ABSTRACT

CLOUSE, HAMILTON SCOTT. Varying Scales of Human Behavior Analysis: A Geometric Approach to Action Study in Multiple Modalities. (Under the direction of Dr. Hamid Krim.)

Remote sensing is the field concerned with studying objects far removed from the observer. All endeavors in this field can be viewed as a process: the existence of a physical phenomenon, it’s observation via sensor(s), information extraction from the sensor data and, finally, the exploitation of that information for some application. Within this field is a set of problems, “layered sensing,” that focuses on the combination of information from several such processes for a common application. Due to differences in sensor data at varying scales of removal from the observed phenomenon, this task in non-trivial.

In this work, we present a complete system that combines information from three such remote sensing processes at different scales and with data from different sensor types into single, coherent representation. This system consists of algorithms that outperform the current state of the art for each process and produce representations of the results that are readlity consumable by each.
Varying Scales of Human Behavior Analysis: A Geometric Approach to Action Study in Multiple Modalities

by
Hamilton Scott Clouse

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Electrical Engineering

Raleigh, North Carolina

2014

APPROVED BY:

Griff Bilbro
Edgar Lobaton

Raju Vatsavai
Hamid Krim
Chair of Advisory Committee
DEDICATION

To my mom.
Hamilton Scott Clouse received his Bachelor of Science degrees in Electrical Engineering and Applied Mathematics from the University of Tennessee at Chattanooga in 2006 and his Master of Science degree in Electrical Engineering from North Carolina State University (NCSU) in 2010. He is currently pursuing a PhD in Electrical Engineering at NCSU while working in the Vision, Information and Statistical Signal Theories and Applications (VISSTA) group. His research interests include signal processing and pattern recognition with applications in machine learning and computer vision.
ACKNOWLEDGEMENTS

Without the help and support (in manifold ways), of my family and friends, among whom I count my nuclear family, my adviser, thesis committee, lab-mates in VISSTA, colleagues at AFRL, and friends both within and without NCSU, I would not have been able to complete this endeavor. Thank you all.
TABLE OF CONTENTS

LIST OF TABLES ................................................................. viii

LIST OF FIGURES ................................................................. ix

Chapter 1: Introduction ......................................................... 1

Chapter 2: Prior Related Research ............................................. 5
  2.1 Image Mosaicking .............................................................. 5
  2.1.1 Feature-driven Analysis .................................................. 6
  2.1.2 Random Sample and Consensus (RANSAC) ............................... 9
  2.1.3 Current Methods for Mosaic Comparison ................................ 10
  2.2 Scene Understanding .......................................................... 11
  2.3 Activity Analysis ............................................................... 11
  2.4 Statistical Tools ............................................................... 12
     2.4.1 Statistical Moments ................................................... 12
     2.4.2 Wilcoxon Signed-Rank Test .......................................... 14
  2.5 Robust Subspace Recovery via Bi-Sparsity Pursuit (RoSuRe-BSP) ...... 16

Chapter 3: Gigapixel Image Mosaicking via Parallel Sparse Tile Decomposition ................................................................. 19
  3.1 Background ....................................................................... 21
  3.1.1 Data ........................................................................... 21
  3.1.2 Problem Formulation ..................................................... 21
  3.2 Image Decomposition .......................................................... 23
3.3 Sparse Subspace Recovery .................................................. 26
3.4 The Structure of the Coefficients ......................................... 27
   3.4.1 Synthetic Data .......................................................... 27
   3.4.2 Tile Translation ......................................................... 27
   3.4.3 Tile Rotation ............................................................ 31
3.5 Results on Captured Data from AWARE-2 .............................. 33
3.6 Evaluation ................................................................. 34
3.7 Summary ................................................................. 38

Chapter 4: Multi-Level Scene Understanding via Hierarchical Classification ........................................... 39
4.1 Methods ........................................................................ 40
   4.1.1 The partitioning of the FOV ......................................... 41
   4.1.2 Description of the ROIs ................................................ 42
   4.1.3 Region Activity Description .......................................... 43
   4.1.4 Training and Classification Framework ........................... 44
4.2 Experimental Testing ....................................................... 44
   4.2.1 Background/Foreground Separation ............................... 45
   4.2.2 The partitioning of the FOV ......................................... 46
   4.2.3 Description of the ROIs ................................................ 46
   4.2.4 Region Activity Description .......................................... 47
4.3 Summary ................................................................. 49

Chapter 5: Activity Analysis and Material Discrimination via LWIR Polarization ............................................. 51
5.1 Introduction .................................................................... 51
5.2 Related Work .............................................................. 52
5.3 Background .................................................................... 56
   5.3.1 Polarization of EM Radiation ....................................... 56
   5.3.2 Physics-based Model of Emission Polarization ............... 60
   5.3.3 Statistical Analysis Tools ............................................. 62
5.4 Data Description .......................................................... 63
LIST OF TABLES

Table 3.1  Table of SNR values for band of strong coefficients in $W$ corresponding to the listed rotations in $I_2$. SNR was computed in the standard way for the test images. [36] .................................................. 33

Table 5.1  Ranges of the defined sub-bands within the IR spectrum band. ........ 65

Table 5.2  Confusion matrix representing the accuracy of the classification performed (in %) via the spectra of the kurtosis time-series characteristics. 80
LIST OF FIGURES

Figure 1.1 Information flow in a Remote Sensing process. The four stages/requirements are: 1) the existence of an object or phenomenon, 2) the collection of data about the object/phenomenon, 3) information extraction through analysis of the data and 3) application of the extracted information. 3

Figure 2.1 Illustration of the sign of the skewness and its relation to the asymmetry of the distribution. 2.1(a) - negative skewness, 2.1(b) - normal distribution, 2.1(c) - positive skewness. 14

