.

An example showing a dismount with an increasing side bend towards the subjects left side is shown in Fig. 21. As the individual moves, the content of the specific spatial cell moves as is seen in the final frame in each row. However, there are also obvious shifts between orientation bins exemplified as the individual's left arm and shoulders shift from being seen as part of the torso, primarily displayed in the the 170° bin (Fig. 21(a)), to more strongly represented in the 130° and 110° bin (Fig. 21(c) and 21(d)). Similarly, stronger

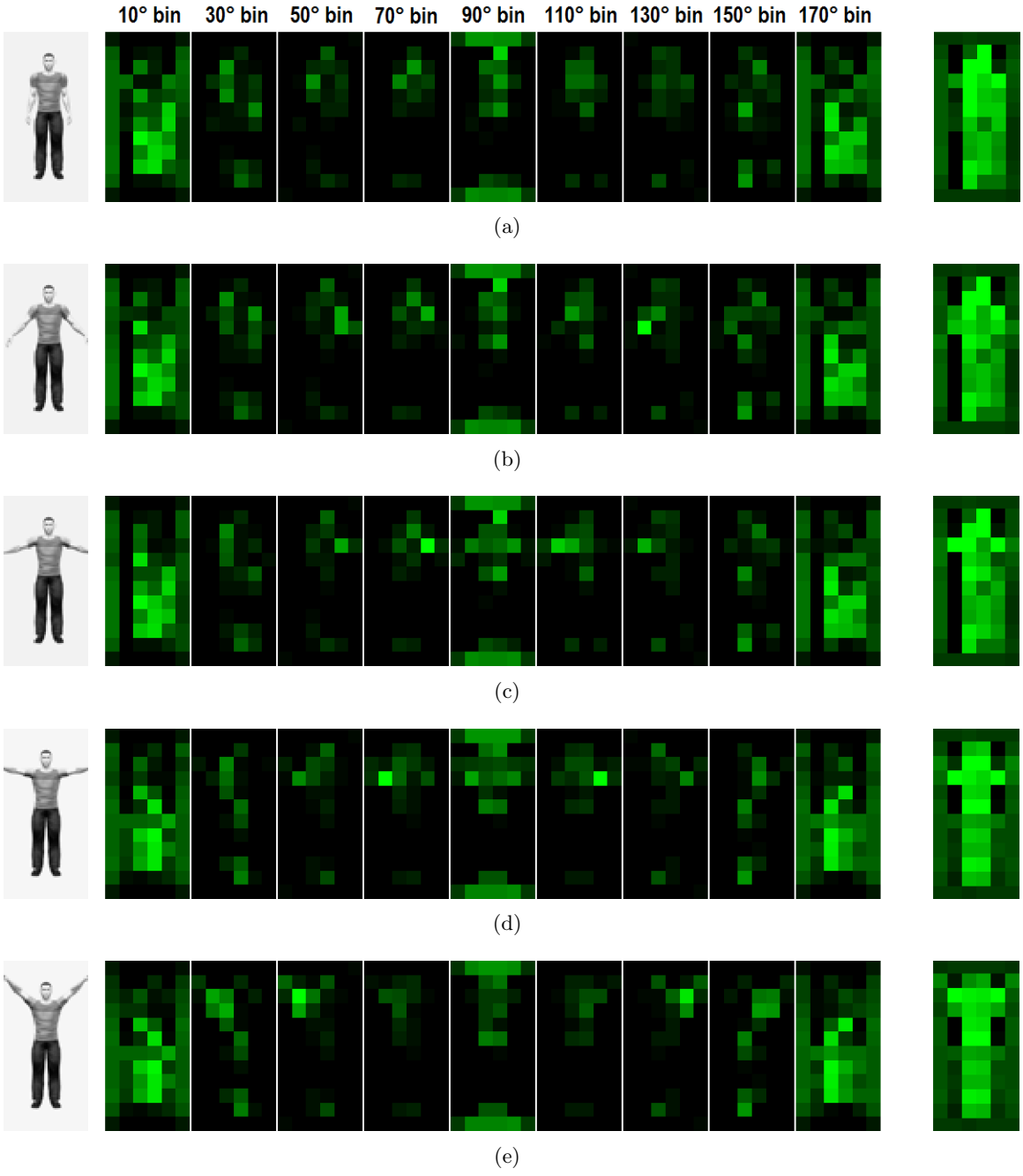
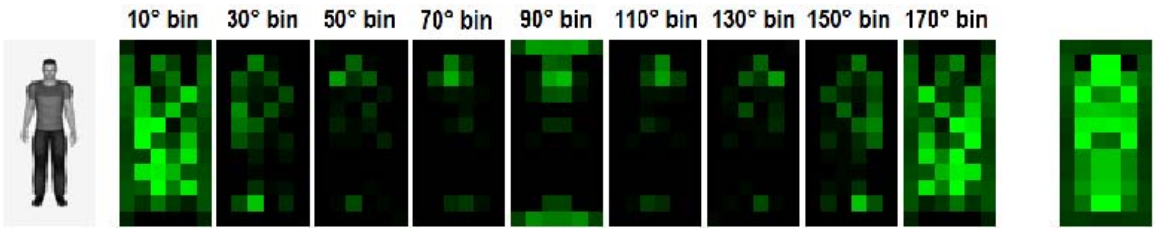
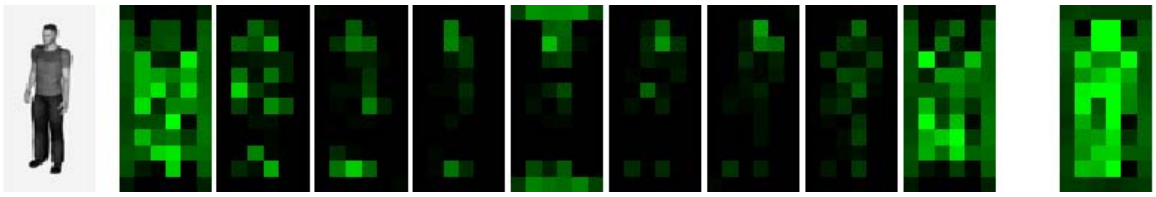


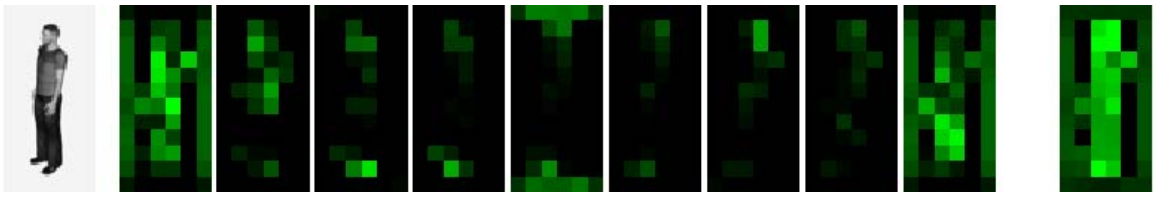
Figure 19. A visualization of HOG features by orientation bin for an arm raise scenario. In each row the original image chip is featured first followed by orientation bins with bin centers at 10°, 30°, 50°,..., 170°. The final frame in each row shows the total cell magnitude. The image chips in each row display a 20° angle of elevation for the camera with a varying arm angles above or below the horizontal: (a) -75° (b)-45° (c) -15° (d) 15° (e) 45°.



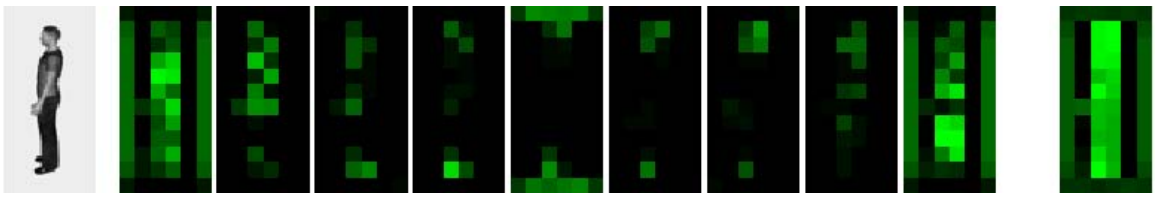
(a)



(b)



(c)



(d)

Figure 20. A visualization of HOG vectors by orientation bin for a rotation scenario. In these images, the camera angle has a 20° angle of elevation with a varying azimuth angle of: (a)0°, (b)30°, (c)60°, and (d)90°.

gradients orthogonal to the shoulders are also seen shifting from the 90° bin to the 110° and 130° bins as the severity of the side bend increases.

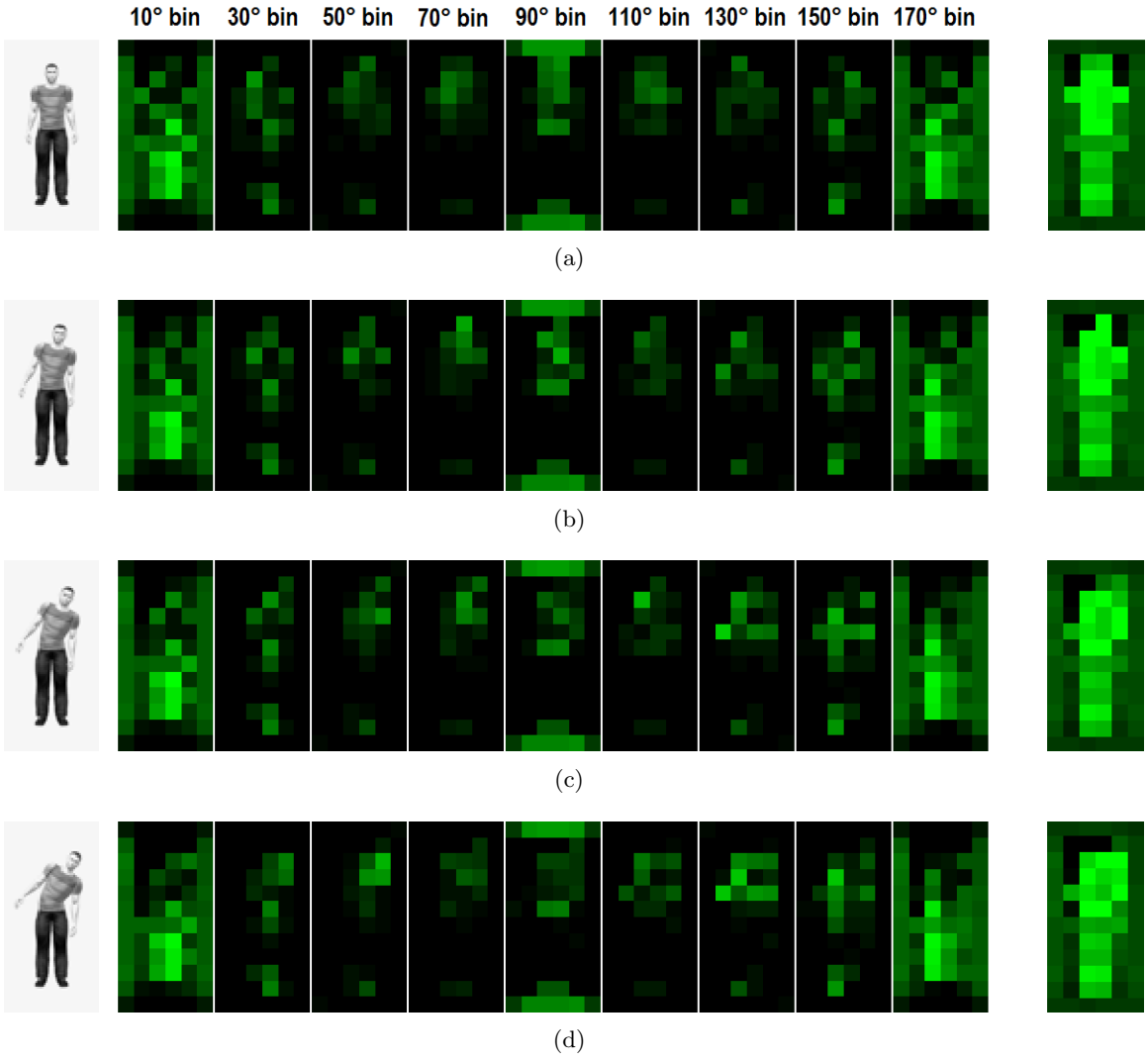


Figure 21. A visualization of HOG features by orientation bin for a side bend scenario. In each row the original image chip is featured first followed by orientation bins with bin centers at 10°, 30°, 50°,..., 170°. **The final frame in each row shows the total cell magnitude. The image chips in each row display a** 10° **angle of elevation with a varying side bend angles to the right of approximately: (a)** 0° **(b)**10° **(c)** 20° **(d)** 30°.

Clearly, slight changes in pose and camera aspect angle have a noticeable effect on HOG features as seen by analyzing the features according to orientation bin in a restructured cell format. However, the significance that the changes in HOG features have upon detections relies completely on the distribution of the SVM weights.

3.3.3 Relationship Between HOG Vectors, SVM Weights, and Prediction Strength.

As changes in image patches are seen to affect HOG vectors, it is of interest to observe spatially how these features interact with the SVM weights (the linear combination of the support vectors defined as **w** in Eq. 14) for the detector trained in [5]. Fig. 22 displays the vector of SVM weights in a format similar to that of Fig. 17 in order to identify the portions of the HOG feature that are critical to establish a positive detection.

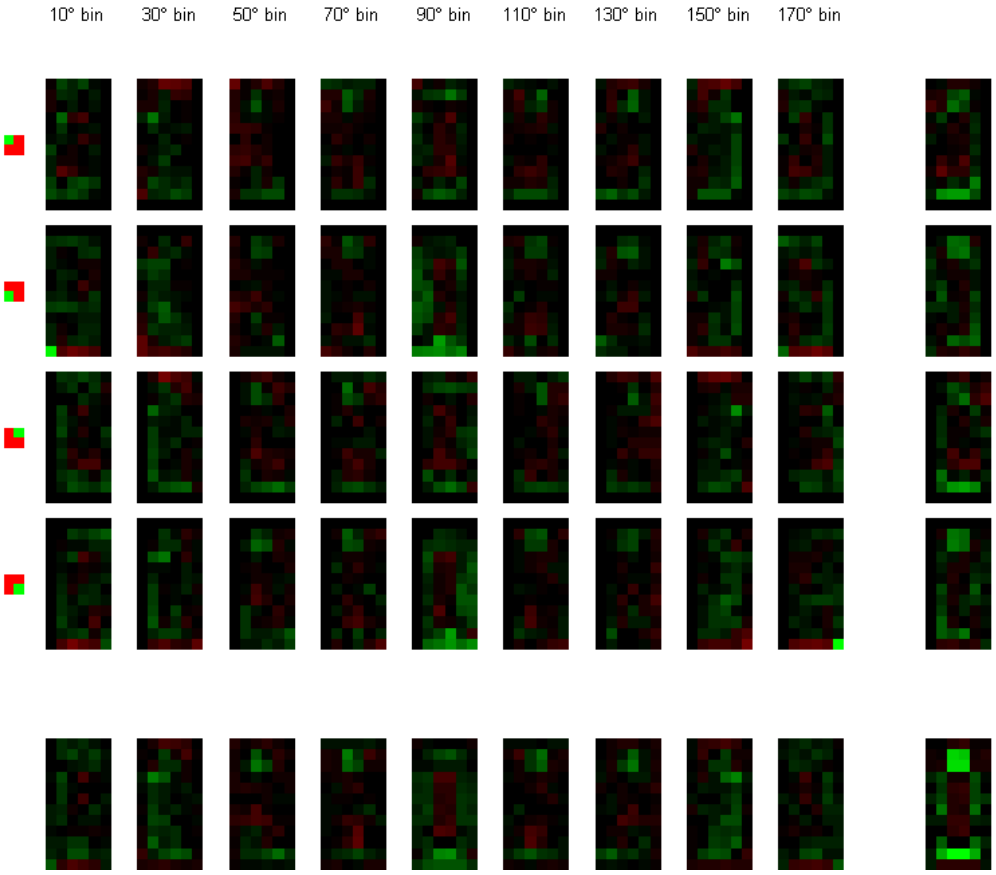


Figure 22. Spatial mapping of the SVM element weights are arranged in the locations corresponding to their respective HOG elements (with orientation bins tiled across and originating block position tiled vertically. The block position for each row is indicated on the left side of each row as a green square occupying one of four positions on a red background. In the remainder of the figure, bright red corresponds to strong negative weighting and bright green representing strong positive. For simplicity in viewing, the 36 frames can be averaged across orientation angle (far right column) or originating block position (bottom row). The bottom rightmost frame displays an averaged SVM weight according to cell position and indicates a heavy positive weights around the outline of an individual.

In Fig. 22, the relative weights corresponding to each HOG element are shown spatially.

In particular, the bottom rightmost frame shows the clear support of a standing individual and highlights the importance of spatial location, as edges around the individual appear with strong weights. One can observe that the outline of the body contributes heavily to positive prediction values, where contrast changes within the torso region detracts from the prediction strength. Consequently, it is important not to have too much variation in the trunk of the body; proper alignment of the torso is key so that the silhouette lines up correctly. The 90° bin appears to have a heavy impact, highlighting the vertical differences between the ground versus feet as well as head and shoulders versus the background. The HOG features can be multiplied by their respective SVM weights to observe contributions to their prediction value, emphasizing how directional magnitude changes contribute to the overall prediction score.

When HOG features from the arm extension examples are weighted by the SVM from [5] as in Fig. 23, centered positions of the head and feet are seen to contribute heavily to the prediction score, while vertical contrast in the torso due to wrinkles or a pattern on the shirt detracts from the the score. The locations of the right and left shoulders in Fig. 23(a) and 23(b) in the 30° and 150° orientation yield positive contributions to their score that are not noticed in the remaining rows. As the dismount raises his arms away from his torso, negative score components appear in the 10° and 170° to either side of the torso. Additional negative score components appear in the last four rows according to the location and orientation of the forearms and hands (right and left respectively): Fig. 23(b):50° and 130°, Fig. 23(c):70° and 110°, Fig. 23(d):110° and 70°, and Fig. 23(e):130° and 50°. The net effect of all of the score contributions results in total prediction scores of: 1.198, 0.568, 0.235, -0.167, and -2.158 (respectively according to row). Consequently, standing poses with arms held above the horizontal are likely to receive prediction scores below a detection threshold, η_t, of 0.

When the HOG features in the rotation example are multiplied by the SVM weights as in Fig. 24, the centered position of the head and feet, as well as the side positions of the arms, form the primary basis for the prediction score. Figs. 24(a) and 24(b) maintain high prediction scores do to the placement of their arms at their sides. However as the dismount

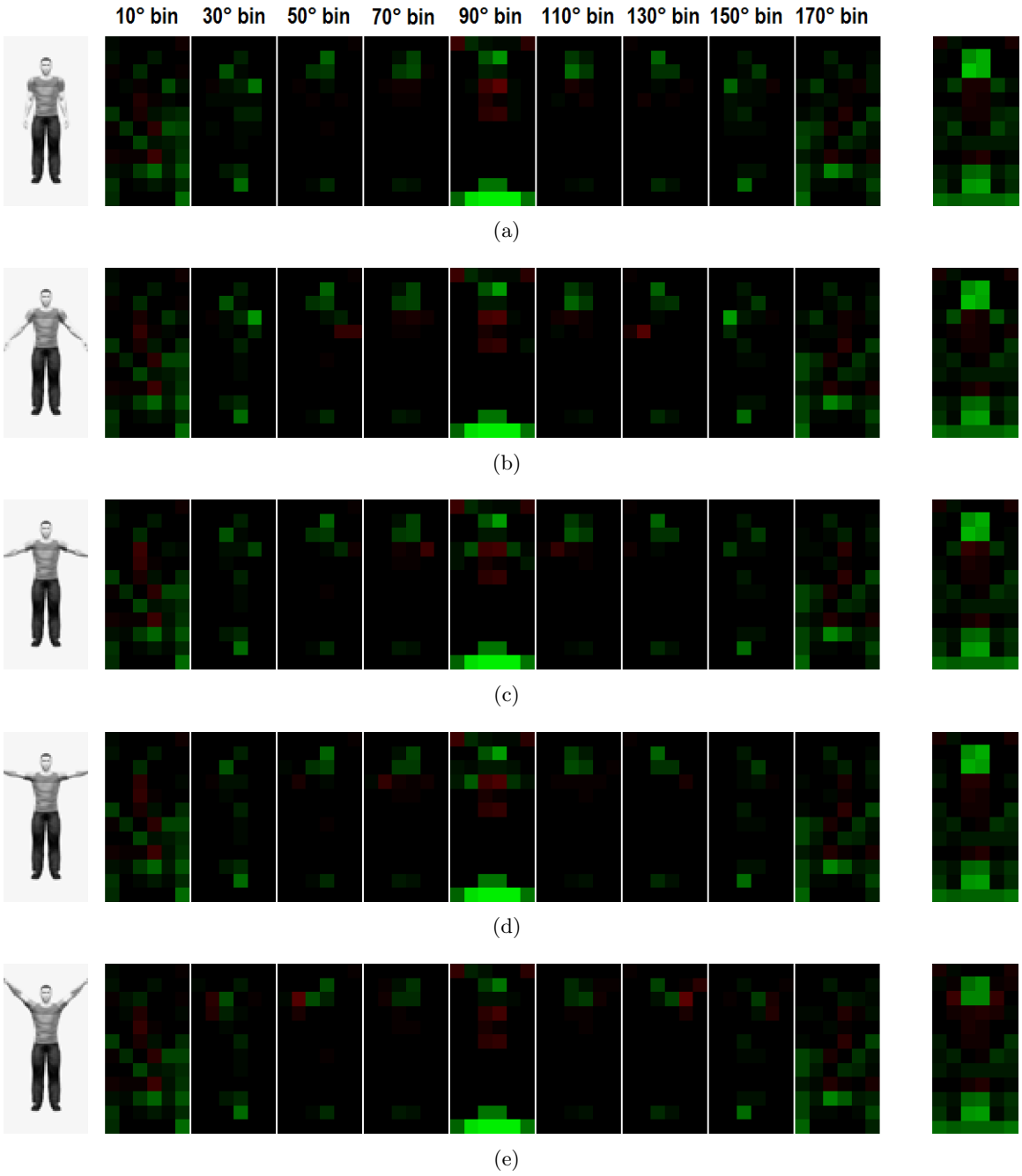


Figure 23. A visualization of weighted HOG features by orientation bin for an arm raise scenario. In each row the original image chip is featured first followed by orientation bins with bin centers at 10°, 30°, 50°,..., 170°. The final frame in each row shows the total cell magnitude. The image chips in each row display a 10° angle of elevation with a varying arm angles above or below the horizontal. Each patch with varying arm angle receives a distinct prediction score: (a) -75° : 1.198 (b) -45° : 0.568 (c) -15° : 0.235 (d) 15° : -0.167 (e) 45° : -2.158.

rotates to 60° and 90° in Figs. 24(c) and 24(d), the prediction score quickly drops as the arms merge into the body and the dismount's silhouette decreases in width.

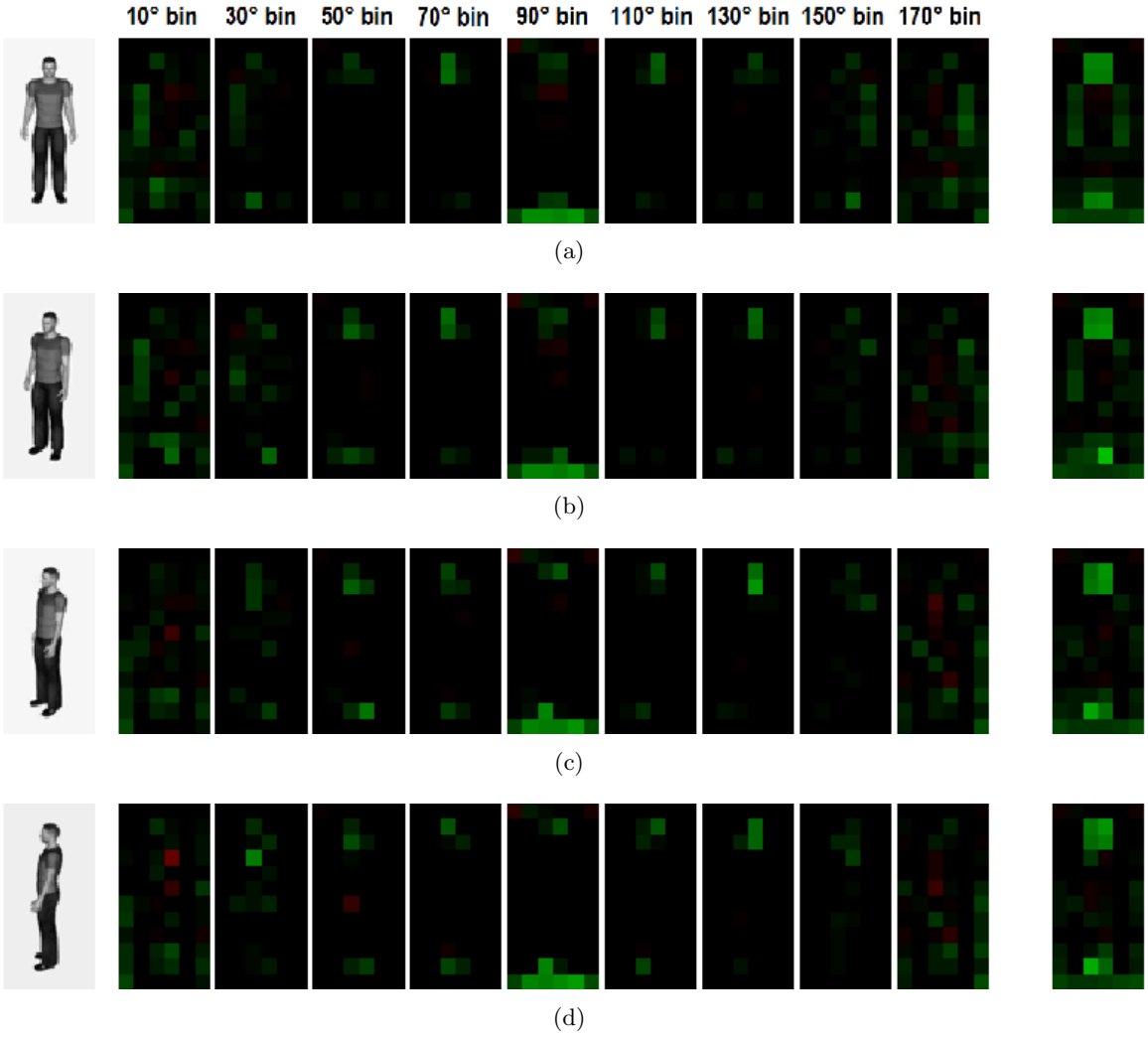


Figure 24. A visualization of HOG vectors by orientation bin for a rotation scenario. In these images, the camera angle has a 20° angle of elevation with a varying azimuth angle. The prediction strengths for these changing azimuth angles: (a) 0°: 3.446 (b) 30°: 3.1314 (c) 60°: 0.9813 (d) 90°: 0.529

The same head, arm, and feet locations play a critical role in the prediction score for side bends as seen in Fig. 25. The head shifting out of the correct spatial location profoundly detriments the prediction score. The drop in prediction scores is even more profound in the case of cuing based off of the position of the head.

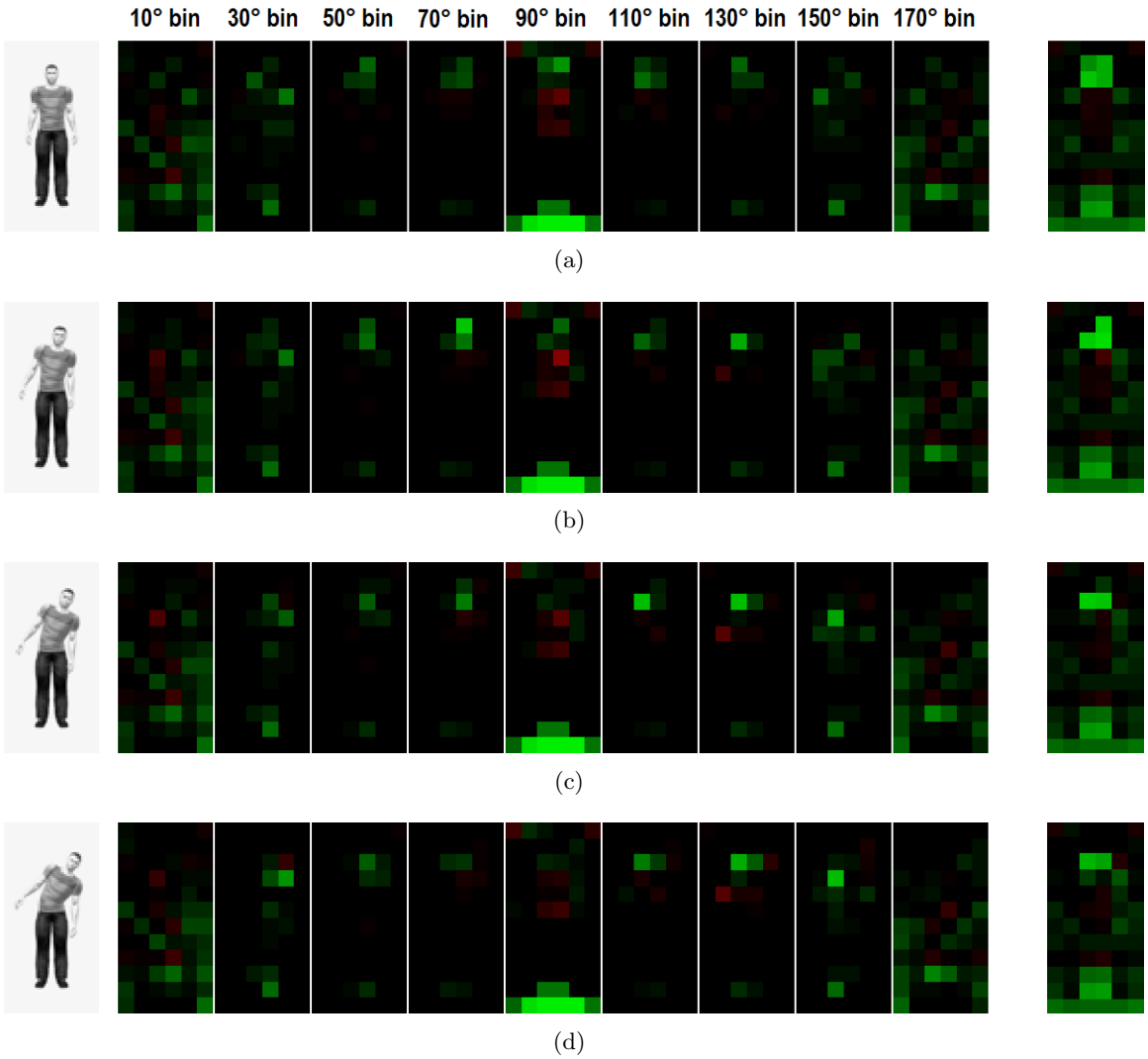


Figure 25. A visualization of weighted HOG features by orientation bin for a side bend scenario. In each row the original image chip is featured first followed by orientation bins with bin centers at 10°, 30°, 50°,..., 170°. The final frame in each row shows the total cell magnitude as weighted by the SVM. The image chips in each row display a 10° angle of elevation with a varying side bend angles to the right. The prediction score of each patch is dependent on the relative degree of the side angle: (a) 0° : 1.200 (b)10° : 0.803 (c) 20° : -0.045 (d) 30° : -0.398.

3.4 Identifying the Limitations in the Brooks Detector

The work in the previous sections gives clear indications that the HOG features are impacted by changes in pose and camera aspect angle as alterations are manifested in the orientation of shapes within the image. In order to characterize the limitations in the Brooks [5] detector, a broad variety of poses and diverse camera angles are tested (over 25,000 in all). The same techniques described earlier are used to test the support of other detectors. The use of three dimensional models of dismounts are once again of particular use to automatically render images of an individual from specified camera angles.

Instead of rendering images from every possible camera angle, this thesis effort concentrates on views in which skin is visible on the dismounts' faces (as this is a requisite for cuing the Brooks [5] detector). For simplicity, the azimuth angle is chosen to range from $[-90°, 90°]$ (in $5°$ increments) from the left side view of the dismount all the way to the right side view. Similarly, the camera elevation angle ranges from $[0°, 50°]$ in $5°$ increments. This span of $50°$ is representative of the camera angles offered by a low flying unmanned aerial vehicle (UAV) or a building mounted camera. A view of the relation between camera position and the dismount is shown in Fig. 26.

For analysis, three dimensional dismount models are viewed and rendered from 407 different camera angles corresponding to the locations indicated in the previous paragraph. Each pose is rendered into a 350×500 pixel image from a constant camera distance with one of many urban backgrounds. The images are then passed to the dismount detector system with their accompanying skin detection masks. As the dismount detector searches through the image, an additional process is added to automatically save the resized 96×48 pixel image patch (and corresponding HOG feature) that yields the maximum prediction value allowing for horizontal shifts as explained in Section 3.1.

A series of crouching poses are shown in Figs. 27, 28, 29, 30, and 31 to indicate the maximum detection strength for each camera position. Fig. 27 shows a dismount crouching very low to the ground which only yields prediction values above the detection threshold in a narrow region from the frontal camera positions with an angle of elevation below $20°$.

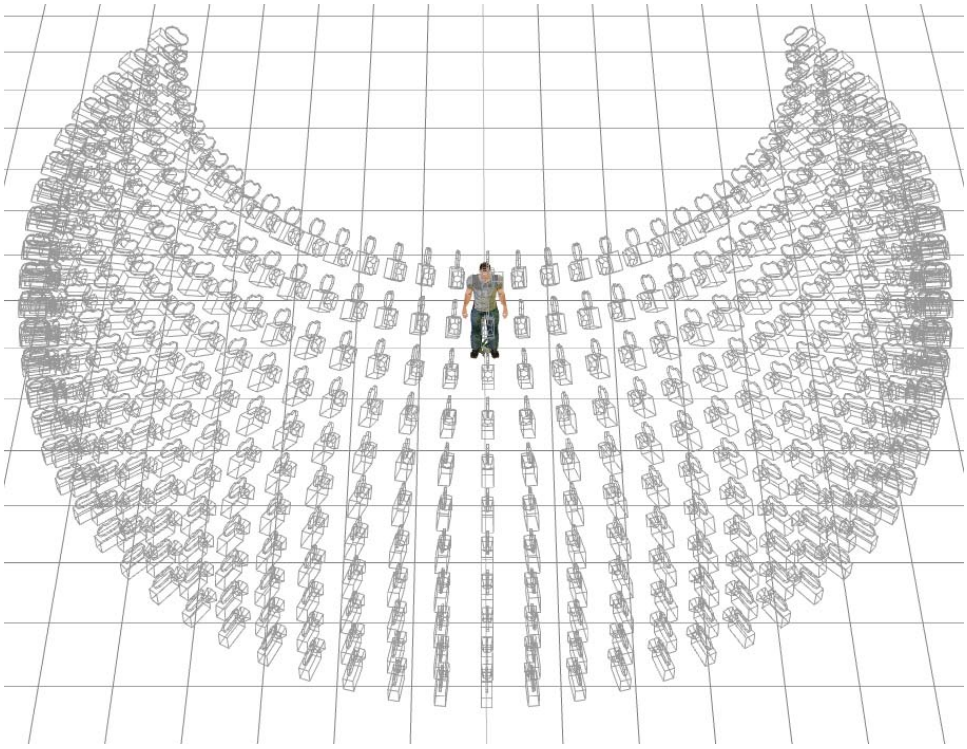
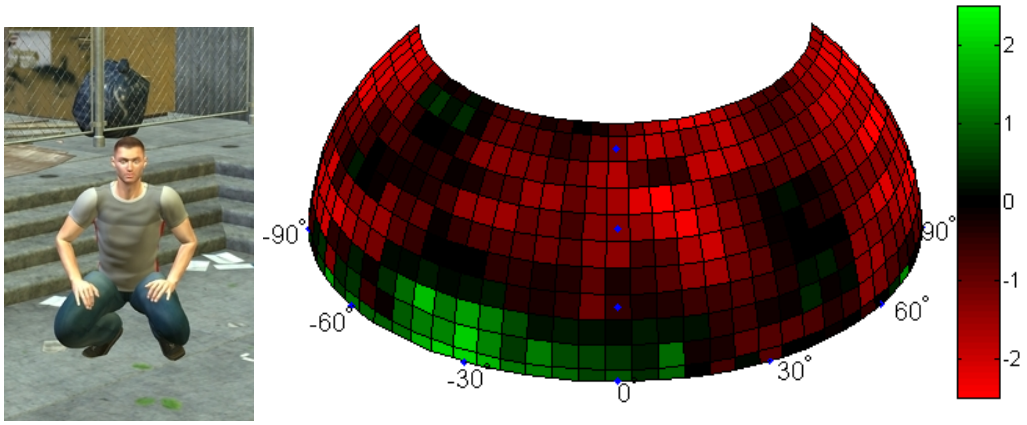


Figure 26. The placement of 407 cameras are shown to span an azimuth angle from $[-90°, 90°]$ (in $5°$ increments) with an angle of elevation ranging from $[0°, 50°]$ in $5°$ increments.

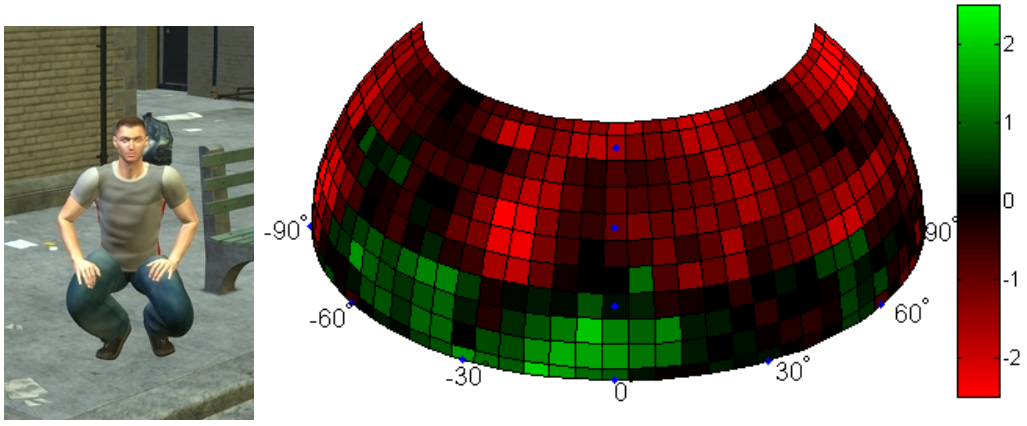
As the individual moves into a higher crouching position as pictured in Fig. 28 and 29, the dismount's body trunk elongates and consequently, is detected in an increased range of azimuth and elevation angles. Fig. 30 shows a very high crouch and is detected in a greater range of camera elevation angles than previously seen. The dismount shown in Fig. 31 exhibits only a very slight crouch and is easily passable for a standing dismount. This pose configuration provides little difficulty for the detector system and is detected at near the detection threshold for almost all camera locations.

Other selected hemispherical plots featuring several selected poses, highlight areas of difficulty in obtaining positive detection with the SVM trained in [5]. A sitting pose shown in Fig. 32 exemplifies a problem for the SVM: dismounts with open or varying leg positions. While the torso appears as a standing dismount from a frontal view, the spread apart legs throw off results and only yields a few detections near a threshold of $\eta_t = 0$. The kneeling pose in Fig. 33 yields good detections when viewed in a frontal view, however, as it is viewed from the side, the contracted width of the dismount as well as the protruding leg hamper



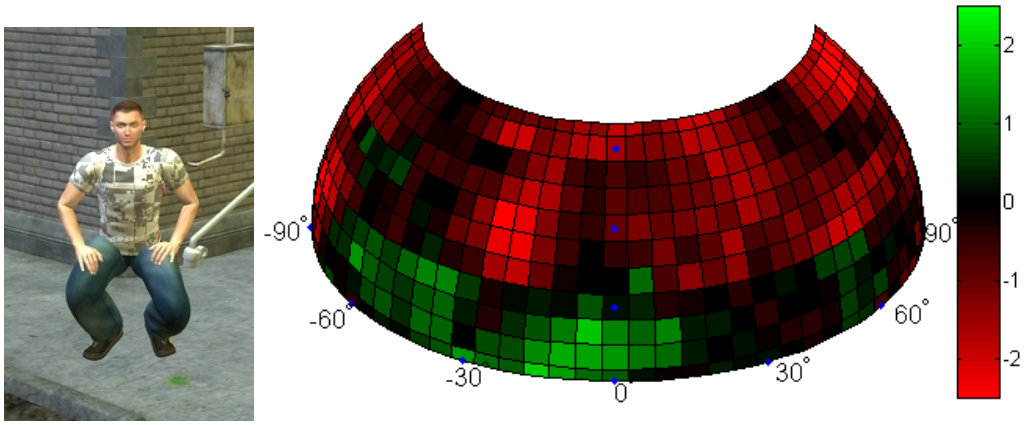
(a) (b)

Figure 27. (a) Snapshot of 3D-model demonstrating a crouching pose (b) hemispherical plot displaying the prediction strength over 407 different camera angles. The color scale is fixed from -3 to 3 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience.



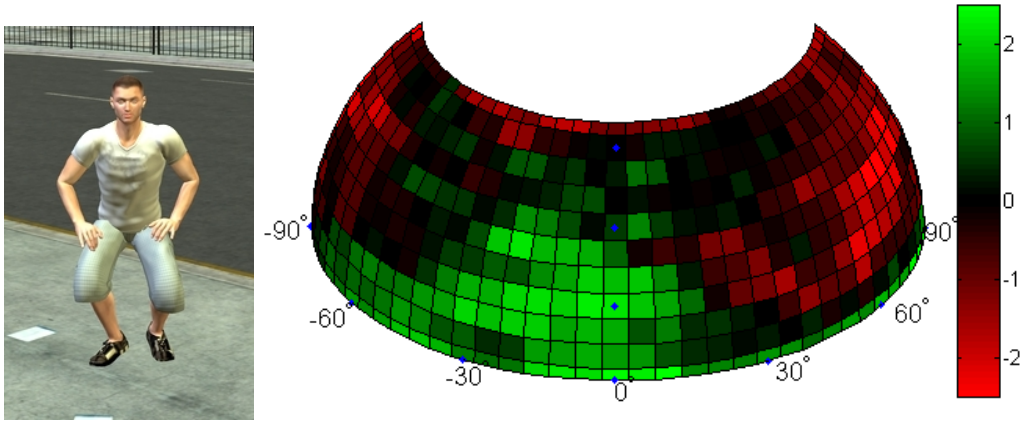
(a) (b)

Figure 28. (a) Snapshot of 3D-model demonstrating a crouching pose (b) hemispherical plot displaying the prediction strength over 407 different camera angles. The color scale is fixed from -3 to 3 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience.



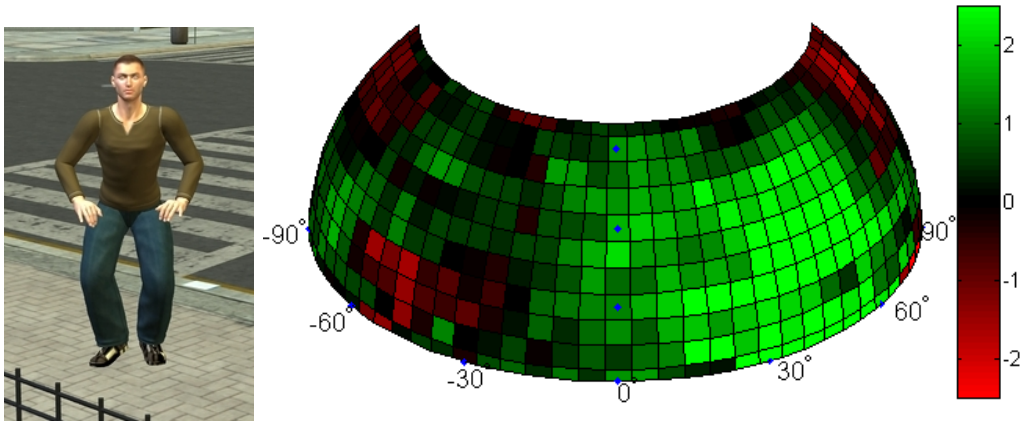
(a) (b)

Figure 29. (a) Snapshot of 3D-model demonstrating a crouching pose (b) hemispherical plot displaying the prediction strength over 407 different camera angles. The color scale is fixed from -3 to 3 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience.



(a) (b)

Figure 30. (a) Snapshot of 3D-model demonstrating a crouching pose (b) hemispherical plot displaying the prediction strength over 407 different camera angles. The color scale is fixed from -3 to 3 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience.



(a) (b)

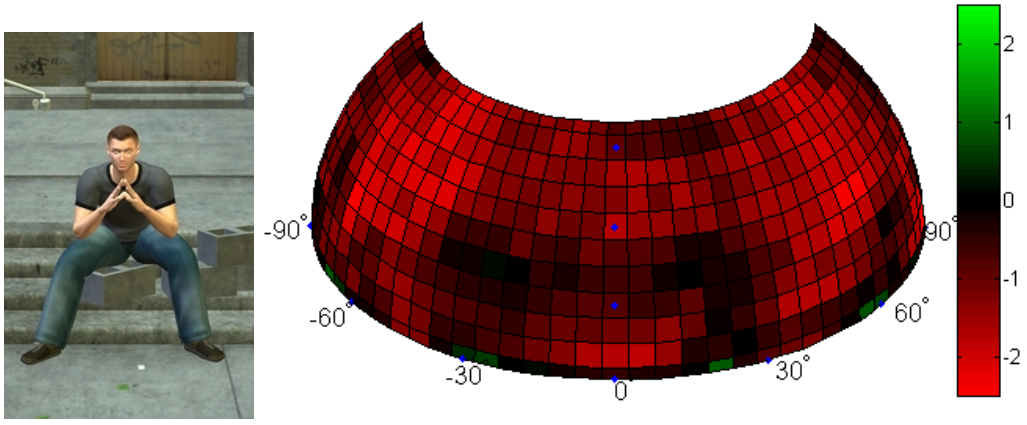
Figure 31. (a) Snapshot of 3D-model demonstrating a crouching pose (b) hemispherical plot displaying the prediction strength over 407 different camera angles. The color scale is fixed from -3 to 3 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience.

any side detections.

When the Brooks [5] detector is equipped with tools allowing for controlled horizontal shifts, it is able to detect several additional dismounts that are in near standing positions (with an upright torso and legs directly beneath them). However, even for these near standing positions, the detector has difficulty when the cameras azimuth angle exceeds 50° (or falls below −50°). While many crouching and squatting poses appear similar to standing poses, prediction scores are reduced due to their compressed height-to-width ratio. This mis-proportion is only accentuated as the angle of elevation increases; a significant cut off is seen above 15°.

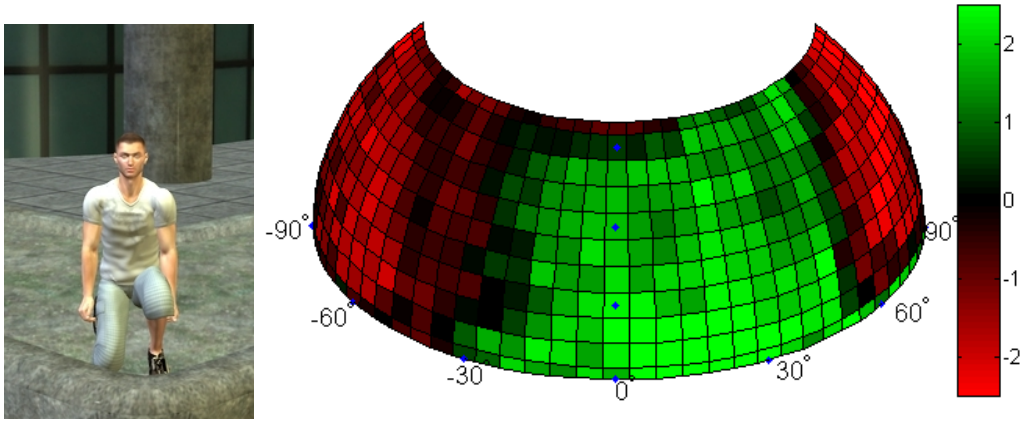
3.5 Extending the Dismount Detector's Ability to Recognize Poses

As seen in previous sections, the support of the current support vector machine trained in [5] simply does not extend to all common poses or camera angles. This section outlines methods for extending the modified Brooks [5] detector's ability to recognize poses.



(a) (b)

Figure 32. (a) Snapshot of 3D-model demonstrating a crouching pose (b) hemispherical plot displaying the prediction strength over 407 different camera angles. The color scale is fixed from -3 to 3 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience.



(a) (b)

Figure 33. (a) Snapshot of 3D-model demonstrating a crouching pose (b) hemispherical plot displaying the prediction strength over 407 different camera angles. The color scale is fixed from -3 to 3 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience.

3.5.1 Related Work.

There are multiple techniques proposed in the literature for the detection of dismounts in various poses. The work of [25] proposes a part template matching technique, where the silhouette of a dismount is extracted, segmented, into a top, mid, and bottom region, then compared to various hierarchical templates for detection and pose determination. Once a dismount is located, the work of [1] focuses on estimating 3D pose by performing background subtraction, locating joint centers, and describing pose as the angle between joints. These sources provide a great deal of insight, however most of the testing is limited to sets of dismounts in similar upright standing or walking poses from ground level camera angles. A more extensive range of poses is pursued by [38] who clusters various silhouettes of individuals and trains separate detectors based off of pose groupings.

3.5.2 Extending Detection Through Additional SVMs.

The technique of using multiple SVMs to detect dismounts in varying configurations as suggested in [38] is promising as it allows a streamlined augmentation of the improved Brooks [5] detector. However, it is not immediately clear which types and groupings of training data are necessary to train additional SVMs. In order to counter this dilemma, an initial sampling of "live data" is collected with the Peskosky [31] imager over several poses to gain a basic understanding. For each of these "live data" cubes, the basic skeleton of the dismount, their outline, and the detected skin region corresponding to the face are extracted and shown in Fig. 34. The poses shown represent diversity in the location of feet, relative width of torso, position of arms, and relative distance between feet and head. Additional computer generated poses are modeled to extend the variety of poses and camera angles.

3.5.2.1 Clustering A Variety of Poses.

After sufficient training data are amassed, it is necessary to determine how different poses relate to each other and to logically group similar pose/camera angle configurations. Silhouettes offer a logical starting point to analyze pose configurations. As previously

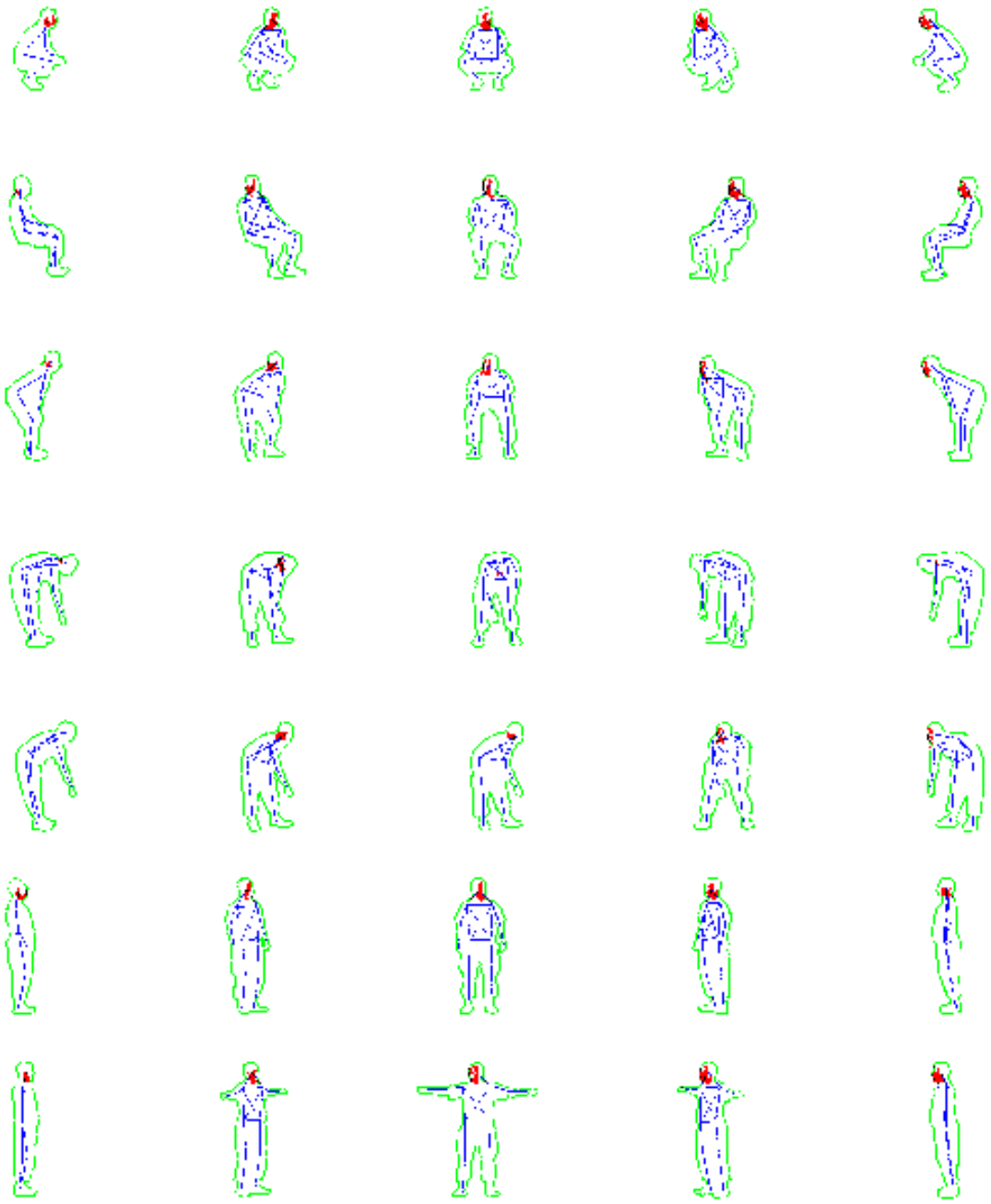


Figure 34. A conglomeration of outlines(green), internal skeletons(blue), and head-skin detections(red) for an imaged dismount. For each of these cases, the sun (illumination source) is on the left.

discussed, for each computer generated pose rendering, a separate rendering (or positioning frame) is created with the dismount featuring accented head skin. If the background for each of these separate renderings is benign, they are easily manipulated to convert the positioning frame into a silhouette of the dismount. Since the detector cues off of skin, in addition to maintaining information about the body outline and shape, it is also pertinent to track the shape and relative position of the head skin island within the silhouette. The black and white silhouette is then averaged with the black and white mask of the skin island to form a frame displaying black for background, gray for the body shape, and white for the head as seen in Fig. 35(d).

However, before any comparisons can be made, the bounding boxes (with width bb_{dx} and height bb_{dy}) surrounding the silhouettes are identified and are used to segment critical portions of the composite skin and silhouette frames. To reduce the complexity of the comparison, frames are resized to normalized 36×24 pixel "silhouette chips" A scale value, S, is chosen such that

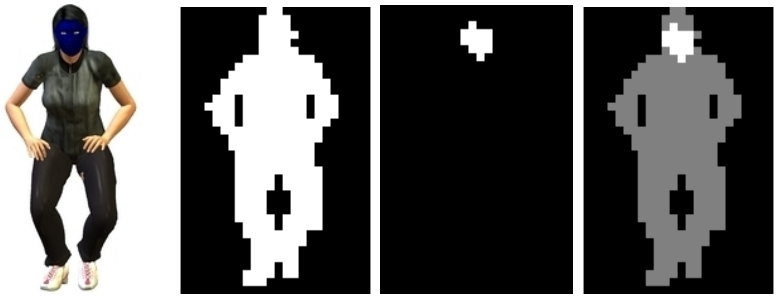
$$S = max\left(\frac{bb_{dx}}{24}, \frac{bb_{dy}}{36}\right). \tag{18}$$

This scale factor is used to determine the total amount of padding necessary in each direction (bb_{xp}, bb_{yp}) to achieve the desired "silhouette chip" ratio ($24 : 36$) as:

$$bb_{xp} = (S \times 24) - bb_{dx}, \tag{19}$$

$$bb_{yp} = (S \times 36) - bb_{dy}. \tag{20}$$

The extra offsets defined by bb_{yp} and bb_{xp} are evenly distributed between the two corresponding sides of the bounding box, to grow a new subsection that has the correct proportions. (In the event that the newly defined bounding box would be outside the range of the composite frame, new values for S, bb_{xp}, and bb_{yp} are computed to remain within the range of the image frame. Initially, the set of dismounts with extended arms appear to be extreme outliers as their disproportionately wide arm-spans greatly increase their vertical padding. In order to promote similarity, the dismounts with arm extensions are formed into patches using the full range of their height and slightly cropping the ends of their arms.)



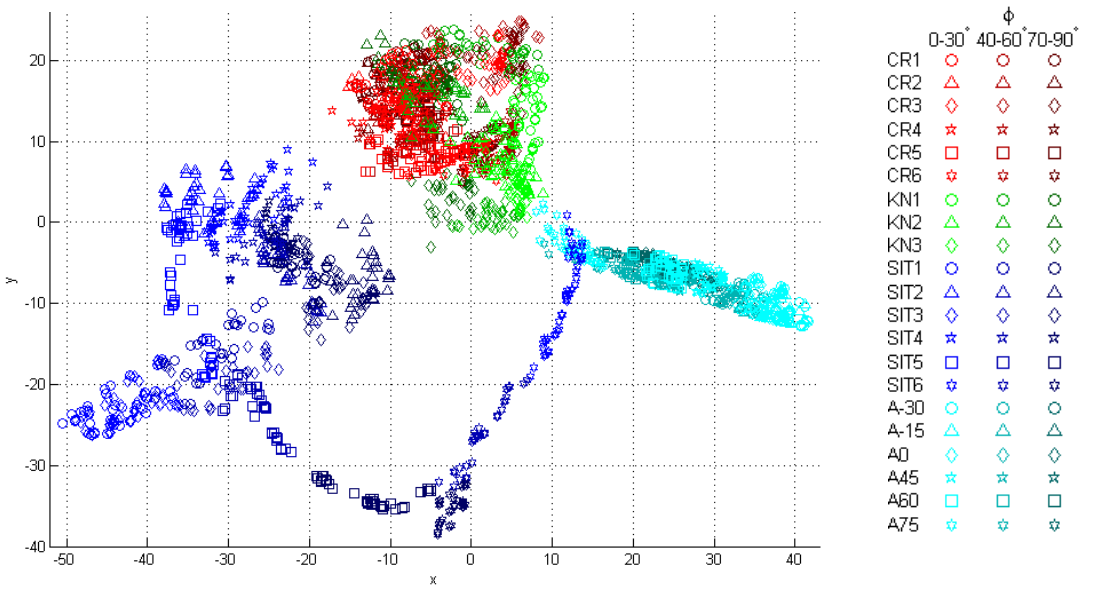
(a) (b) (c) (d)

Figure 35. (a) A "blue masked" version of a simulated dismount. (b) A 36×24 pixel silhouette corresponding to the dismount. (c) The location of the head skin island within the same 36×24 pixel boundaries. (d) Combination of the head skin island and silhouette information.

After the content contained within the bounding box is resized by $\frac{1}{S}$, these newly computed "silhouette chips" are then vectorized and clustered using ISOMAP [35] as recommended by [38]. This process performs two key functions. First, it measures the geodesic instead of linear distances between various samples, preserving more of a shape context [4] [6]. Second, ISOMAP allows for the desired degree of dimensionality reduction. As this vectorized chip is clustered, we choose to embed the high (864) dimensional data into a five dimensional manifold to allow for as many degrees of freedom in-between samples while still being representative of the data when viewed in only two or three dimensions. In the following figures, the dimensions displayed are those with the highest degree of variance. A 2D view of the ISOMAP clustering (Fig. 36) provides a general view of the groupings by pose type.

A 3D view of the ISOMAP clustered silhouette data in Fig. 37 reveals densely represented regions and approximately 10 groupings of data points. K-means is used to best represent these regions in the 5 dimensional ISOMAP space. After experimenting with various numbers of sample means, the use of 10 means seems to best represent the groupings without creating unnecessary clusters.

The placement of each cluster mean in Fig. 38 sheds light on the relationship of the various poses. The data points are grouped by which cluster they fall closest to as seen in Fig. 39 with partial commentary in Table 1. An extended display of the silhouettes and their assigned clusters is available in Appendix B.



(a) (b)

Figure 36. (a) Labeled data is seen with 2 of the 5 Isomap clustering dimensions. (b) A legend identifying varying crouching, kneeling, sitting, and arm extension poses as well as azimuth angle range is provided for convenience.

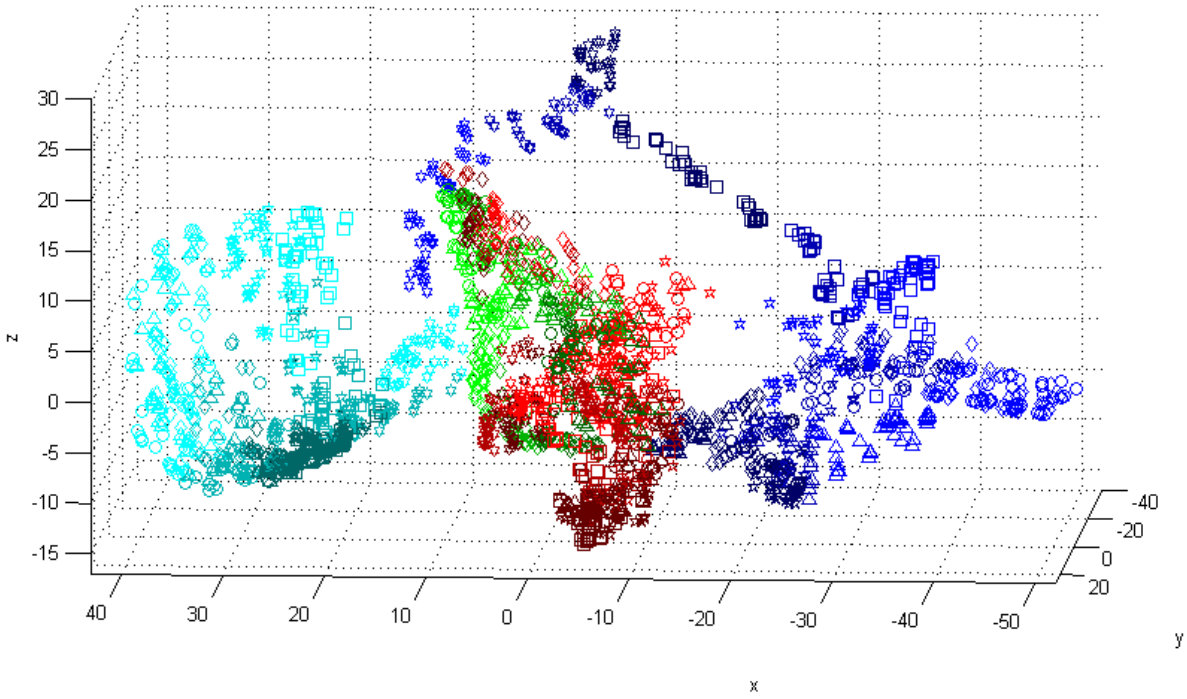


Figure 37. Labeled data is seen with 3 of the 5 Isomap clustering dimensions. The legend from 36(b) applies to this figure identifying data points by pose and azimuth angle range.

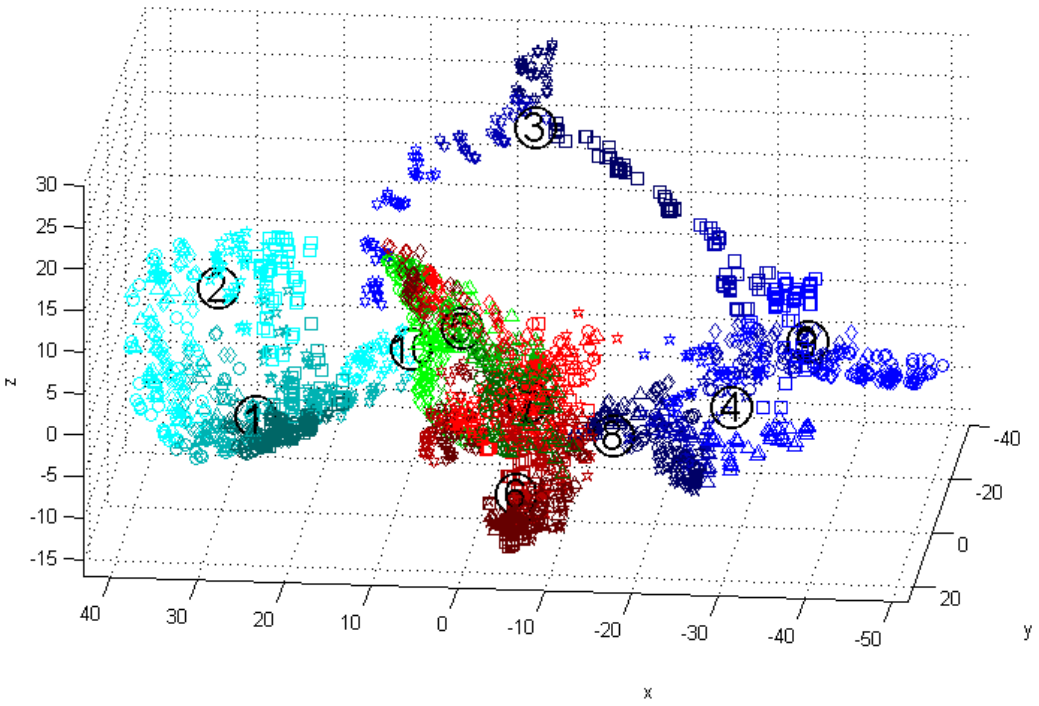


Figure 38. Labeled data is seen with 3 of the 5 Isomap clustering dimensions as well as 10 K-means learned representative mean vectors. The legend from 36(b) applies to this figure identifying data points by pose and azimuth angle range.

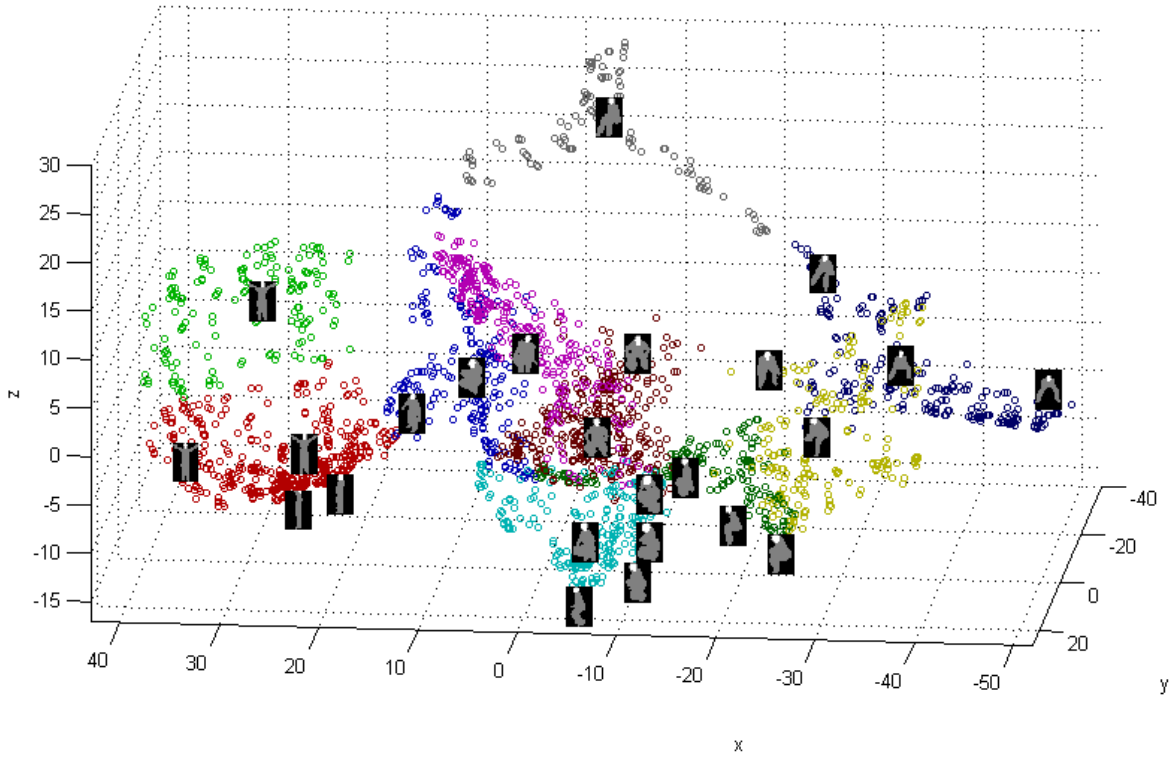


Figure 39. Pose data is seen with 3 of the 5 Isomap clustering dimensions grouped by color according to the closest mean vector. Selected silhouette chips are superimposed on this plot to provide contextual understanding of the clusters.

Table 1. Generalities about the poses and camera angle are observed for each cluster

Cluster	Poses
1	A tight distribution of standing poses seen from a $40 - 90°$ azimuth angle. Even though these dismounts display various degrees of arm extensions, they all look quite similar from an intermediate or side view.
2	Frontal views of standing poses showing a wide body outline.
3	Intermediate and side views of sitting dismounts (leaning back).
4	Frontal and intermediate views of forward leaning, sitting poses with spread legs.
5	Sitting and crouching poses where a knee or leg extends to the left of the dismount
6	Side views of all stages of crouching poses.
7	Forward views of crouching poses predominantly characterized by a wide placement of hands on hips.
8	Side crouching and sitting poses where a leg or foot protrudes below the rear.
9	Front and slight side angles of sitting dismounts with spread legs.
10	Several different frontal view of poses are included together predominantly featuring: higher crouches, kneeling, and one sitting pose all featuring erect back positions.

Judging from a hemispherical plot of prediction values from a representative crouching pose over 407 different camera angles, it is reasonable to expect a SVM to support a range of $40°$ horizontally. This assertion is also supported by the relative locations of the head skin islands in all of the chips. If crouching dismounts are examined, there appear to be three major distributions (in the vertical range from $[0, 30°]$): $[65, 90°]$ in the horizontal direction (as well as the mirror image $[-90, -65°]$), $[35, 65°]$ in the horizontal direction ($[-65, -35°]$), and $[0, 35°]$ in horizontal direction ($[-35 - 0°]$).

3.5.2.2 Generating Training Chips for New SVMs.

After the range of desired poses have been identified to build the new SVM, image patches with dismounts featuring these poses are collected. The Brooks [5] dismount detector was trained from an existing image repository featuring aligned and scaled images of pedestrians in a cityscape. This was sufficient for training the skin cued detector due to the fairly uniform positioning of the dismounts' heads in these images. However, in the absence of a repository with an extensive variety of poses that maintains the location of the

dismounts' faces, it becomes necessary to amass a new repository of desired poses.

In order to reduce redundancy in SVM generation, this thesis proposes to train SVMs based on patches with individuals facing forward or to the left. All other mirror images (with negative azimuth angles) are flipped horizontally. Besides necessitating fewer alternating SVMs, this modification of image patches provides a greater number of training samples for each SVM.

3.5.2.3 Modify the Cuing Mechanism and Incorporate into Multi-SVM Detector.

As additional SVMs are created, it is important to note that their positive training chips may present unique head skin island locations within the 96×48 pixel chip. As the training data for each new SVMs is accumulated, the average location of the top of the head skin island is calculated based off of the positive samples and is set as the Δv value (as described in Section 2.3.3.2); the variable Δw is used to describe the skin island centroid offset from the left side of the patch. SVMs built to detect side poses also carry a secondary set of parameters with a horizontally flipped Δw.

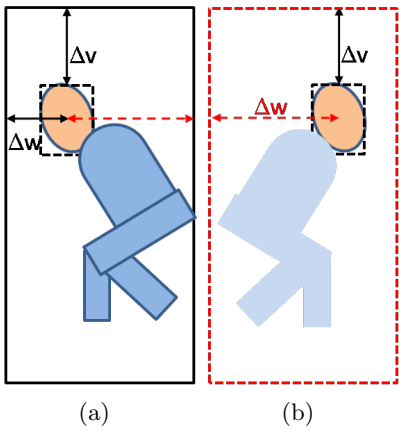


Figure 40. (a) The Δv, Δw distances are shown for an image patch as vertical and horizontal black lines. The red line indicates a secondary value for Δw. (b) A right facing dismount in a similar crouching pose would be captured in a patch utilizing the secondary value for Δw. Note: patches generated with the secondary Δw value are reflected horizontally before the remainder of the detection process continues (as seen in Fig. 41)

The detector system is modified to accommodate multiple SVMs as shown in Fig. 41.

As there is no known pose information associated with test skin islands, each SVM is tried on the same skin island, ultimately selecting the best match. The displayed multi-SVM detector shows the incorporation of two SVMs: a standing position SVM and a side crouching SVM (used twice for mirror images); however, the structure may be extended to incorporate any number of SVMs. The multi-SVM dismount detector is structured such that any given image yields N distinct skin islands with unique centroids. Each of these distinct skin islands has the potential of generating many different patches at various offsets and scales. SVMs apply their own Δv and Δw values to generate prediction windows. The number of prediction windows produced from each skin island may vary with Δv, Δw, and the edges of the image. Accordingly, the total number of patches generated from the ith set of Δv and Δw parameters can be defined as M_i.

Fig. 41 demonstrates how the detector attempts two different patch generation schemes for use with the side crouching SVM, then flips the patches generated with the secondary set of parameters. The next stage in the cascaded dismount process transforms the image patches into HOG features. HOG features are then classified with their respective SVMs to receive prediction values. The prediction value yielding the greatest result (above the detection threshold) is preserved and is used to identify the dismount (suppressing any non maximum detections resulting from the same skin island). The multi-SVM detector tracks which SVMs are chosen to identify the dismount, supplying additional information to the user. Fig. 41 identifies K standing dismounts and L dismounts in a side crouch.

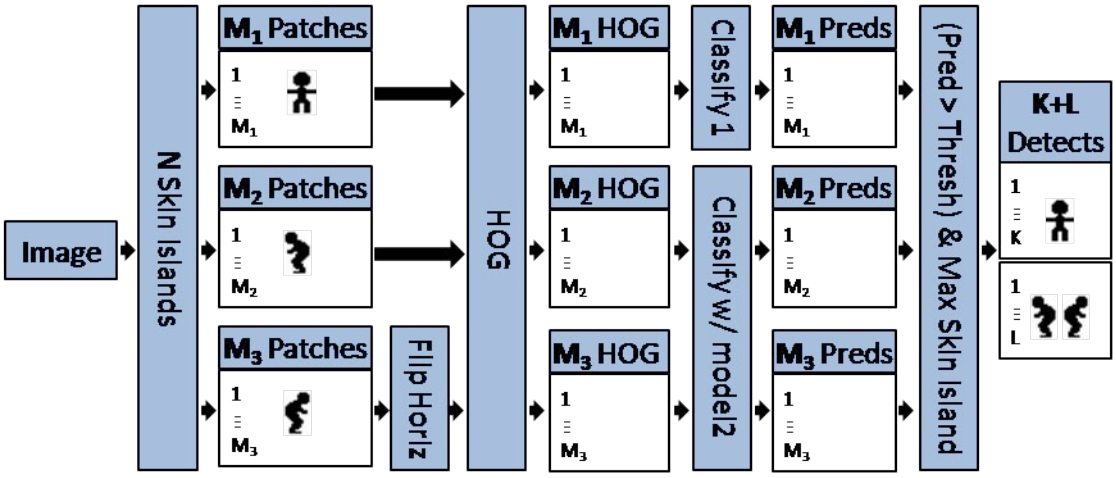


Figure 41. Each SVM applies their own set of Δv, Δw parameters to form prediction windows around around skin islands generated from an image. The resulting HOG features are classified with their corresponding SVMs to obtain prediction values for each patch. Maximal predictions (above the threshold) are reported for each skin island.

4. Experimental Results and Analyses

This chapter documents the specific testing processes as well as the experimental results obtained while analyzing improvements suggested throughout this thesis. First, a description is given of the data sets used during the training and testing processes. Second, a brief description regarding the use and scoring of a particular data set is presented. Third, the impact of the improvements to the Brooks [5] detector are individually discussed. Fourth, specifics are detailed regarding the training and testing of the SVM for crouching poses. Finally, the effectiveness of the completed multi-SVM dismount detector is discussed.

4.1 Data Sources

Three primary data sources are used throughout this thesis: a set of panchromatic visible imagery from [13], a collection of computer modeled dismount poses developed in this thesis, and a set of multispectral imagery obtained from the Peskosky [31] imager.

4.1.1 Daimler Benchmark Training Set.

Selections of the Daimler Benchmark set are used in part for the training and testing process of dismount detectors in this thesis. The data set is a collection of panchromatic visible imagery provided by [13] that features 15,660 aligned and scaled image patches centered around pedestrians in a cityscape, as well as 6,744 images absent of any dismounts (for negative training). The Daimler Benchmark set was sufficient to train the Brooks [5] detector due to the fairly uniform positioning of the dismounts' heads in these images. However, in the absence of a repository with an extensive enough array of poses that maintains the location of the dismounts' faces, it becomes necessary to amass a new repository of desired poses when examining non-standing dismounts.

4.1.2 3D Generated Data.

As mentioned in Section 3.3.1, 2D color images rendered from realistic human 3D models present a useful source of testing and training data. The DAZ Studio^{TM}3D modeling

software [18] is used to amass a set of over 25,000 (500×350 pixel) renderings, each featuring a dismount with a different urban background scene from a variety of city models. In addition to the variations in the dismount's pose and camera aspect angle, the dismount gender, body proportions, and clothing are also varied to include as much variety as possible. The positive set includes 56 sets of distinctive poses viewed over 407 labeled camera angles, with each set accompanied by an image frame for locating the position of the dismount and to simulate the position of their head skin island. These 56 sets include 25 crouching, 15 sitting, 3 kneeling, and 16 near standing pose sets (including variations in arm position, side bends, and walking stances). Hundreds of other pose sequences are produced from limited camera angles. A set of 151 480×640 pixel images is also produced that features a diversity of backgrounds and objects that are free from the presence of dismounts for the purposes of negative training.

The positioning frames corresponding to each image in the positive set are used to extract the locations of the bounding box around the dismounts (as in Section 3.5.2.1) as well as the location of the head skin island as derived in Eqn. 17.

Instead of forming tight 36×24 pixel silhouette chips as in Section 3.5.2.1, the bounding boxes around the dismounts are used to center the individuals into a 96×48 pixel patch, including a 12 pixel border on top and bottom.

4.1.3 Live Data Collections.

The Peskosky [31] imager is utilized in this research effort (over the course of five separate data collections) to gather 285 images of dismounts in various poses representative of typical human motion. These poses are roughly grouped into seven pose categories (listed in order of pose height): standing, bending, crouching (high), sitting (forward), sitting (back), kneeling, and crouching (low). The imagery is taken from a range of angles of elevation between $[0°, 6°]$ and azimuth angles $[-90°, 90°]$. The poses are further broken down by approximate azimuth angle and fitted into the closest of three bins: frontal, mid, and side views. The number of poses fitting into each subgroup in shown in Table 2. The five distinct data collections are imaged during different times of day with various solar conditions and

diverse backgrounds. The "complete set" is used for most stages of the detector testing; however, a subset of poses is also used to assess improvements in the crouching SVM. The "potential crouching" subset includes 93 non-standing poses such as sitting, leaning, bending over, kneeling, and crouching, that may be detected by the side crouching SVM.

Table 2. The live collection data is grouped by pose and approximate camera view.

	standing	bending	high crouch	sit forward	sit back	kneel	low crouch
Front	22	6	4	4	9	5	7
Mid	26	15	11	9	12	12	15
Side	30	27	20	10	15	12	13

4.2 Scoring Live Data Testing

The live data set is used to test three different dismount detectors: the Brooks [5] detector, the improved Brooks [5] detector described in Section 3.1, and the multi-SVM dismount detector (offering an additional SVM for side crouching poses). The Brooks [5] detector is used as a performance baseline as it was previously shown to perform on par with [13]'s premier HOG based dismount detector.

As all three detector systems are tested on the complete image set, the original images are overlaid with prediction boxes suggesting the location of a dismount. For scoring purposes, these prediction boxes are hand labeled, identifying each box as containing a dismount or a false alarm. Since the poses do not appear in fixed proportions and are not centered around the head skin island, there is wide variety in the relationship of displayed prediction boxes to the location of the actual dismount, necessitating a refined definition of a positive detection. For these purposes, a prediction box is defined as a positive detection if it contains a majority of the body: the face and trunk of the dismount (allowing the cropping of the arms or lower legs), offering sufficient visual data to indicate the location of the individual in the image. The remaining prediction boxes are labeled as false alarms and are used to evaluate the dismount detector's ability to reject erroneous portions of the image.

4.3 Skin Island Based Improvements to the Brooks [5] Detector

The use of multispectral skin detection provides a richness of data that is beneficial for providing context to the image scene. The size and location of head skin islands provides key information to aid in achieving better detections while further rejecting false alarms.

4.3.1 Suppression of False Alarms by Median Filter.

The application of a median filter to eliminate the effects of stray pixels, as explained in Section 3.1, is tested on the complete live data set to assess dismount detection and overall performance improvements. For this test, an 8×8 pixel median filter is utilized. (While median filters generally employ odd-sized filters, the use of an even 8×8 filter does not produce a problem due to the size and shape of skin islands, yet yields good experimental results.) A plot showing the probability of detection (P_D) verses the number of false positives per frame (FPPF) compares the Brooks [5] detector with and without the benefit of a median filter (Fig. 42).

The results are quite dramatic, as the median filter suppresses false alarms by more than one order of magnitude. However, the number of false alarms generated by the original Brooks [5] detector reveal significantly higher numbers of false alarms than expected. While the extent of the false alarm suppression is impressive, a portion of the success is likely due to imperfections in the imager and registration artifacts. In order to avoid any potential for bias in the remainder of the testing, a median filter is applied to all following baseline Brooks [5] detector results.

4.3.2 Additional Performance Gains Through Leveraging Skin Island Data.

In addition to the suppression of false alarms through use of a median filter, three additional methods of leveraging skin island data (as mentioned in section 3.1) are used to form an improved Brooks [5] detector to provide improved performance from the baseline Brooks [5] detector with median filtering.

First, the number of prediction windows generated from a single skin island are con-

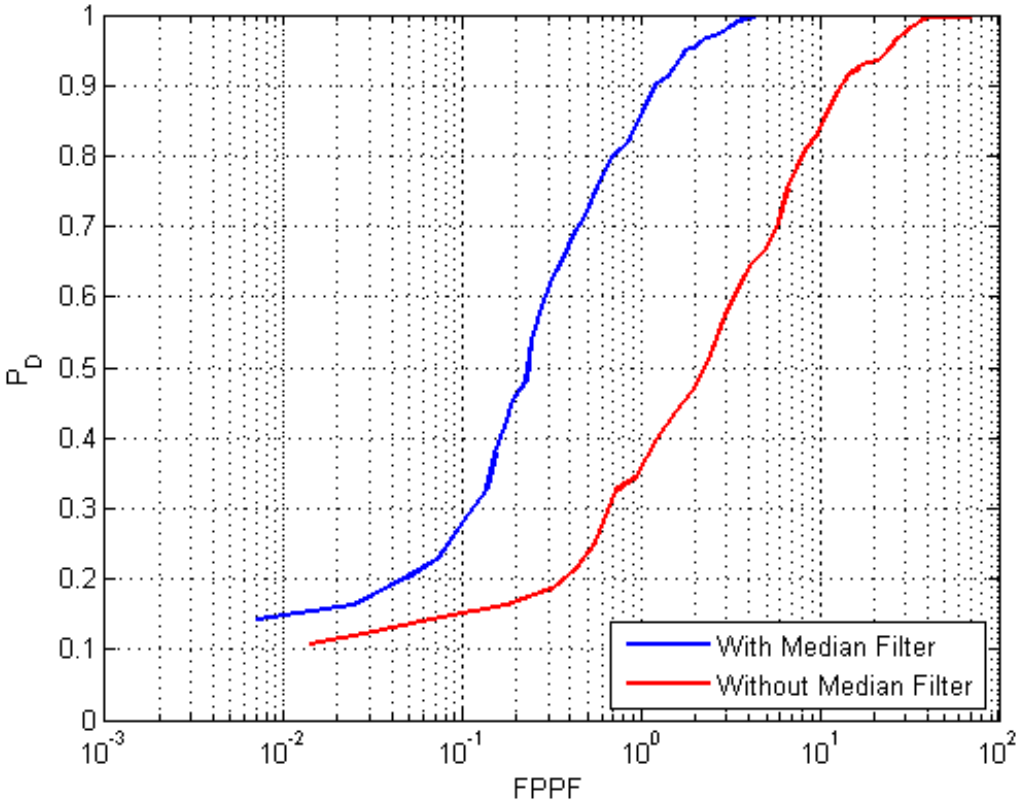


Figure 42. Probability of detection vs false positives per frame are displayed (with a logarithmic x-axis) for the Brooks [5] detector without median filtering (red) and with median filtering (blue).

strained by tracking the skin islands used in the cuing process. Limiting each skin island to produce only the prediction window that yields the maximum prediction score drastically reduces the number of false alarms and redundant detections especially prevalent in high probability of detection regions. This method is used instead of using the coverage statistic as explained in [5].

Second, false alarms are further constrained by correlating the skin island height with human anthropometrics. Minimum and maximum ratios of head to body height $R_{min} = 4$ and $R_{max} = 16$ are used in the improved Brooks [5] detector to restrict the dimensions of prediction windows generated around each skin island. These values accommodate a typical standing individual with a ratio of $1 : 8$ while allowing for a shorter or crouching individual, or even a standing individual who's full head height is only partially captured

in the detected head skin island.

Third, the range of possible detections are widened by allowing limited horizontal shifts based off of skin island width. The use of just 13 possible (negative and positive) offsets from the image patch center in increments of 0.25 times the skin island width allows for a visible increase in performance over the stand alone Brooks [5] detector.

The benefit of the three discussed practices of leveraging the available skin island data provide considerable performance gains as seen in Fig. 43. While the baseline detector enjoyed the added benefit of the median filter, the suppression efforts of the improved Brooks [5] detector provide approximately one order of magnitude of increased false alarm suppression for most probability of detection thresholds. The detection abilities of the improved Brooks [5] detector are curtailed however, at a 95% probability of detection. At higher ranges (with a sufficiently low detection threshold, η_t), the baseline Brooks [5] detector generates enough prediction windows to include all the dismounts in the live test set. However many of these detections occur in smaller portions of the prediction window and do not tightly fit the dismount (as addressed later in this chapter). Many of the larger prediction windows with lower prediction scores are suppressed in the improved Brooks [5] detector in favor of prediction windows with higher scores that originate from the same skin island.

4.4 SVM for Side Crouching Poses

4.4.1 Identifying Critical Training Patches.

Training patches are extracted from 20 of the 25 crouching poses in the 3D generated data set. Four sets are reserved for testing, while ISOMAP clustering reveals the last set of crouching poses as being too dissimilar to the remainder of the poses for comparison (closer related to a kneeling position).

Next, the 407 patches corresponding to each set are labeled according to camera angle. As the SVM is trained based on patches with individuals facing forward or to the left, images with negative azimuth angles are reflected horizontally.

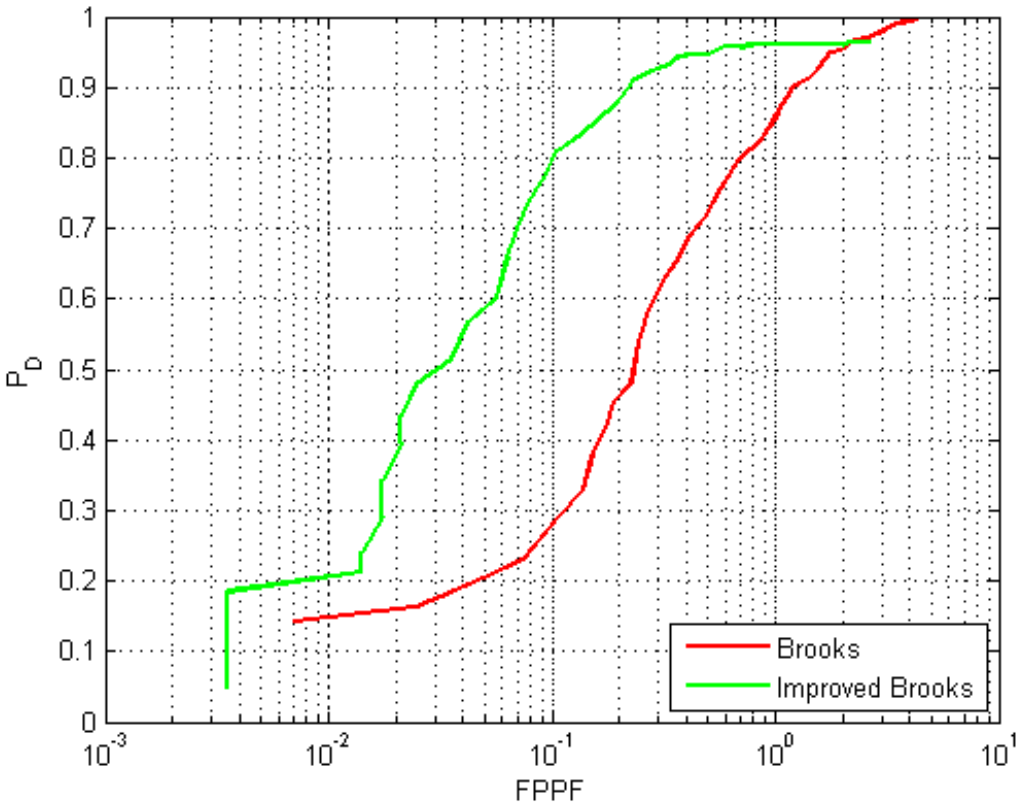
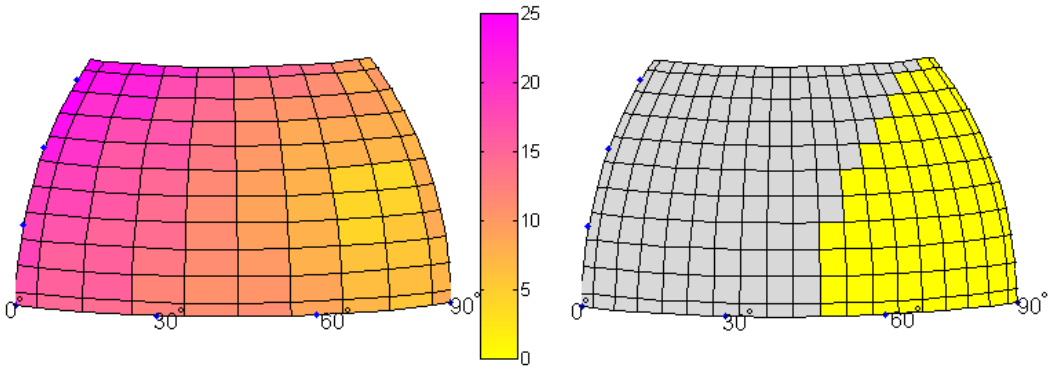


Figure 43. Probability of detection vs false positives per frame are displayed (with a logarithmic x-axis). Multiple detections from the same dismount are counted as false alarms.

The clustering of silhouette chips from Section 3.5.2.1 is used to identify poses that are spatially most similar, such as the tight grouping of crouching poses located closest to the sixth mean vector shown in Fig. 38. Individual poses are analyzed for trends over a sampling of camera angles. Fig. 44(a) identifies 110 camera positions of a representative crouching pose and displays the total distance (in ISOMAP space) between each pose and the sixth mean vector (in a hemispherical plot). (Additional poses are analyzed in Appendix C.) The intuition from Fig. 44(a) leads to an interpolated distribution of 76 camera angles (with positive azimuth angles) that consistently produce visually similar silhouettes (shown in Fig. 44(b)). Images of crouching poses over this set of camera angles (as well as corresponding negative azimuth angles) are selected for training a new SVM for side crouching poses (henceforth the side crouching SVM).



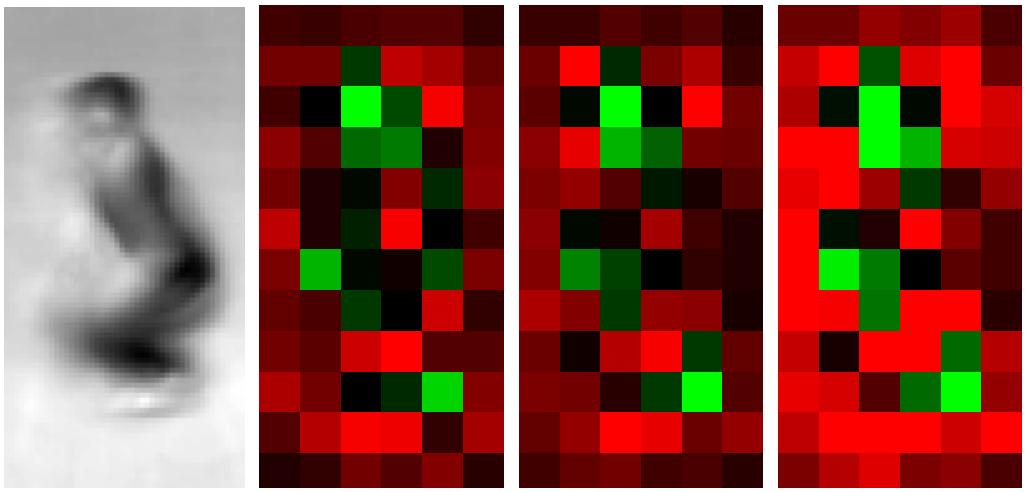
(a) (b)

Figure 44. Hemispherical plots are displayed over an azimuth angle range of $[0°, 90°]$ and elevation angle range of $[0°, 50°]$. Blue dots along the angle of elevation at $15°$, $30°$, and $45°$ are shown for convenience. (a) The similarity of a representative crouching pose to the closest mean vector is shown in a hemispherical plot by the Euclidean distance from "mean vector 6" to each of the 110 camera angles in five dimensional ISOMAP space. The color scale is fixed from 0 to 25 . (b) An interpolated distribution of 76 camera angles (yellow) represent closely related views of crouching poses. Camera angles represented in gray are generally insufficiently similar.

4.4.2 Training the Side Crouching SVM.

There are 3040 patches accumulated from the available 3D generated data set used for training the side crouching SVM. The various crouching positions are seen overlaid on top of each other when the training patches are spatially averaged as in Fig. 45(a). The dominant figure in the center represents a mid crouch; the ghosted outlines of a lower crouch, as well as several other positions are visible in the surrounding area. Training of the SVM requires negative examples to typify the class of objects to reject. Negative patches (3040 in total) are randomly selected from the Daimler Benchmark and computer generated negative sets of imagery. A Matlab® adaptation of SVM-Light [20] with a third degree polynomial kernel is then used to train the new SVMs. The first round of training produces a SVM with weights represented in Fig. 46, with an expanded view of the combined weights by cell in Fig. 45(b), which shows strong weights to align with the edges of the crouching silhouette.

New random selections from the the Daimler Benchmark and 3D Generated negative sets are accumulated until 3040 more samples are found that register as false alarms when tested on the first iteration of the SVM. The new negative patch set indicates deficiencies in the hyperplane defining the decision surface. Accordingly, a tighter decision surface and



(a) (b) (c) (d)

Figure 45. (a) The average of the positive training ships (b) Autoscaled SVM weights after the first round of training (c) Autoscaled SVM weights after the second round of training (d) SVM weights after the second round of training using same scaling factor as in (a)

a more precise SVM is formed when the new set of negative patches is used in a second iteration of training. The resulting SVM weights are seen in Fig. 47, with an expanded view of the combined weights by cell in Fig. 45(c). The expanded view indicates a tightening of weights, especially around the head and thighs. Due to processing constraints, only two rounds of iterative training are performed.

4.4.3 Crouching SVM on Software Generated Data.

A total of 354 new crouching patches from the 3D generated data set as well as the same number of negative patches (originating from the Daimler Benchmark and 3D generated data sets) are used to test the second iteration of the crouching SVM. A receiver operating characteristic (ROC) curve is calculated to examine the probability of detecting a dismount versus the probability of registering a false alarm. Data points on the ROC curve are established by varying the prediction threshold of the detector and counting the number of detections and false alarms over a sample set for each threshold value. The resulting ROC curve shows complete separation of the negative and positive sets, establishing 100% detection with no false alarms. The SVM is then tested on a progressively harder set of imagery composed exclusively of patches that register as false alarms when tested on the

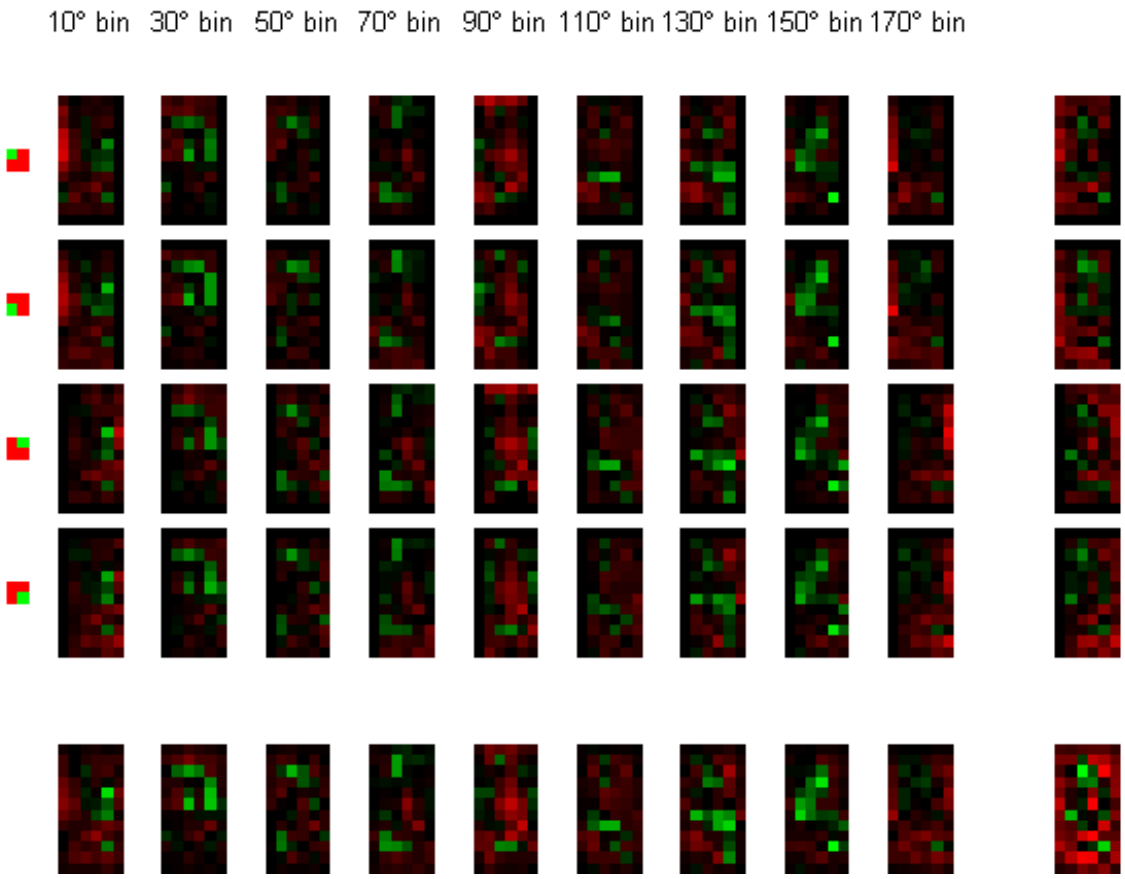


Figure 46. The first iteration of the side crouching SVM. A spatial mapping of SVM weights (corresponding to HOG features) is divided into orientation bins across the top and originating block position down the left. The block position for each row is indicated on the left as a green square occupying one of four positions on a red background. In the remainder of the figure, brighter shades of green corresponds to larger values. For simplicity in viewing, the 36 frames can be averaged across orientation angle (far right column) or originating block position (bottom row). The bottom right frame displays an averaged cell magnitude representation of the SVM weights.

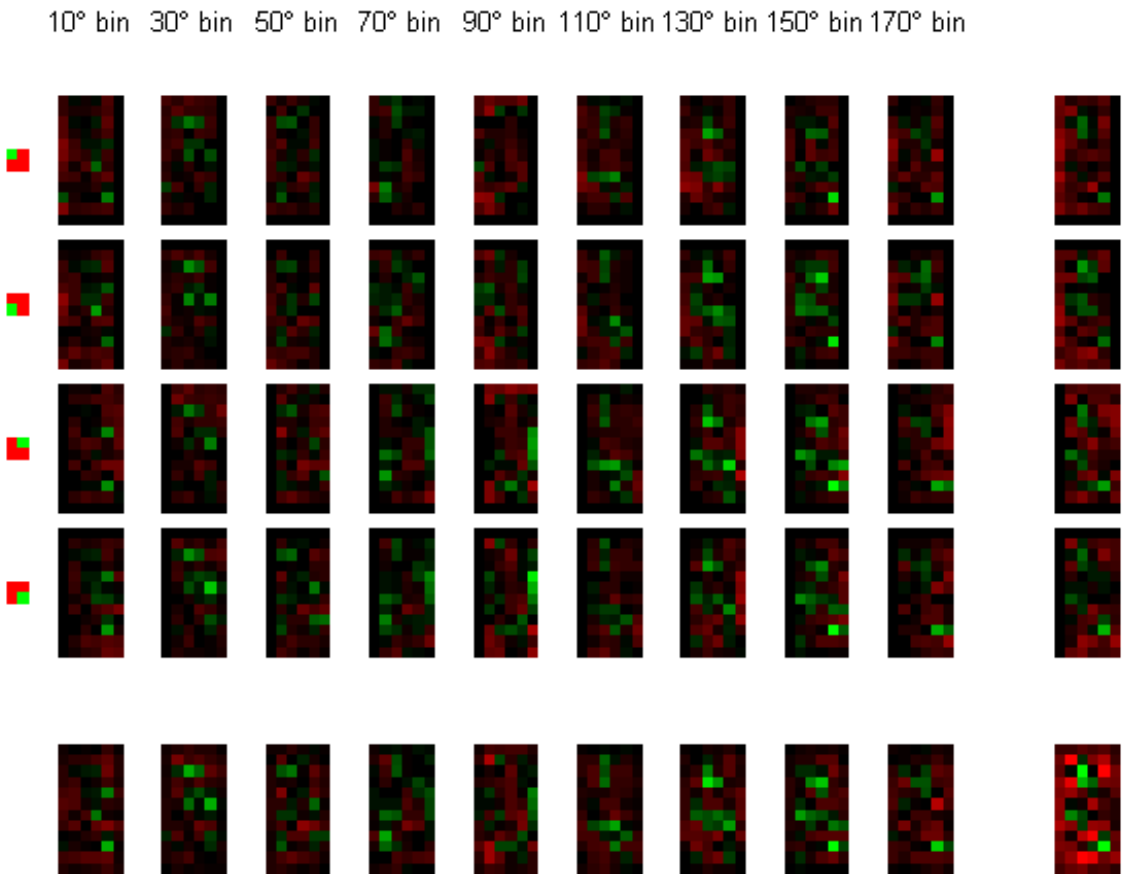


Figure 47. The second iteration of the side crouching SVM. A spatial mapping of SVM weights (corresponding to HOG features) is divided into orientation bins across the top and originating block position down the left. The block position for each row is indicated on the left as a green square occupying one of four positions on a red background. In the remainder of the figure, brighter shades of green corresponds to larger values. For simplicity in viewing, the 36 frames can be averaged across orientation angle (far right column) or originating block position (bottom row). The bottom right frame displays an averaged cell magnitude representation of the SVM weights.

previous iteration of the SVM. The resulting ROC curve similarly shows complete separation with all known samples being distinguishable from the negative samples. (Due to the trivial nature of these ROC curves, no figures are inserted.) Achieving a perfect ROC curve for the first set of negative data is understandable as the sample size is small and there is likely a large distinction between a crouching dismount and a random grab from an image. A possible explanation for the perfect ROC curve for the progressively harder negative set lies in the source of the image patches. While no patches are re-used between training and testing, the negative testing set is defined by the same limiting parameters as the negative training set used to create the second iteration of the SVM. Another key factor to the perfect performance are the uniformly high prediction scores (above 0.93) for all the positive testing patches.

The second iteration of the side crouching SVM is additionally tested on the training data, this time utilizing the cuing process of the dismount detector with a $\Delta w = 16$. As the cuing process is added, the prediction scores drop dramatically, but still provides enough separation to clearly distinguish between side crouching poses and samples belonging to the negative set.

4.4.4 Crouching SVM on Live Data.

When the SVM for side crouching poses is tested on data from the live collection set, it produces prediction windows that cue correctly around crouching dismounts, however it yields low prediction scores. While the detector still provides separation between members of positive and negative sets, the low prediction values are indicative of an anticipated problem: narrowness of the training data. Due to the sparsity of training samples, the SVM becomes over fitted and does not support the full range of diversity present in the training set. As the amount of available training data remains a limitation, this thesis compensates by utilizing the coarser, less restrictive first iteration of the crouching SVM to establish results.

Correct cuing of the detector plays a large role in prediction strength. The issue of triggering becomes an even harder problem for the case of crouching poses, since the head

is not in a centered position as in standing poses, but appears in a wide range of locations with respect to the body. Extensive testing is performed on the potential crouching pose set to determine good values for Δv and Δw. A good vertical offset of $\Delta v = 15$ is experimentally determined (results not shown here). Fig. 48 displays the effects of four different values of Δw over the "promising crouching" subset of poses that ultimately shows a preference of $\Delta v = 16$. A (Δv, Δw) parameter set of $(15, 16)$ is subsequently used when the crouching SVM is incorporated into the multi-SVM detector.

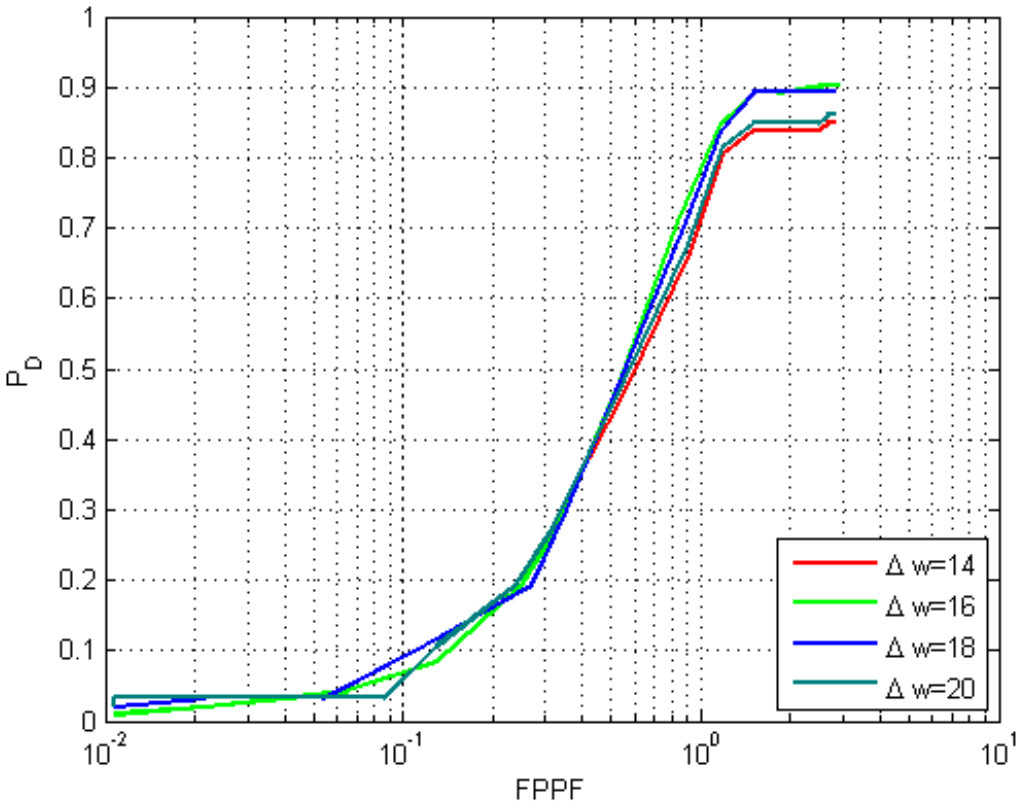


Figure 48. Probability of detection vs false positives per frame are displayed (with a logarithmic x-axis). A trade-off study is performed to identify the best value of Δw for crouching poses.

4.5 Effectiveness of Multi-SVM Dismount Detector

In order to assess the accuracy of the new cascading dismount detector, the number of correct detections, missed detections, and false positives are tracked over the live testing

set. These results are compared against the Brooks [5] detector (with the added benefit of a median filter) and the improved Brooks [5] detector. Three comparisons are made based off of differing philosophies regarding detections. Section 4.5.1 represents a philosophy of not penalizing multiple detections originating from the same dismount whereas Section 4.5.2 treats the additional detections as false alarms. Section 4.5.3 implements a stricter detection definition as the three detector systems are assessed.

4.5.1 No Penalty for Multiple Detections.

Some testing proceedures in [5] are based off of a user's indifference toward multiple detections cued around the same dismount, arguing that these additional prediction windows can be easily mitigated by a user. In Fig. 49, the probability of detection on a linear axis is compared against the average number of false alarms per frame. As only one dismount is located in each image, a maximum of one detection from each image is counted toward the total probability of detection. As the multiple detections can be visually excluded with minimal effort, they not included in the number of false alarms.

The multi-SVM detector shows drastic improvements to the Brooks [5] detector, while capturing many of the detections missed by the improved Brooks [5] detector above the 95% P_D range. The multi-SVM detector also features performance gains over the two other detectors in the 0.05 FPPF range and under.

4.5.2 Penalize Multiple Detections.

The second comparison features a user's preference against the additional visual data from multiple detections. As the extra windows represent additional confusers limiting the effectiveness of the end user, Fig. 50 treats them as false alarms. The results of the second comparison are similar to that of Fig. 49, only showing a general detriment to the effectiveness of the Brooks [5] detector, while the other two detectors perform as previously.

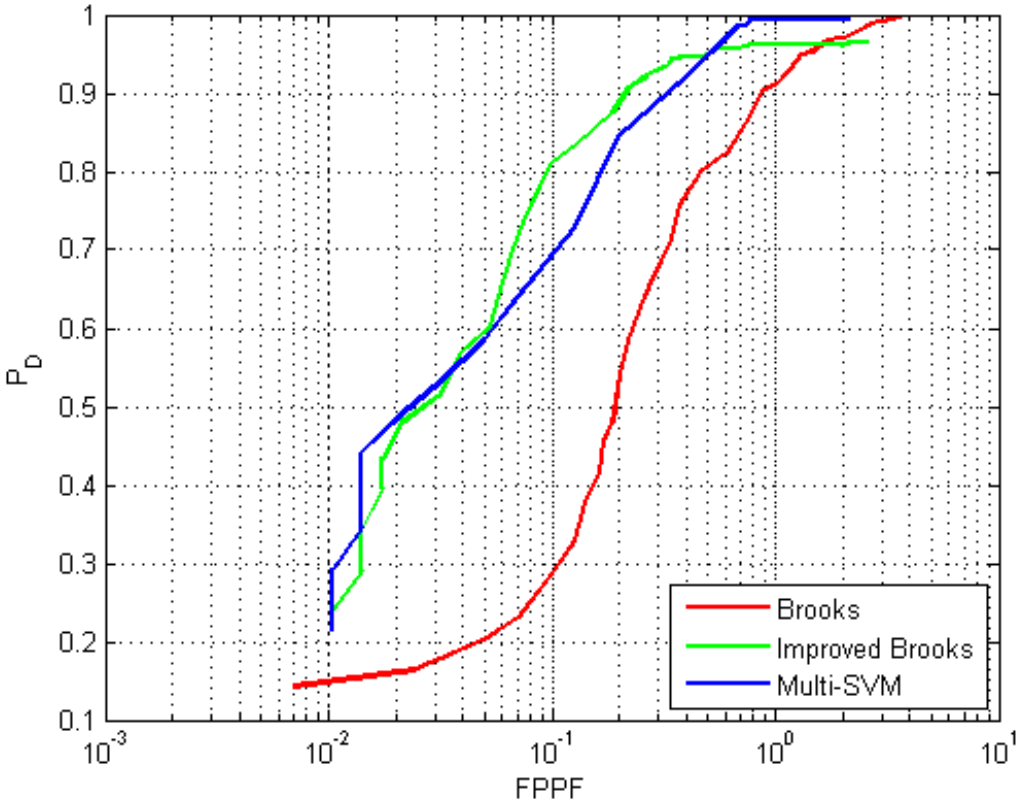


Figure 49. Probability of detection vs false positives per frame are displayed (with a logarithmic x-axis). Multiple detections from the same dismount are discarded.

4.5.3 Strict Detection Definition.

Discussion of the first comparison mentioned inexactness in the prediction windows generated from the Brooks [5] detector. Hence, a new term: "strong detection" is defined, requiring prediction windows to fully including dismounts (allowing only for cropping of the hands and feet). Additionally, strong detections are defined to demand a central position of the dismount within the prediction window, with less than a 20% border in any direction. Fig. 51 displays the probability of strong detections versus the number of false positives per frame. The stringent requirement of strong detections limits the Brooks [5] detector's maximum P_D to 81% even with 7 FPPF. The improved Brooks [5] detector is able to achieve a higher maximum P_D at a much lower FPPF rate. The multi-SVM detector is able to achieve a higher probability of maximum detection while offering the best false

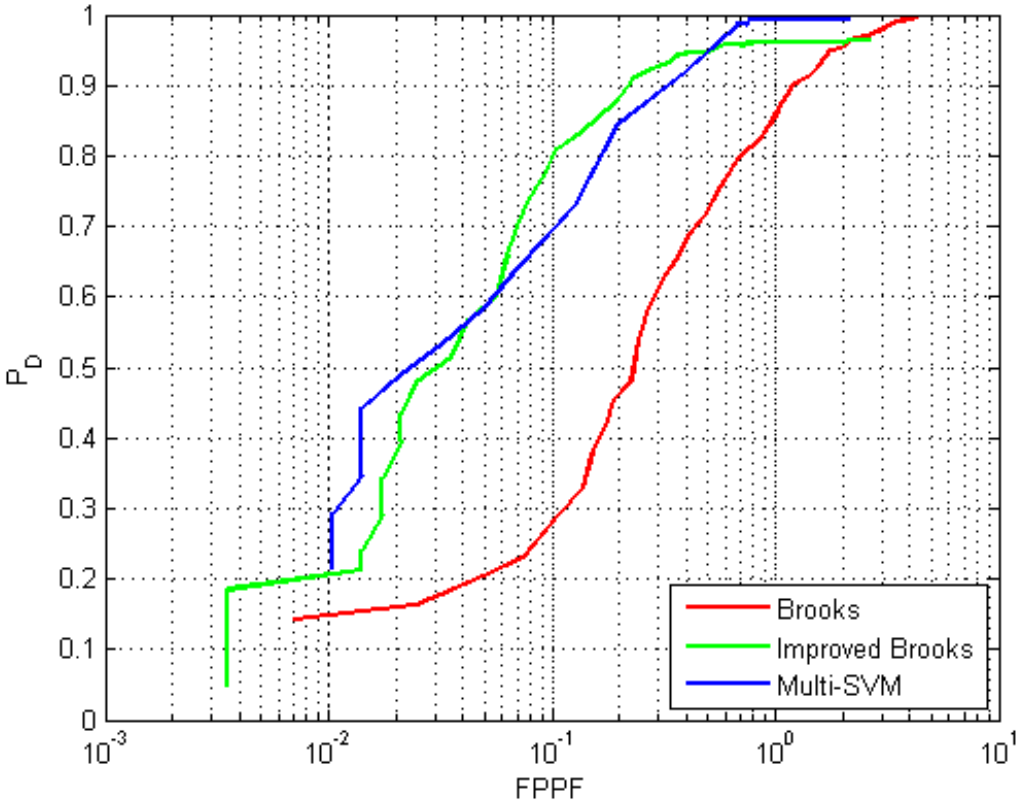


Figure 50. Probability of detection vs false positives per frame are displayed (with a logarithmic x-axis). Multiple detections are treated as false alarms.

positive rejection rate in all but the midrange of threshold values.

4.6 Effectiveness of Multi-SVM Detector on Target Pose Groups

The performance of the multi-SVM detector is further evaluated on a subset of the live test data featuring a 50/50 split of standing and crouching poses. After this comparison, sample detection windows from the baseline and improved Brooks [5] detectors as well as the multi-SVM detector are displayed to illustrate the detection improvements witnessed in live data examples.

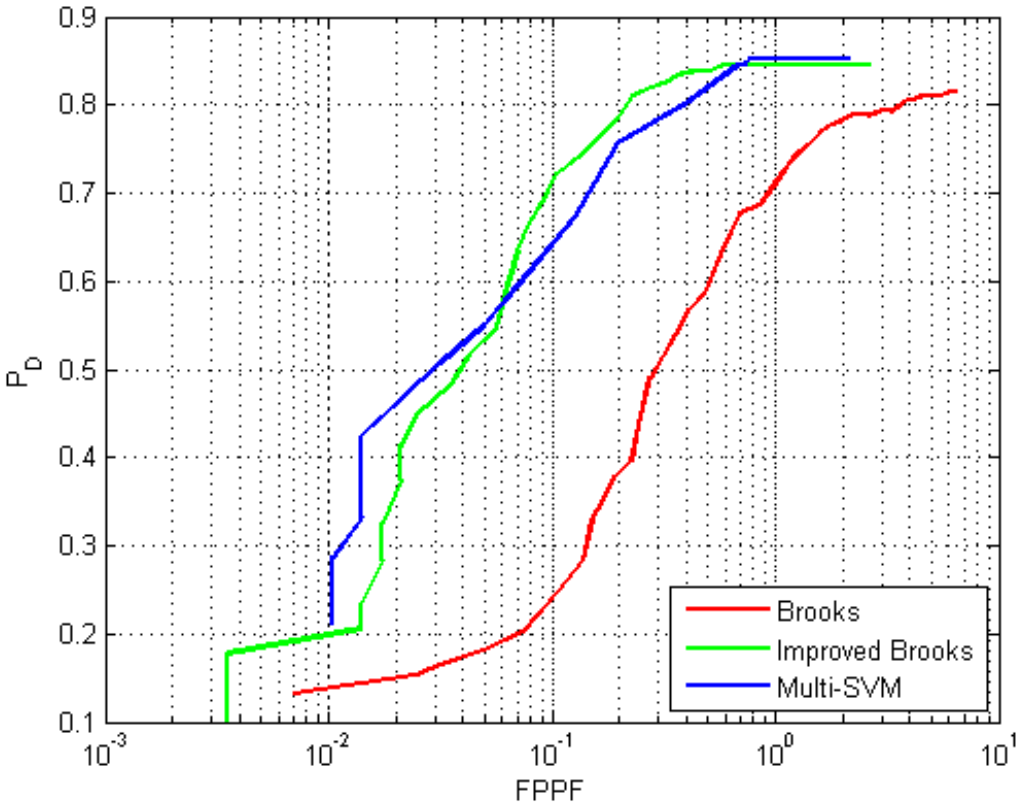


Figure 51. Probability of strong detections vs false positives per frame are displayed (with a logarithmic x-axis). Multiple detections are treated as false alarms.

4.6.1 Target Poses With Penalized Multiple Detections.

The analyses from Section 4.5 over the full live data set indicates several P_D regions where the multi-SVM detector generates more false positives per frame than the improved Brooks [5] detector. Additional false alarms occur because supplemental SVMs provide new ways for each skin island (even those not pertaining to dismounts) to yield detections above the threshold η_t. The impact of these additional false alarms is generally mitigated by the improved detection of the new target class of poses. However, a minority representation of the target class of poses may not allow the performance gains to be fully realized. A 50/50 split of standing and near crouching poses provides an opportunity to evaluate the multi-SVM detector on pose groups it is trained to detect. In Fig. 52, the multi-SVM detector is seen to offer superior performance to the baseline and improved Brooks [5] detectors as it

achieves a higher P_D at each FPPF rate.

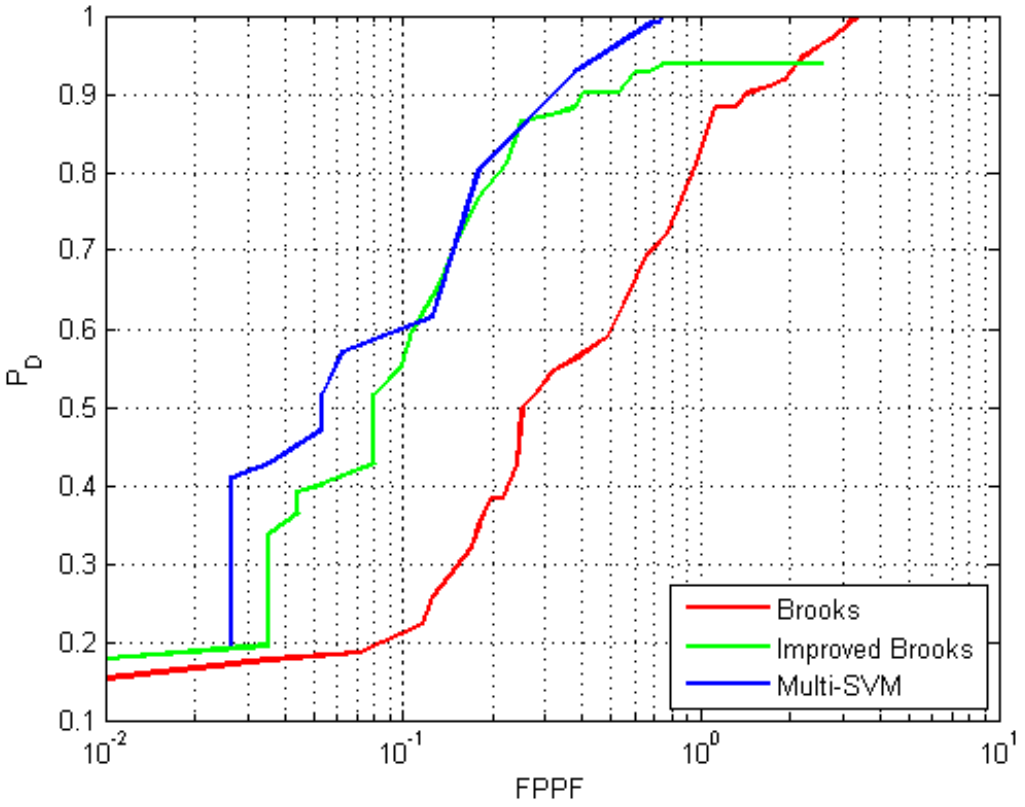
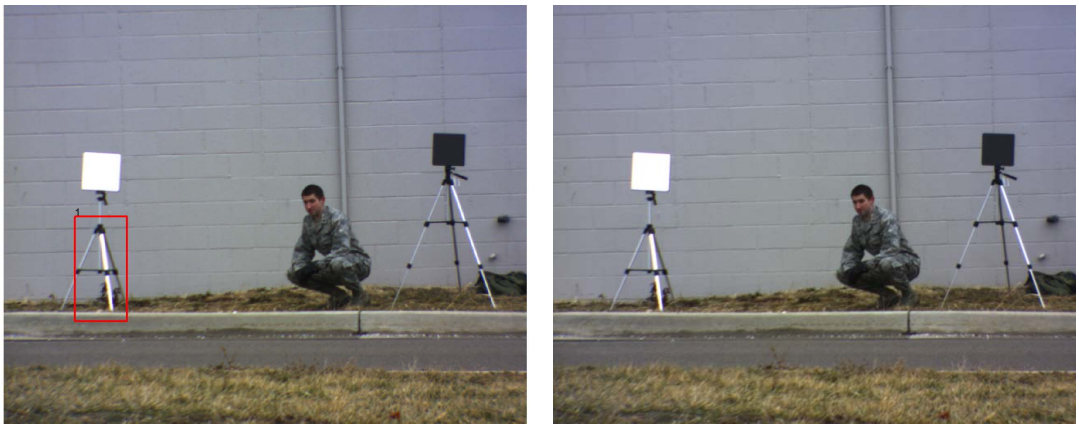


Figure 52. Probability of detection vs false positives per frame are displayed (with a logarithmic x-axis). Multiple detections are treated as false alarms.

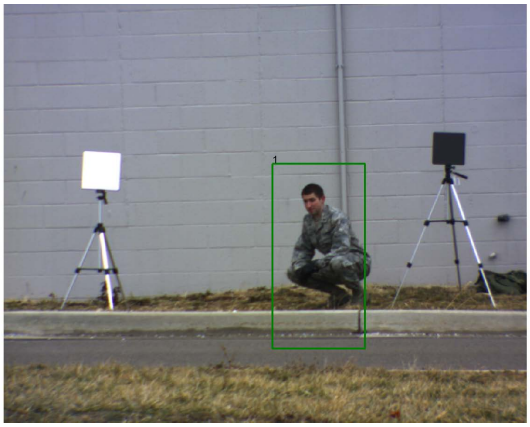
4.6.2 Visual Output of Detector Systems.

The Brooks [5] detector and the multi-SVM detector are configured to visually display boxes around prediction windows yielding values above the detection threshold η_t. The visual outputs for all three detectors are shown in Fig. 53 for a representative crouching pose. The Brooks [5] detector in Fig. 53(a) is unable to detect the crouching dismount, however, it generates a false alarm around the vertical structure of a tripod. The improved Brooks [5] detector suppresses the false alarm from the tripod, but does not detect the crouching dismount. The multi-SVM detector in Fig. 53(c) easily identifies the dismount in a side crouching pose, while rejecting any potential false alarms. These visual outputs illustrate an example of the added abilities of the multi-SVM detector to detect additional

poses while providing for increased false alarm suppression.



(a) (b)



(c)

Figure 53. (a) A false alarm around a tripod is generated by the Brooks [5] detector is outlined in red. (b) The improved Brooks [5] detector suppresses the false alarm from the tripod, but does not detect the crouching dismount. (c) The multi-SVM detector finds the crouching dismount outlined in green as well as suppresses the previous false alarm around the tripod.

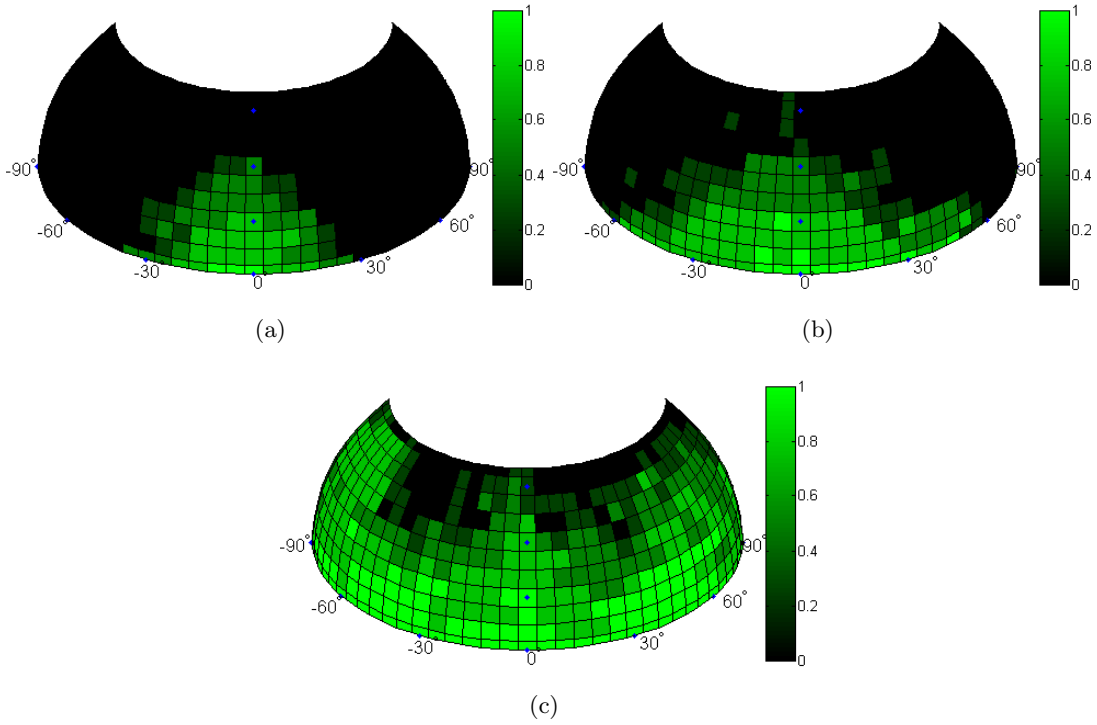
4.7 Additional Testing

Additional testing is performed to analyze the effectiveness of the multi-SVM over a wide sweep of camera angles. Four sets of computer generated crouching imagery reserved for testing are used to compare the multi-SVM detector to the improved and baseline Brooks [5] detectors.

As the only skin islands present in this testing properly align to the heads of the dismounts, no false alarms are triggered off of background objects. In order to reflect this lack

of confusers, thresholds for establishing detections are chosen for a low false positive per frame rate of 0.05.

Positive detections for each camera position are tallied and averaged to represent the probability of detecting a crouching pose in a hemispherical representation as seen in Fig. 54. The Brooks [5] detector indicates limited coverage for frontal positions at an angle of elevation below 30°. The ability of the improved Brooks [5] detector to apply logical horizontal offsets, and its lower FPPF rate lead to added detections as the azimuth angle extends to 55°. The multi-SVM detector performs best of all, as it features the lowest number of false alarms per frame and provides the added capabilities of detecting crouching poses from a side angle (e.g. 60°, 90°). In these examples at a low FPPF rate, the multi-SVM detector, when only equipped with one additional SVM (for side angles), improves the overall ability to detect a crouching dismount over a $[-90°, 90°]$ azimuth, $[0°, 50°]$ elevation angle range from 7.99% to 55.65% (12.5% to 73.5% for angles of elevation 30° or below).



(a) (b)

(c)

Figure 54. Hemispherical plots display the probability of predicting crouching poses over 407 different camera angles. Blue dots along the angle of elevation are located at 15°, 30°, and 45° for convenience in reading the plot (a) Brooks [5] Detector (b) Improved Brooks [5] Detector (c) Multi-SVM Detector

4.8 Chapter Highlights

This chapter shows the dramatic success of leveraging available skin island information, as the improved Brooks [5] detector suppresses an order of magnitude of false alarms when compared to the baseline Brooks [5] detector. These savings are in addition to nearly one more order of magnitude in FPPF suppression from the median filter. The multi-SVM detector, when augmented with only one additional SVM and tested on a broad spread of dismount poses, is shown to provide improved performance over the baseline and improved Brooks [5] detectors, especially at high P_D values. The multi-SVM detector excels in the detection of crouching poses over a wide range of camera angles and improves the probability of detection 7-fold when compared to the Brooks [5] detector. These results indicate the viability of using computer generated data to train multiple SVMs for the detection of dismounts over a variety of poses and camera aspect angles.

5. Conclusions and Future Work

This chapter begins by presenting a summary of methods used, and conclusions obtained, as a result of this thesis. Next, avenues for further study and experimentation are identified for the continuation of this research effort. The chapter concludes by highlighting notable contributions to histograms of oriented gradients (HOG) visualization and the field of dismount detection.

5.1 Summary of Methods and Conclusions

The overarching goal of this thesis is to provide better tools, methodology, and insight related to the detection of dismounts in imagery. As such, the impact of image changes on HOG features is explored in order to identify limitations in a given detector system. A HOG-based skin cued detector from [5] is refined leveraging available skin information and is outfitted with an additional SVM to extend its detection coverage.

The first focus of this thesis analyzes the effect that articulations in dismount pose and changes in camera aspect angle have on HOG features. In order to assess the potential changes in HOG features, a novel visualization method is developed to display the entirety of any HOG feature in spatial context (arranged by orientation bin). As multiple sequences of motion are analyzed, HOG features are shown to depend on the position and angle of the edges of each shape (i.e., head, shoulders, arms, etc.) in the image. The role of these dependencies gains meaning in association with the support vector machine (SVM) used to classify the HOG features. The novel visualization method is used to display the HOG features scaled by the SVM weights to indicate components critical to prediction score formation.

After promoting an understanding of the relationship between pose, HOG feature, SVM weight, and prediction score, the Brooks [5] detector is examined to reveal limitations in its ability to detect dismounts in imagery. The limitations are characterized through testing with a group of image patches, constructed to represent complete ranges of motion and feature a series of tightly controlled adjustments in pose and camera aspect angle. For the

typical threshold value of 0, the Brooks [5] detector is capable of identifying dismounts in frontal poses when their body is centered and positioned immediately below their head, appearing as a solid trunk from head to feet. However, these detections often fail for dismounts with elevated arms and shoulders, contracted body height, or with widely spread legs. As the angle of elevation increases for these optimal cases, detections generally die off at 25°. The Brooks [5] detector further encounters difficulty for many side poses past a 60° azimuth angle, as the individual's profile generally does not fully fill the space occupied by their width. Additional modifications to the detector system are made, leveraging the position and size of detected skin islands. These improvements allow for intelligently spaced horizontal offsets to increase the likelihood of detections, and suppress false alarms by constraining the size of detection windows and limiting the number of detection windows generated from each skin island. These improvements reduce the number of false alarms from the original Brooks [5] system by over one order of magnitude.

This thesis shows how limitations and "holes" in a dismount detector's coverage can be rectified through the training of additional SVMs which are incorporated into a multi-SVM dismount detector. In specific, a range of closely related side crouching poses is identified and used to tackle poor detection regions in the Brooks [5] detector. Computer generated images of dismounts are used to train a side crouching SVM. The resulting SVM performs well on similar computer generated data, but experiences universally lower prediction scores when tested on live data. To compensate for the narrow base of training, a looser version of the SVM is utilized and incorporated into a multi-SVM dismount detector.

The multi-SVM dismount detector is tested on a set of 285 multispectral images obtained through the Peskosky imager [31] featuring a wide variety of dismount poses. The multi-SVM detector significantly outperforms the baseline Brooks [5] detector, improving the false alarm rate from approximately 1.77 to 0.53 false positives per frame at a 95° probability of detection. When the definition of detections is restricted to require a close fitting prediction window, the multi-SVM detector identifies an additional 5% of dismounts that would otherwise go undetected, regardless of the allowable number of false positives per frame. The multi-SVM detector as shown represents the added benefit of additional SVMs

and skin island-based false alarm suppression. It is anticipated that inclusion of increased training data and additional SVMs will increase the performance gains shown.

5.2 Future Work

There are multiple areas for further work related to this research effort. The need for additional data is presented, several suggestions are made for improving the existing detector system, and a related field of work is mentioned for future combination with skin cued detections.

5.2.1 More Robust Data Sets.

The limited availability of hyperspectral dismount data presented limitations in testing and training. An expanded library of dismount poses obtained from the intended operational detection system imager would better correlate testing and training data as well as provide increased diversity in pose. The product of an extensive hyperspectral dismount pose database would yield more robust detection results.

5.2.2 Improvements to the Cuing Mechanism.

The quality of the detection results are a product of the accuracy in each step of the detection process, however, two overarching factors are largely responsible: correct cuing and adequate SVM training data. Correct cuing is a problem in aligning training data as well as determining horizontal and vertical offsets necessary for the best formation of prediction windows. Cuing is a particular problem over wide ranges of camera angles, as a dismounts hair occupies differing portions of the head. An improved mechanism for aligning computer generated training samples could calculate a skin island location dependent on the 3D location of a dismount and the present camera angle. Similarly, test samples could be cued based off of a spread of horizontal and vertical offsets calculated to support the range of elevation and azimuth angles.

5.2.3 Better Models of Skin Detections.

The simulated head skin islands created to cue model generated data represent the full shape of the human face. However, a fully defined detection of the face is not often typical from real world data collection. A set of collected data for a specific illumination angle, shown in Fig. 34, highlights the best skin detections for each dismount in red. Distorted skin detection results occurred when the dismount was facing away from the sun (located on the left of the dismount). The work in [22] could be used to model skin based off of the position of the illumination source. The creation of more realistic skin islands could be of use in training as well as testing as it could be incorporated in the computer generated dismount models to improve the training set.

5.2.4 Additional SVMs.

This thesis provided the methodology for forming supplemental SVMs for a multi-SVM detector system. While constraints on available training data only allowed for the creation of one additional SVM, the training and inclusion of additional SVMs focusing on distinct poses, such as sitting, would likely result in significant performance gains.

5.2.5 Adjusting the Weight of HOG Features According to Object Shape.

Due to the sparsity of training samples, HOG features could be weighted by the expected shape of the dismount inside the image patch as described by [38] using the term: object weighted appearance model (OWAM). This method for prompting stronger gradients around the dismount is encouraging for future training.

5.2.6 Adjusting Individual SVM Thresholds.

Multiple SVMs present varying distributions of prediction values for dismounts and in scene confusers. In order to make the SVM prediction results comparable, it is possible to adjust the dynamic range of the prediction scores to similarly align distributions of detections and false alarms. This adjustment in the scale of prediction scores could similarly

be used to weight the confidence in a particular SVM and provide an additional parameter that could be tuned for a specific scenario.

5.2.7 Incorporation of Clothing Detection.

Work exploring the detection of clothing could be incorporated into the detector system to promote a "whole person detection". This idea fits in with other research in [8] who is interested in the detection of clothing in hyperspectral imagery. There is great synergy between these two components as areas of a dismount that aren't covered by skin are likely covered by identifiable clothing.

5.3 Contributions

There are relatively few existing methods to aid in the visual inspection of HOG features. This thesis provides a technique to take any HOG feature, extract components associated with the original cell structure, and display it in a spatial context arranged by orientation bin. The use of graduated colors within the orientation frames allows for a clear comparison between changes in orientation bins. This HOG visualization tool allows one to inspect a set of weighted HOG features to better understand limits of a particular SVM.

This thesis promotes the use of model generated data as a tool to assess or train a dismount detector. 3D models are ideal for simulating a wide spread of poses, as they can accurately represent human motion and appearance while offering a fine degree of manipulation and tuning. Complete ranges of motion can be generated as well as fine shifts in position that might be challenging to capture in live data collection. The use of simulated data over a full range of needed poses and camera angles can help to fill the gaps in collected field data. This is especially useful when actual field data is limited or challenging to acquire. Furthermore, we show that training on modeled data, while testing on real data is a viable option.

Most current dismount detectors provide detections for limited camera angles and are mostly tailored to support street level pedestrian detection with a minimal angle of elevation. The work in this thesis concentrates on the appearance of a dismount in not only multiple

poses, but also a full range of camera angles. The understanding and characterization of how HOG features and detections (for a given SVM) change with angle helps to promote a 3D view of HOG features. This exploration of a wide range of camera angles is especially pertinent for an Air Force audience concerned with imaging platforms requiring support over a wide range of aspect angles.

This thesis work has a significant impact for a given dismount detection system as methods are shown to identify the types of detections that correctly register and which dismount pose configurations will evade detection. As the limitations in camera angle are identified, the doctrine and policy regarding the use of the imaging systems can be adjusted. For example, knowing that a current system is only capable of strong detections with an elevation angle less than 40° establishes the effective field of view of a low flying UAV with a dismount detecting platform. This same information could be used similarly to determine maximal effectiveness when mounted on a building.

Once the limitations of a SVM are made known, it is also possible to gather additional training data to retrain the SVM or to implement an additional SVM that can be used in a multi-SVM format to detect dismounts. Diverse SVMs can work well in concert to allow for "full detection" throughout an image scene.

A benefit of the multi-SVM system is the ability to differentiate based off of pose. This pose information constitutes an additional level of contextual information that aids in human measurements and signatures intelligence. The pose of a specific dismount is of potential use in a surveillance or tracking scenario as an operator of the detection system can communicate pose to aid an agent on the ground in identifying an individual of interest.

The baseline Brooks [5] detector demonstrated a promising ability to perform on par with state of the art dismount detectors, while limiting search space complexity and suppressing two orders of magnitude of false alarms. The improvements from this thesis, extend these performance gains, suppressing over one more order of magnitude of false alarms while allowing for an increased variety of dismount poses. The abilities of the multi-SVM structure are especially evident when applied to challenging crouching poses, yielding a 7-fold increase detection probability. These dramatic improvements clearly demonstrate the benefit of the

multi-SVM approach, which can be extended to include other pose configurations.

Appendix A. Rotated Detections

This appendix discusses an additional way that skin islands can be leveraged to yield better detections. While this content is quite promising, it ultimately was not incorporated into the improved Brooks [5] detector discussed in Section 4.3 due to the added computational complexity entailed.

1.1 Adjustment for rotation

When forming search windows around the centroids of the skin islands, it is possible to include some additional context. Since each skin island is treated as if it belongs to the head, it is possible to fit an ellipse to each of these shapes. A computer program can be used to estimate the locations of the major and minor axis of the ellipse. The orientation of these axes can then be used to extract the degree (θ) to which the ellipse is rotated about the origin (seen in Fig. 55). A slight rotation is applied to the image so that the skin segment pertaining to the head is oriented vertically. Since the head is frequently aligned with the torso (as in the examples of side bends), this has the effect of providing a more vertical orientation for the entire dismount, allowing for better detection. The potential for improved detection, however, comes at the cost of added computational complexity.

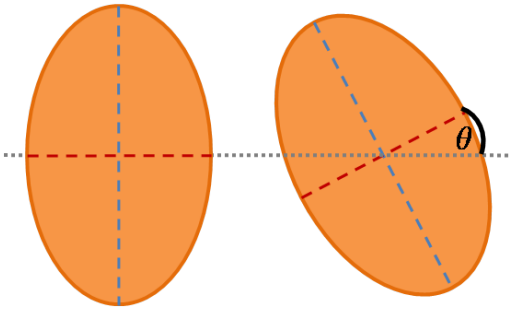


Figure 55. Two ellipses are shown with red minor axes and blue major axes on top of a horizontal gray line. The first ellipse is vertically aligned as its minor axis is parallel to the horizontal. The degree to which the the second ellipse is rotated (θ) can be measured as the angle between its minor axis and the horizontal.

1.2 Rotation Matrices

Rotation Matrices are used to transform sets of points in Euclidean space. For the purposes of this thesis, this is useful to transform coordinates indicating locations of skin island centroids as well as vertexes of image patches and prediction windows. Since these transform uses can be restricted to an xy-plane, an $m \times 2$ set of x, y coordinates, $\mathbf{A}$, can be rotated counterclockwise by the angle θ by computing the product $\mathbf{A} \times \mathbf{R}$, where $\mathbf{R}$ is the rotation matrix

$$\mathbf{R} = \begin{vmatrix} cos\,(\theta) & sin\,(\theta) \\ -sin\,(\theta) & cos\,(\theta) \end{vmatrix}. \tag{21}$$

1.3 Increasing Prediction Strength by Rotation Adjustments

The geometry of the head skin island is used to form a rotated prediction window around a dismount that would normally encounter difficulty in detection. Fig. 56 shows a sequence of dismounts with increasingly extreme side bends. As the prediction windows are cued off of the position of the head, the later side bend examples receive low prediction scores, below or only marginally above the detection threshold. When an ellipse is fitted to the head skin island, the orientation angle can be extracted to rotate the prediction window. The resulting predictions in Fig. 57 show significant increases in prediction score, providing a useful option to extend the coverage of the dismount detector leveraging the properties of the head skin islands. Automatic adjustments for rotation are of particular use for imaging systems that are anticipated to experience rotation around the frontal plane.

Table 3. The orientation angle of head skin islands is used to apply a slight rotation to a dismount. As the head and torso are better aligned dramatic improvements in prediction strength are witnessed.

	frame 1	frame 2	frame 3	frame 4	frame 5	frame 6
Orig. Pred	2.108	1.781	1.395	0.806	0.446	0.269
New Pred	2.201	2.120	2.081	2.057	1.841	1.879

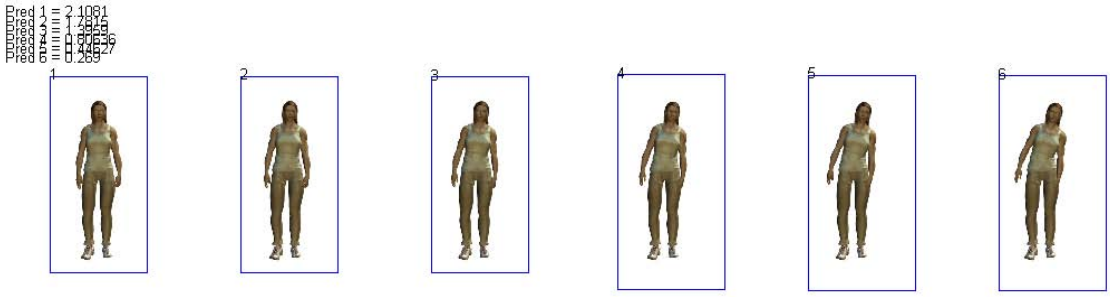


Figure 56. Dismounts are detected in the six frames with lower prediction strengths as the severity of a side bend is increased.

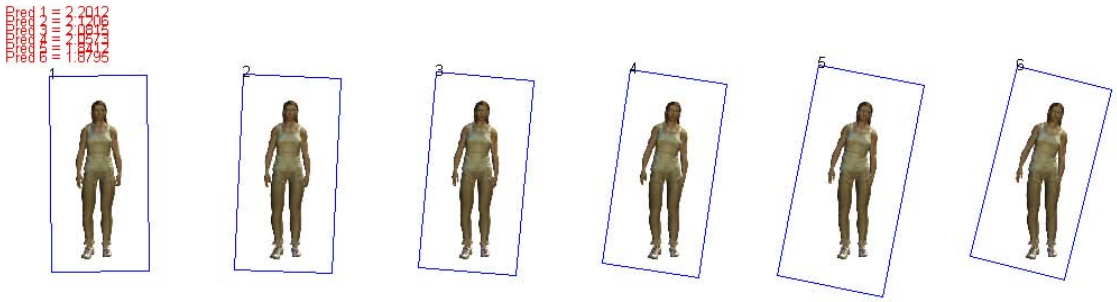


Figure 57

Appendix B. Clustered Silhouette Chips

This appendix documents the ISOMAP clustering of simulated dismount poses and supports the work in Section 3.5.2.1. A complete set of the silhouette chips used for clustering are seen in Fig. 58. While the poses shown are in no way exhaustive, they serve to represent distinctions between poses and orientation angles.

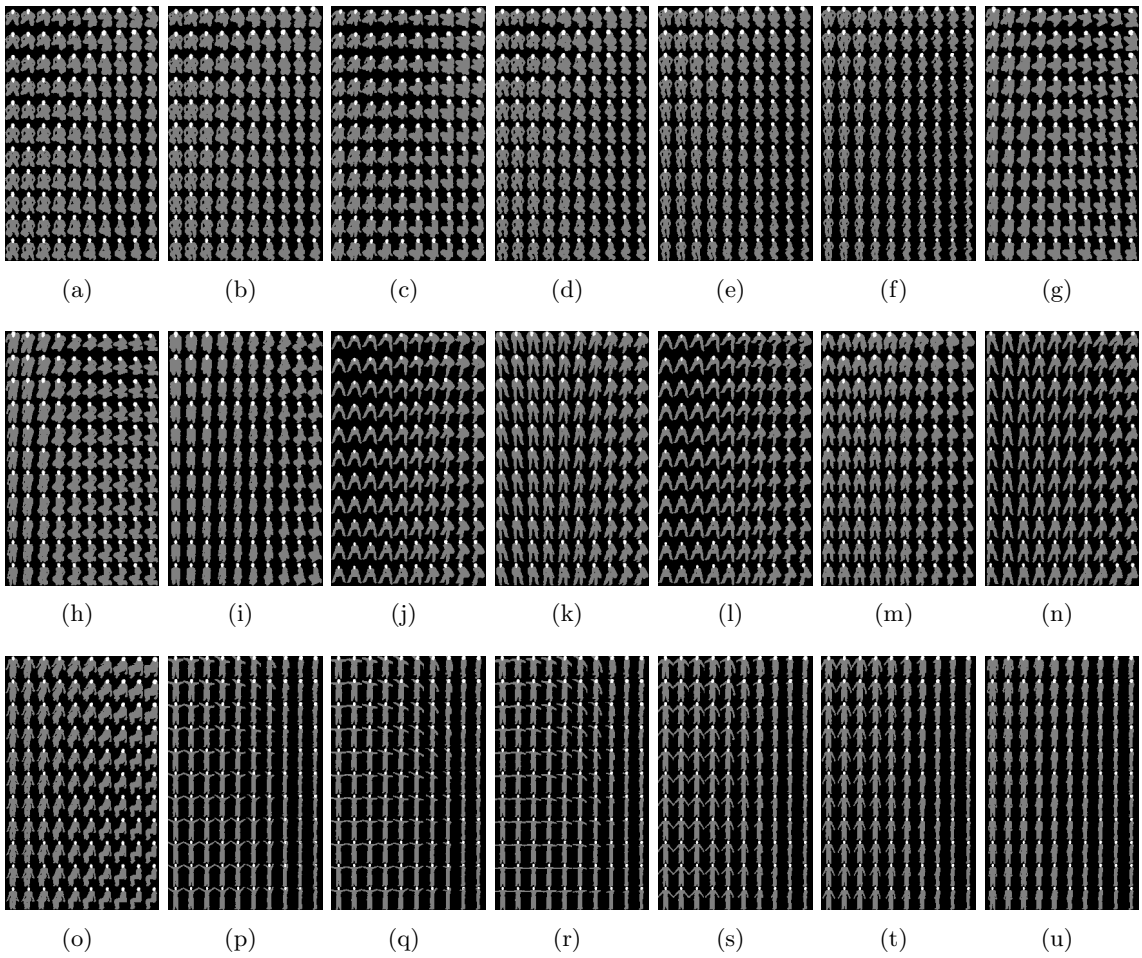


Figure 58. All 21 different poses used in ISOMAP clustering are displayed as pose silhouettes with highlighted head skin regions. Rows of each subplot correspond to evenly spaced angles of elevation in the range $[0°, 50°]$ with columns corresponding to azimuth angle in the range $[0°, 90°]$. The poses can be identified by label: (a) CR1 (b) CR2 (c) CR3 (d) CR4 (e) CR5 (f) CR6 (g) KN1 (h) KN2 (i) KN3 (j) SIT1 (k) SIT2 (l) SIT3 (m) SIT4 (n) SIT5 (o) SIT6 (p) A-30 (q) A-15 (r) A0 (s) A45 (t) A60 (u) A75

K means is used to group the labeled training samples. A choice of $K = 10$ is eventually decided upon (seen in Fig. 59) after comparing with a wide number of values for K. Two close contenders are K values of 9 (Fig. 60) and 11 (Fig. 61).

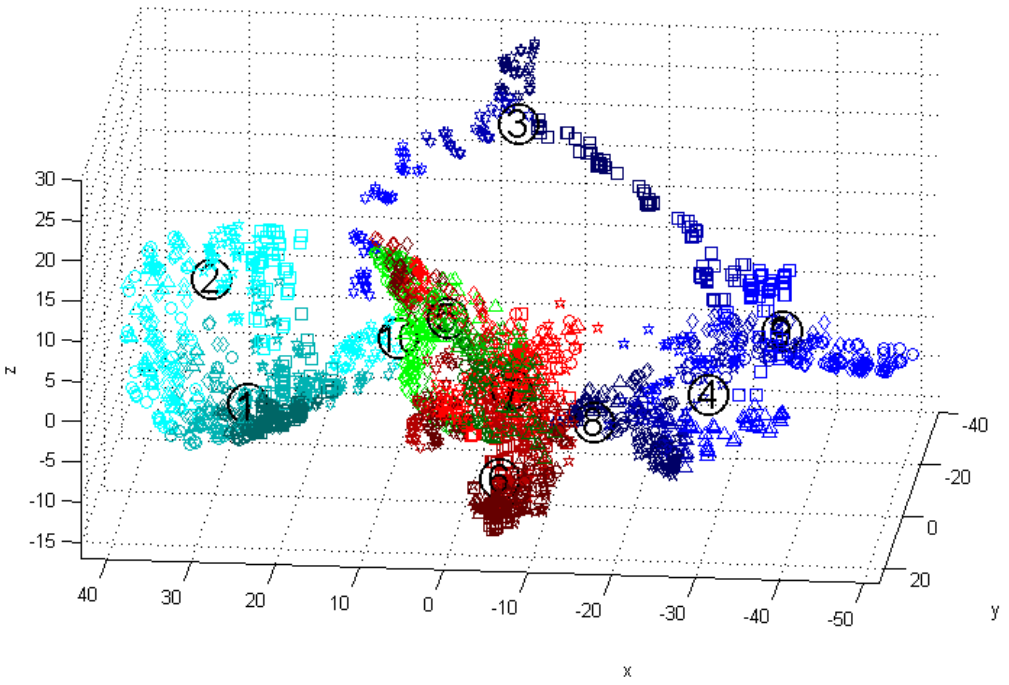


Figure 59. Labeled data is seen with 3 of the 5 Isomap clustering dimensions as well as 10 machine learned representative means. The legend from 36(b) applies to this figure identifying data points by pose and azimuth angle range.

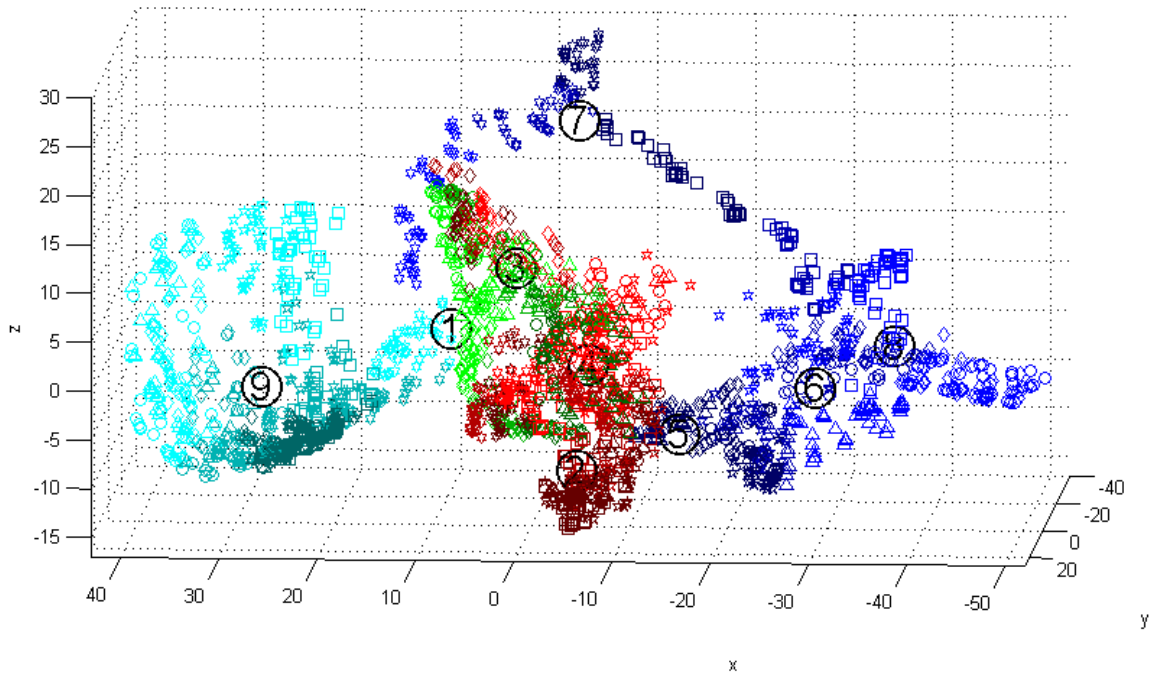


Figure 60. Labeled data is seen with 3 of the 5 Isomap clustering dimensions as well as 9 machine learned representative means. The legend from 36(b) applies to this figure identifying data points by pose and azimuth angle range.

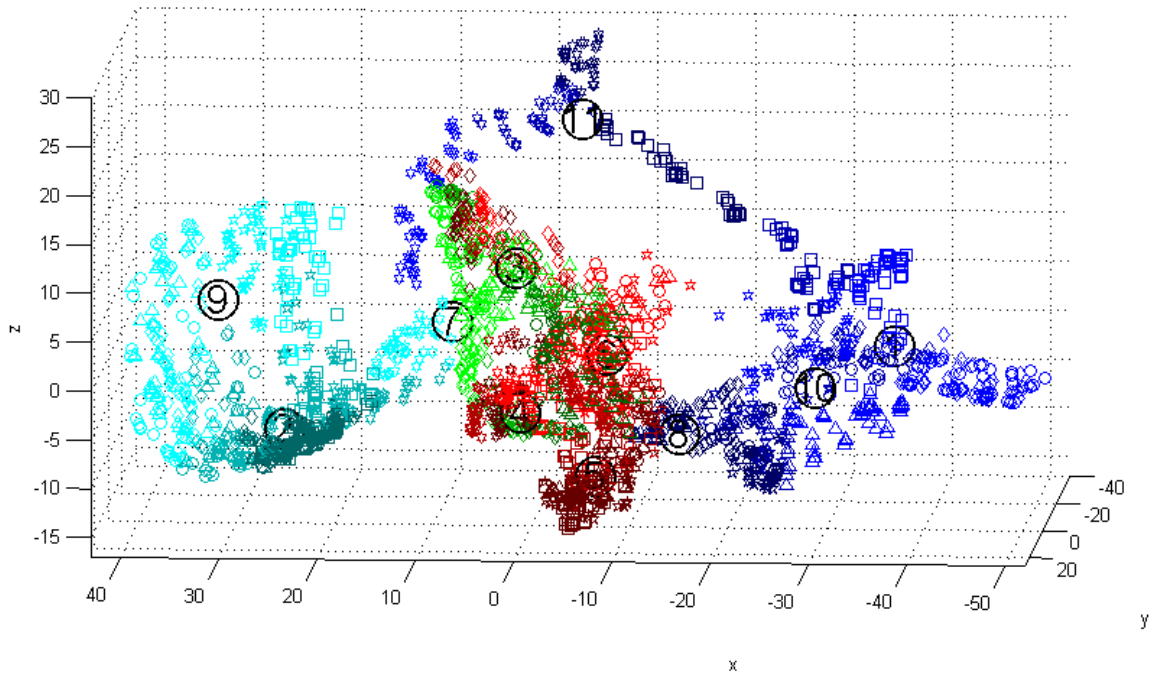


Figure 61. Labeled data is seen with 3 of the 5 Isomap clustering dimensions as well as 11 machine learned representative means. The legend from 36(b) applies to this figure identifying data points by pose and azimuth angle range.

Complete groupings of image patches belonging to each of the ten clusters are displayed in Fig. 62 through 71. Graphical representations of these shilouette chips are arranged according to distance (in ISOMAP space) from the shilouette to the mean vector representing the cluster. This arrangement is organized starting in the top left corner, then moving across rows.

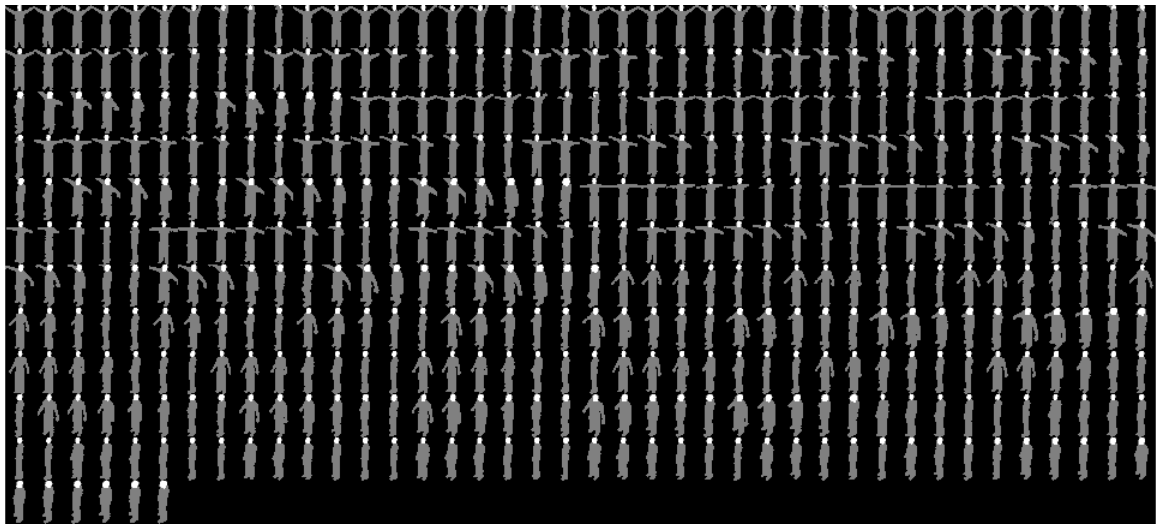


Figure 62. Cluster 1

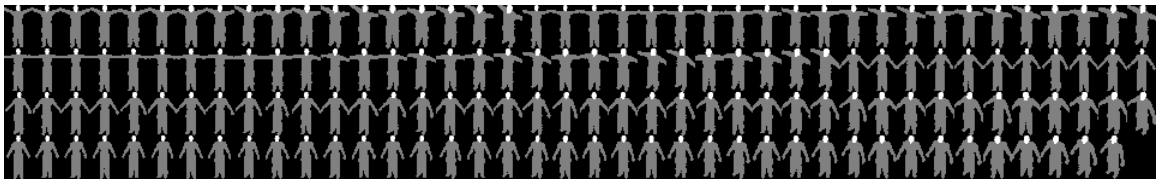


Figure 63. Cluster 2

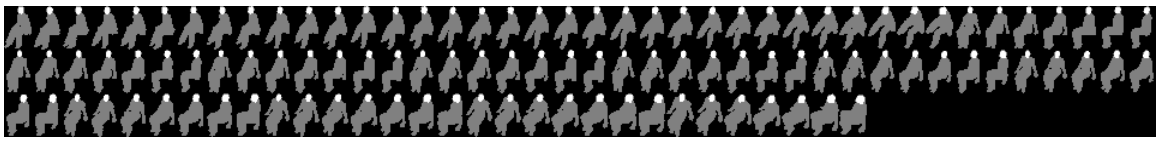


Figure 64. Cluster 3

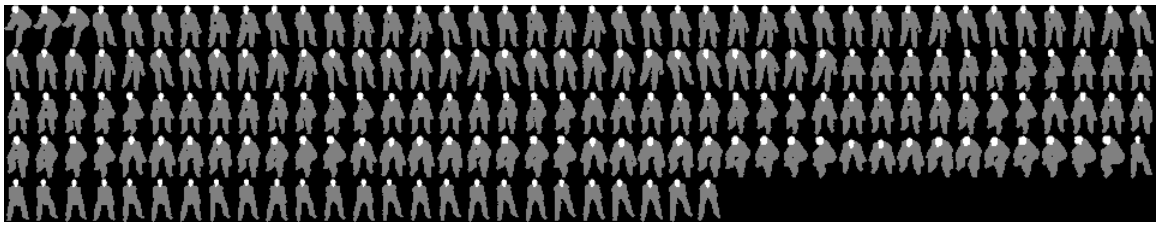


Figure 65. Cluster 4

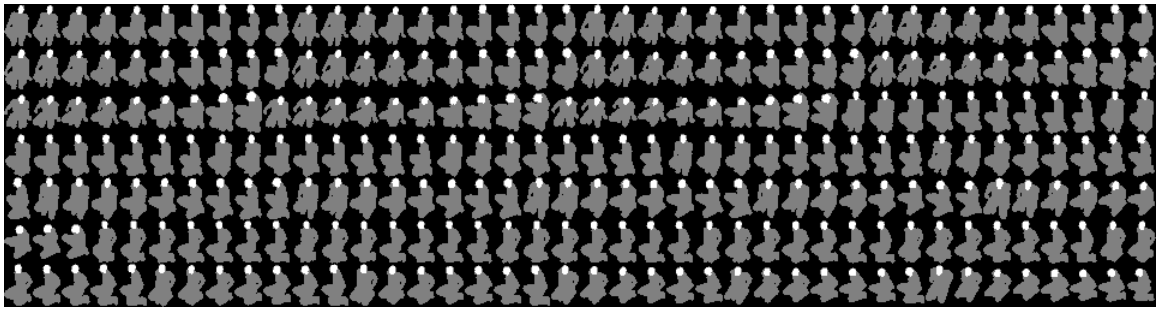


Figure 66. Cluster 5

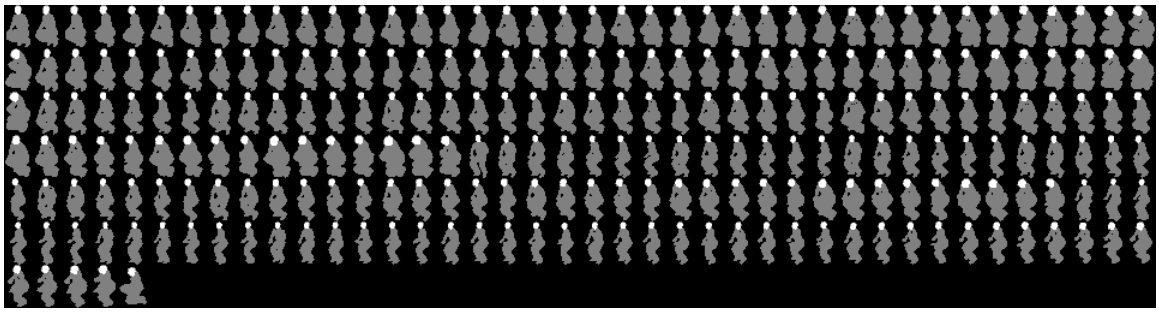


Figure 67. Cluster 6

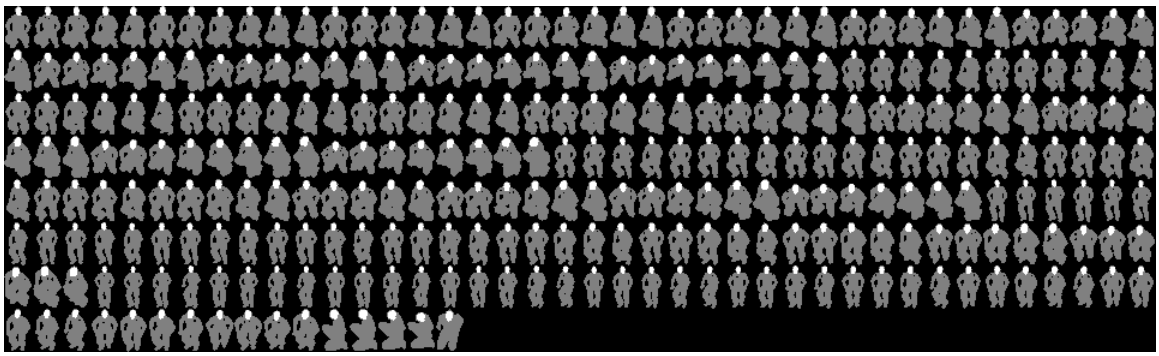


Figure 68. Cluster 7

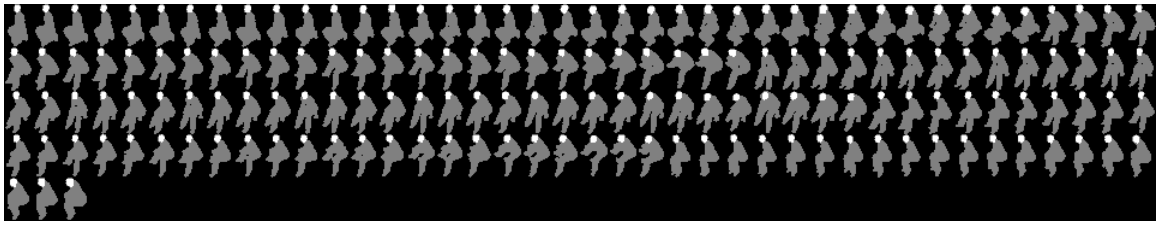


Figure 69. Cluster 8

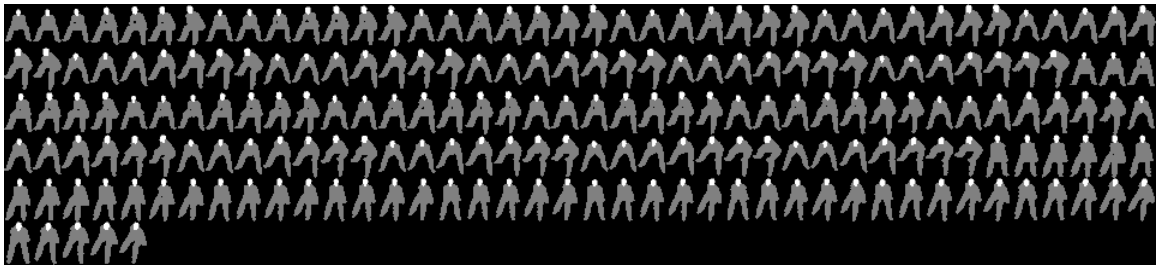


Figure 70. Cluster 9

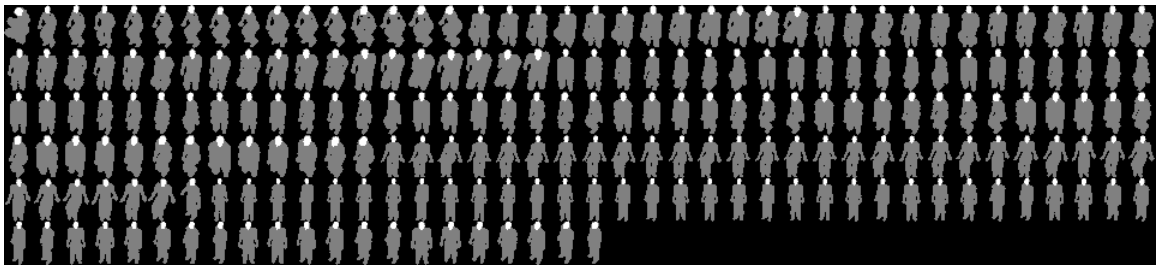


Figure 71. Cluster 10

Appendix C. Pose Similarity Over Camera Angle

This appendix analyzes the similarity of silhouettes within specific pose types and supports the work in Sections 3.5.2.1 and 4.4.1. Generalities about the similarity of a pose exposed to different camera angles can be explored as the Euclidean (ISOMAP) distance from each silhouette chip to a mean vector is plotted in a hemispherical representation. As the colors decay from yellow to pink, natural break points are found that help determine broad ranges of poses and camera angles that yield similar images. These generalities can be applied to identify training samples that can together be used together to form supplemental SVMs.

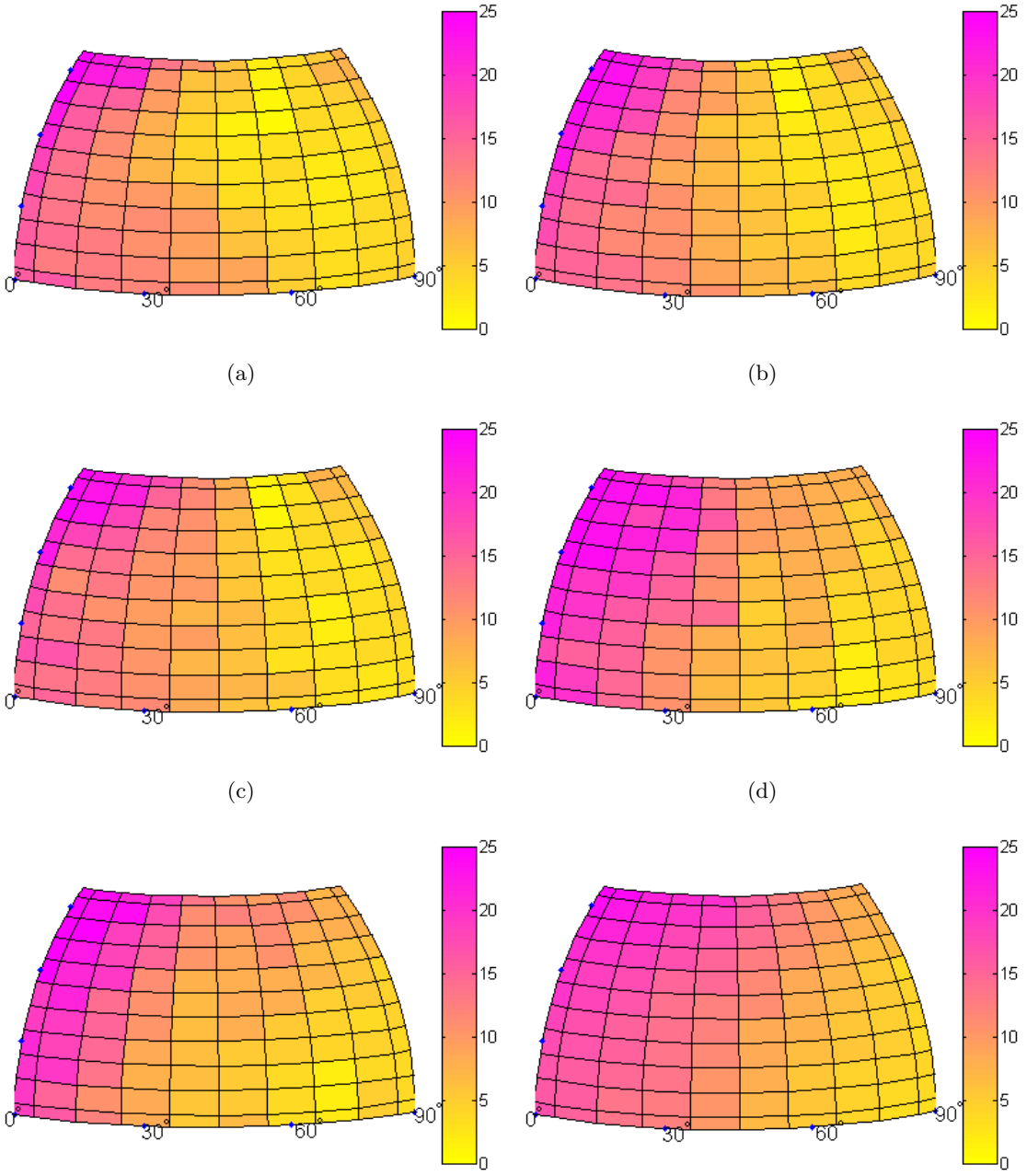


Figure 72. Hemispherical plot displaying the Euclidean distance of each of the 210 clustered poses from each training set in five dimensional space from "mean vector 4". (The color scale is fixed from 0 to 25 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience) (a) A-30 (b) A-15 (c) A0 (d) A45 (e) A60 (f) A75

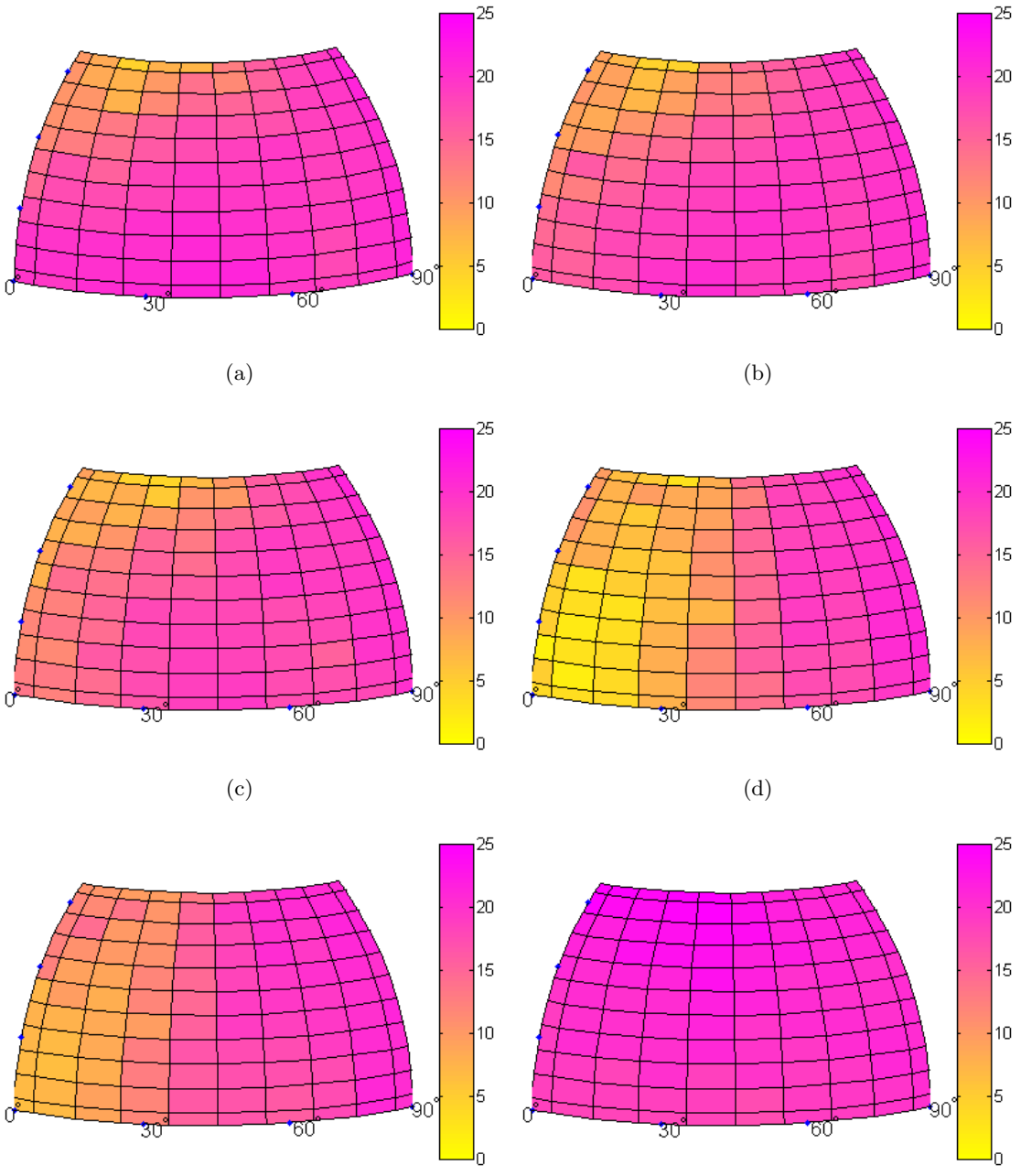
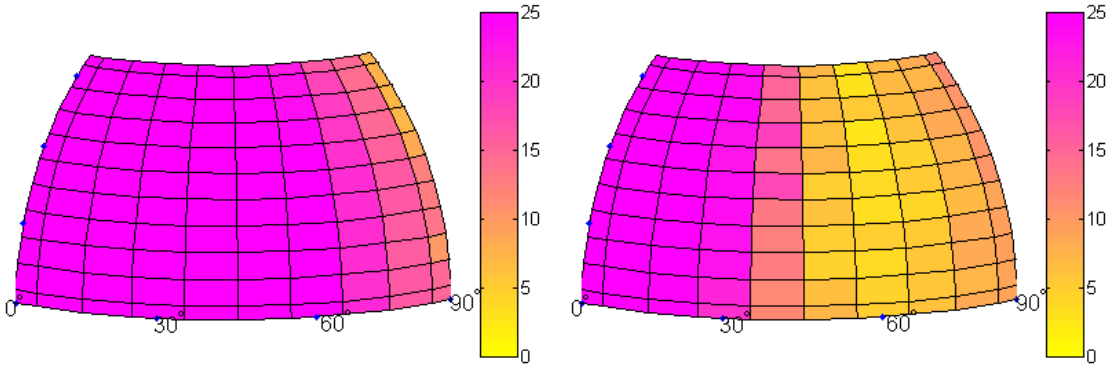


Figure 73. Hemispherical plot displaying the Euclidean distance of each of the 210 clustered poses from each training set in five dimensional space from "mean vector 2". (The color scale is fixed from 0 to 25 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience) (a) A-30 (b) A-15 (c) A0 (d) A45 (e) A60 (f) A75



(a) (b)

Figure 74. Hemispherical plot displaying the Euclidean distance of each of the 210 clustered poses from each training set in five dimensional space from "mean vector 3". (The color scale is fixed from 0 to 25 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience) (a) SIT5 (b) SIT6

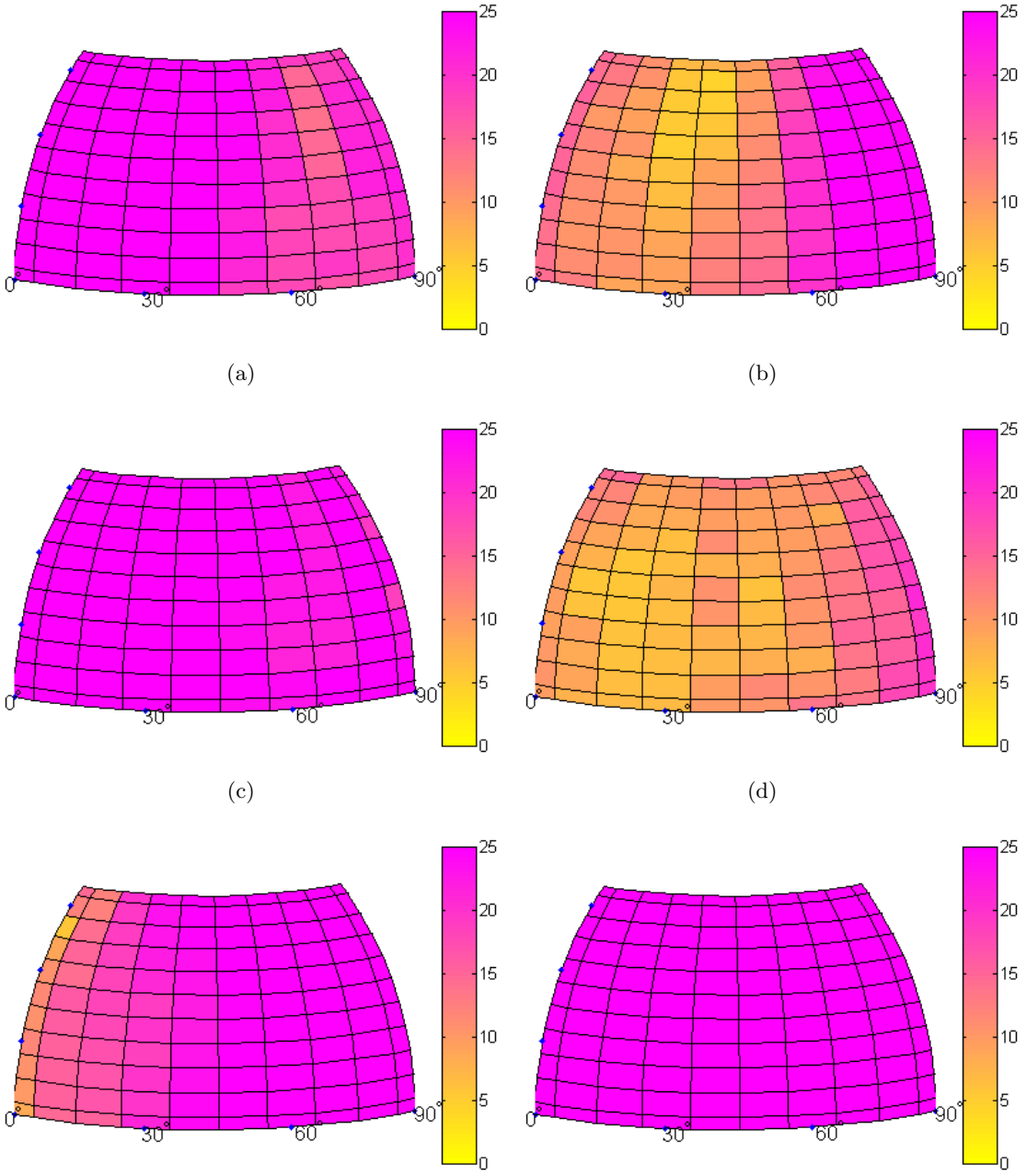


Figure 75. Hemispherical plot displaying the Euclidean distance of each of the 210 clustered poses from each training set in five dimensional space from "mean vector 4". (The color scale is fixed from 0 to 25 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience) (a) SIT1 (b) SIT2 (c) SIT3 (d) SIT4 (e) SIT5 (f) SIT6

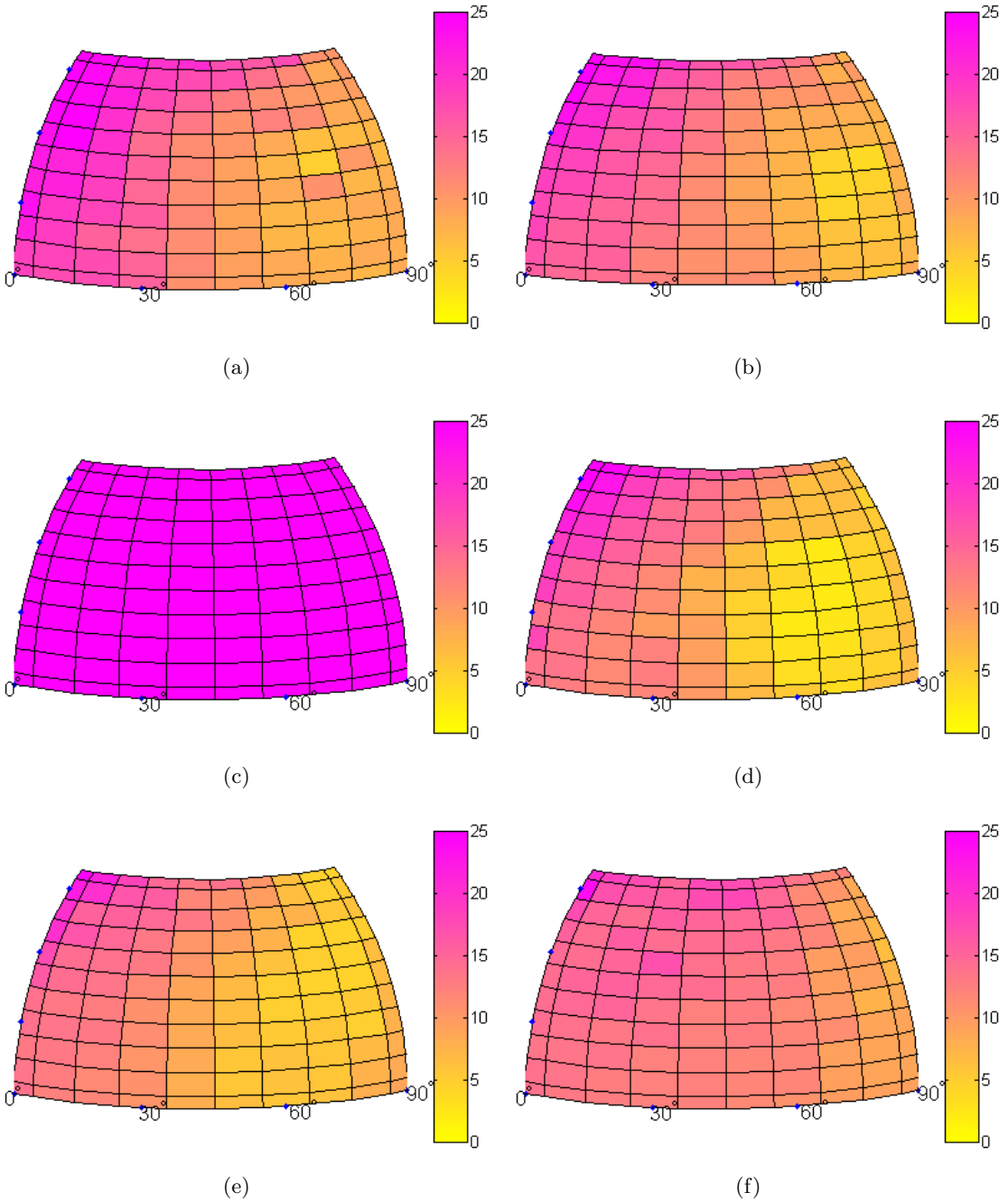


Figure 76. The similarity of representative crouching poses to "mean vector 6" are shown in a Hemispherical plot displaying the Euclidean distance from "mean vector 6" of each of the 210 clustered poses from each training set in five dimensional IsoMap space. (The color scale is fixed from 0 to 25 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience) (a) CR1 (b) CR2 (c) CR3 (d) CR4 (e) CR5 (f) CR6

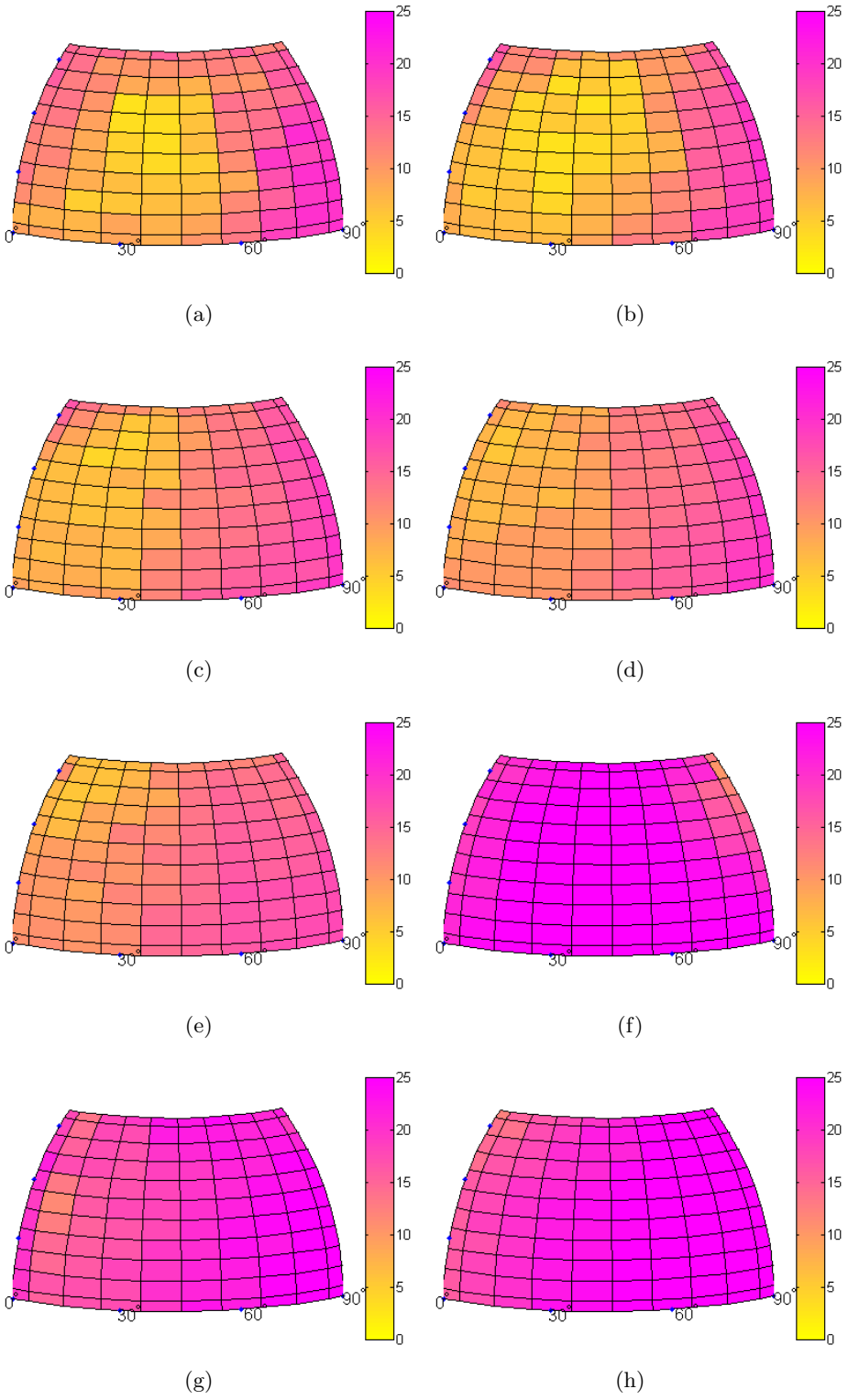


Figure 77. Hemispherical plot displaying the Euclidean distance of each of the 210 clustered poses from each training set in five dimensional space from "mean vector 7". Noteably absent is set CR3, which returned a Euclidean distance of 25 or greater for the 110 poses tested. (The color scale is fixed from 0 to 25 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience) (a) CR1 (b) CR2 (c) CR4 (d) CR5 (e) CR6 (f) KN1 (g) KN2 (h) KN3

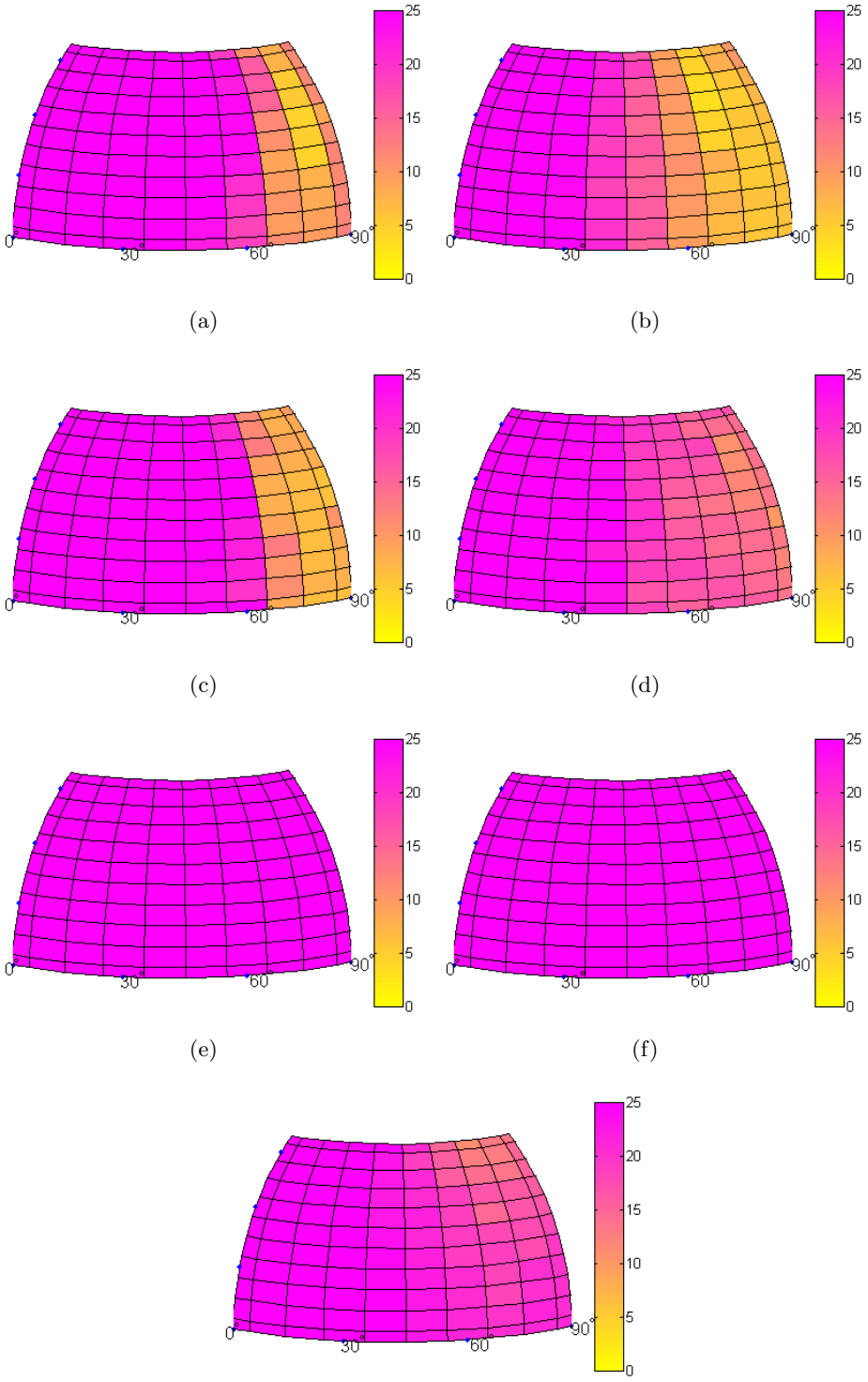


Figure 78. Hemispherical plot displaying the Euclidean distance of each of the 210 clustered poses from each training set in five dimensional space from "mean vector 8". (The color scale is fixed from 0 to 25 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience) (a) SIT1 (b) SIT2 (c) SIT3 (d) SIT4 (e) SIT5 (f) SIT6 (g) KN3

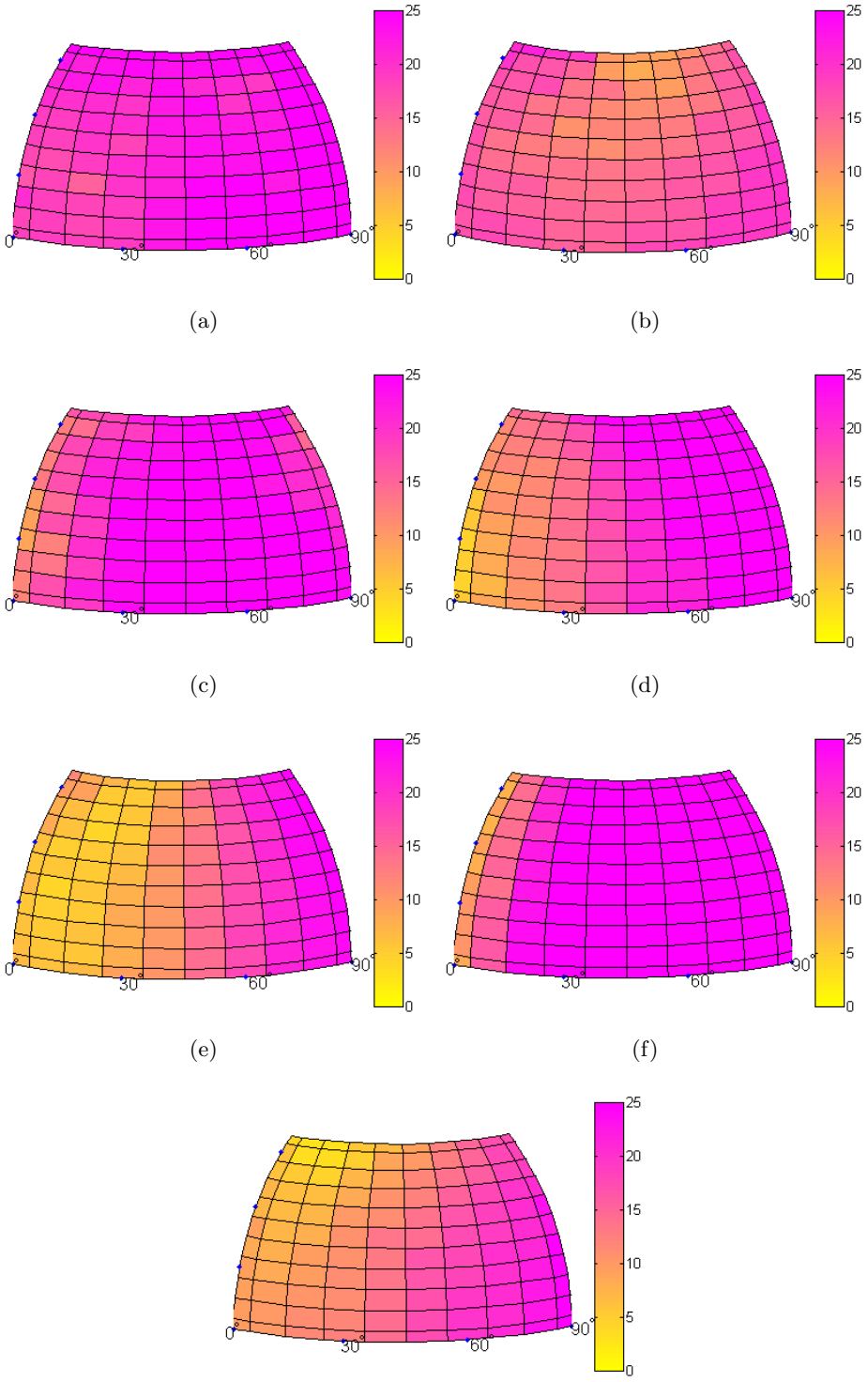


Figure 79. Hemispherical plot displaying the Euclidean distance of each of the 210 clustered poses from each training set in five dimensional space from "mean vector 10". (The color scale is fixed from 0 to 25 and blue dots along the angle of elevation at 15°, 30°, and 45° are shown for convenience) (a) CR5 (b) CR6 (c) KN1 (d) KN2 (e) KN3 (f) SIT6 (g)A75

Bibliography

[1] Agarwal, A. and B. Triggs. "Recovering 3D human pose from monocular images". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(1):44 –58, 2006. ISSN 0162-8828.

[2] Andriluka, M., S. Roth, and B. Schiele. "Pictorial structures revisited: People detection and articulated pose estimation". *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1014 –1021. 2009. ISSN 1063-6919.

[3] Bauer, K. B. "OENG 685 Applied Multivaritae Analysis 1", 2006. Air Force Institute of Technology course notes.

[4] Belongie, S., J. Malik, and J. Puzicha. "Shape matching and object recognition using shape contexts". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):509 –522, April 2002. ISSN 0162-8828.

[5] Brooks, Adam. *Improved multispectral skin detection and its application to search space reduction for dismount detection based on histograms of oriented gradients*. Master's thesis, Air Force Institute of Technology, Mar 2010.

[6] Cao, Hui, Noboru Ohnishi, Yoshinori Takeuchi, Tetsuya Matsumoto, and Hiroaki Kudo. "FAST HUMAN POSE RETRIEVAL USING APPROXIMATE CHAMFER DISTANCE". 2006. URL `http://hdl.handle.net/2237/10437`.

[7] Chen, D., X.B. Cao, Y.W. Xu, H. Qiao, and F.Y. Wang. "A SVM-based classifier with shape and motion features for a pedestrian detection system". *Intelligent Vehicles Symposium, 2006 IEEE*. 0 2006.

[8] Clark, Jeffrey. "Stochastic Feature Selection with Distributed Spacing and its Application to Textile". PhD perspectus presented on 25 August 2010.

[9] Dalal, Navneet and Bill Triggs. "Histograms of Oriented Gradients for Human Detection". *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:886–893, 2005.

[10] Dalal, Navneet, Bill Triggs, and Cordelia Schmid. "Human Detection Using Oriented Histograms of Flow and Appearance". *Computer Vision ECCV 2006*, volume 3952 of *Lecture Notes in Computer Science*, 428–441. Springer Berlin / Heidelberg, 2006. URL `http://dx.doi.org/10.1007/11744047_33`.

[11] Dowdall, Jonathan A., Ioannis A Pavlidis, and George B Bebis. "Face Detection in the Near-IR Spectrum". *Proceedings of Infrared Technology and Applications XXIX*. 2003.

[12] Eismann, M. T. "OENG 633 Hyperspectral Remote Sensing", 2006. Air Force Institute of Technology course notes.

[13] Enzweiler, M. and D.M. Gavrila. "Monocular Pedestrian Detection: Survey and Experiments". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(12):2179 –2195, 2009. ISSN 0162-8828.

[14] Fang, Yajun, K. Yamada, Y. Ninomiya, B.K.P. Horn, and I. Masaki. "A shape-independent method for pedestrian detection with far-infrared images". *Vehicular Technology, IEEE Transactions on*, 53(6):1679 – 1697, 2004. ISSN 0018-9545.

[15] Forsyth, D.A. and M.M. Fleck. "Automatic Detection of Human Nudes". *International Journal of Computer Vision*, 32:63–77, 1999. ISSN 0920-5691. URL `http://dx.doi.org/10.1023/A:1008145029462`. 10.1023/A:1008145029462.

[16] Gavrila, D. M. and S. Munder. "Multi-cue pedestrian detection and tracking from a moving vehicle". *International Journal of Computer Vision (est: February-March*, 73:41–59, 2007.

[17] Hastie, T., R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, corrected edition, July 2003. ISBN 0387952845.

[18] Inc., DAZ Productions. "DAZ Studio^{TM}3D". URL `http://www.daz3d.com`.

[19] Jengo, C.M. and J. LaVeigne. "Sensor performance comparison of HyperSpecTIR instruments 1 and 2". *Aerospace Conference, 2004. Proceedings. 2004 IEEE*, volume 3, 6 vol. (xvi+4192). 2004. ISSN 1095-323X.

[20] Joachims, Thorsten. *Making large-scale support vector machine learning practical*, 169–184. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-19416-3. URL `http://portal.acm.org/citation.cfm?id=299094.299104`.

[21] Kilgore, George A. and Rand P. Whillock. "Skin detection sensor", May 2007. URL `http://www.freepatentsonline.com/y2007/0106160.html`.

[22] Koch, Brad. *A multispectral bidirectional reflectance distribution function study of human skin for improved dismount detection*. Master's thesis, Air Force Institute of Technology, Mar 2011.

[23] Laboratory, Air Force Research. *Modtran4 version 3, revision 1*. Technical report, Space Vehicles Directorate, Air Force Materiel Command, Hanscom AFB, Mass. 01731-3010, 2001.

[24] Leibe, B., E. Seemann, and B. Schiele. "Pedestrian detection in crowded scenes". 1:878 – 885 vol. 1, 2005. ISSN 1063-6919.

[25] Lin, Zhe and L.S. Davis. "Shape-Based Human Detection and Segmentation via Hierarchical Part-Template Matching". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(4):604 –618, 2010. ISSN 0162-8828.

[26] Lowe, David G. "Distinctive Image Features from Scale-Invariant Keypoints". *Int. J. Comput. Vision*, 60:91–110, November 2004. ISSN 0920-5691. URL `http://portal.acm.org/citation.cfm?id=993451.996342`.

[27] Mohan, A., C. Papageorgiou, and T. Poggio. "Example-based object detection in images by components". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(4):349 –361, April 2001. ISSN 0162-8828.

[28] News, Defense. "New USAF Leaders Lay Out Top Priorities", Aug 2008.

[29] Nunez, Abel S. *A physical model of human skin and its application for search and rescue*. Ph.D. thesis, Air Force Institute of Technology, 2009.

[30] Nunez, A.S. and M.J. Mendenhall. "Detection of Human Skin in Near Infrared Hyperspectral Imagery". *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International*, volume 2, II–621 –II–624. 2008.

[31] Peskosky, Keith. *Design of a monocular multispectral skin detection, melanin estimation, and false alarm suppression system*. Master's thesis, Air Force Institute of Technology, Mar 2010.

[32] Ramanan, D., D.A. Forsyth, and A. Zisserman. "Strike a pose: tracking people by finding stylized poses". *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, 271 – 278 vol. 1. 2005. ISSN 1063-6919.

[33] Security, Global. "Northrop Grumman Successfully Demonstrates VADER Dismount Detection", Feb 2010.

[34] Shashua, A., Y. Gdalyahu, and G. Hayun. "Pedestrian detection for driving assistance systems: single-frame classification and system level performance". *Intelligent Vehicles Symposium, 2004 IEEE*, 1 – 6. 2004.

[35] Tenenbaum, Joshua B., Vin de Silva, and John C. Langford. "A Global Geometric Framework for Nonlinear Dimensionality Reduction". *Science*, 290(5500):2319–2323, 2000. URL `http://www.sciencemag.org/content/290/5500/2319.abstract`.

[36] Viola, P., M.J. Jones, and D. Snow. "Detecting pedestrians using patterns of motion and appearance". *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 734 –741 vol.2. 2003.

[37] Wu, Bo and R. Nevatia. "Tracking of Multiple, Partially Occluded Humans based on Static Body Part Detection". *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, 951 – 958. 2006. ISSN 1063-6919.

[38] Zhang, Li, Bo Wu, and R. Nevatia. "Detection and Tracking of Multiple Humans with Extensive Pose Articulation". *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 1 –8. 2007. ISSN 1550-5499.

[39] Zhu, Qiang, Mei-Chen Yeh, Kwang-Ting Cheng, and S. Avidan. "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients". *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. 2006.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
24–03–2011	Master's Thesis	Aug 2009 — Mar 2011

4. TITLE AND SUBTITLE

OVERCOMING POSE LIMITATIONS OF A SKIN-CUED HISTOGRAMS OF ORIENTED GRADIENTS DISMOUNT DETECTOR THROUGH CONTEXTUAL USE OF SKIN ISLANDS AND MULTIPLE SUPPORT VECTOR MACHINES

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Jonathon R. Climer, 2d Lt, USAF

5d. PROJECT NUMBER

10ENG300

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GE/ENG/11-05

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/RHPA (Julia Parakkat)
2800 Q Street, B824
WPAFB, USA 45433
937-255-0605, Julia.Parakkat@wpafb.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

AFRL/RHPA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This thesis provides a novel visualization method to analyze the impact that articulations in dismount pose and camera aspect angle have on histograms of oriented gradients (HOG) features and eventual detections. Insights from these relationships are used to identify limitations in a state of the art skin cued HOG dismount detector's ability to detect poses not in a standard upright stances. Improvements to detector performance are made by further leveraging available skin information, reducing false detections by an additional order of magnitude. In addition, a method is outlined for training supplemental support vector machines (SVMs) from computer generated data, for detecting a wider range of poses and camera configurations. The multi-SVM structure yields a 7-fold increase detection probability when applied to challenging crouching poses. These dramatic improvements clearly demonstrate the viability of such an approach, which can be extended to include other pose configurations.

15. SUBJECT TERMS

HOG, Dismount, Pose Articulations, Skin-Cued, SVM

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Michael Mendenhall, Maj, USAF (ENG)
U	U	U	UU	130	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636, x4614; Michael.Mendenhall@afit.edu