Figure 2.2 Illustration of the kurtosis and its relation to the peakedness of the distribution. 2.2(a) - low kurtosis, 2.2(b) - normal distribution, 2.2(c) - high kurtosis. 14

Figure 3.1 Mosaic output from seven images using current state of the art tools. 20
Figure 3.2 Two example high-resolution images from the micro-camera array. 22
Figure 3.3 Depictions of the AWARE-2 camera array. 23
Figure 3.4 The graph of neighboring images is formed by considering the overlap/intersection between images as the prerequisite for edges between representative nodes. 24
Figure 3.5 Partitioning of tile into minimal regions. 25
Figure 3.6 Test image from which synthetic sub-images, or tiles, were generated. [23] 27
Figure 3.7 Masks indicating the portions of the test image, Figure 3.6, used to create synthetic tiles. 29
Figure 3.8  Coefficient matrices $W_c$ and $W_r$ are computed in Equations (3.12) & (3.13). Lighter colors indicate high values. Blocks in the matrix, delimited by artificially inserted horizontal and vertical lines, are the representation of the sub-image corresponding to the column in the space of the sub-image corresponding to the row.

Figure 3.9  Image decomposition used for rotation tests.

Figure 3.10  Subsets of the coefficient matrix $W$ corresponding to various rotations of $I_2$.

Figure 3.11  Computed mosaic of football scene.

Figure 3.12  Computed mosaic of graduation scene #1. The lack of a micro-camera in the array caused a blank space on the left side of the produced mosaic.

Figure 3.13  Computed mosaic of graduation scene #2. The lack of a micro-camera in the array caused a blank space in the middle of the produced mosaic.

Figure 3.14  Computed mosaic of park scene.

Figure 3.15  Comparison of other work to our computed mosaic via a sampled portion.

Figure 4.1  Example frame of $X(x, y, t)$ showing courtyard from MSEE Data.

Figure 4.2  Close-up of example frame showing results of (left) background, $B(x, y, t)$, and (right) foreground, $F(x, y, t)$, separation via RoSuRe-DSP.

Figure 4.3  Region boundaries produced from local maxima of $D$.

Figure 4.4  Similarity between regions indicated by the number of matches, normalized.

Figure 4.5  Activity density time-series for regions $V_2$ and $V_{10}$, as labeled in Figure 4.3.

Figure 4.6  Similarity between activity density signatures, normalized.

Figure 5.1  The E-field component of an EM wave rotates about the transverse axis and thus traces an ellipse in the plane of rotation. This wave, $E$, can be decomposed into orthogonal components $E_x$ and $E_y$. 
The polarization ellipse with the labelled parameters (From [22]). $E_{0x}$ and $E_{0y}$ are the amplitudes of the corresponding orthogonal constituent waves, $\xi$ and $\eta$ are the major and minor axes of the ellipse, $\psi$ is the angle of polarization and $\chi$ is the angle of ellipticity.

The degenerate states of the polarization ellipse with conditions (From [22]). $E_{0x}$ and $E_{0y}$ are the amplitudes of the corresponding orthogonal constituent waves, and $\delta$ is the phase difference between those waves. From top-left, proceeding clockwise: linear horizontal polarized (LHP), linear vertical polarized (LVP), linear $-45^\circ$ polarized (L-45P), left circular polarized (LCP), right circular polarized (RCP), linear $+45^\circ$ polarized (L+45P).

Superpixel polarizing filter pattern of micro-polarimeter focal-plane array. Starting at the top left and proceeding clockwise, the orientations are $0^\circ$, $+45^\circ$, $90^\circ$ and $-45^\circ$.

The Infrared band of the EM spectrum with radiance curves for reflections and thermal emissions. Emissions clearly dominate the spectrum in the LWIR (From [47]).

$P(\theta)$ curves for Uranium glass (shown in red) and Tungsten (shown in blue).

Scatter plot of data resulting from sampling $P(\theta)$ curves for Uranium glass and Tungsten and the addition of Gaussian noise.

All DoLP measurements in a given space-time volume as a function of the partial derivative, with respect to time (frame).

Points in the $n \times k$ parameter space with two clusters evident.

Material curves resulting from using cluster centers as parameter values. Data points colored to indicate true material-type (Uranium glass as red and Tungsten as blue).

Resulting $\theta$ trajectory from similarly classified data points. Arrows indicate errors in the classification indicated by marked deviations from the otherwise “smooth” trajectory.
Figure 5.12 Higher order statistical moments for each frame of a section of the walking simulation. ................................. 74
Figure 5.13 Correspondence between frames of the simulated walking video and the kurtosis value in each frame. ......................... 75
Figure 5.14 Correspondence between frames of the simulated running video and the kurtosis value in each frame. ......................... 76
Figure 5.15 Representative higher order statistical moments for each frame of a section of the collected data. ................................. 78
Figure 5.16 Kurtosis time-series sections for all actions as performed by the same individual. ................................. 79
Chapter 1

Introduction

Information is ubiquitous and may come in a variety of forms. Particularly, electronic information is available in vast amounts and is generated by a plethora of sensors and sources. For decision-making, especially, more information is always better. This immediately leads to the question of how to cope with a likely data deluge. Classically, the integration of such abundant information has required many individuals, specialized in their particular process or data-type, to work in tandem to analyze all of the individual data streams, and then to report to a single individual or group who will perform further analysis, on the whole, to draw conclusions, and act accordingly. Now, with the immense amount of available data at our disposal, and the need for real-time decision-making becoming more critical, we can neither supply the manpower nor the time to perform analysis in the classical way. One proposed solution of such an impending juggernaut is layered sensing.

Defined in [16], layered sensing is a framework for aggregating and presenting information. As the name implies, the framework assumes multiple “sensors,” or information sources, as input. The result of exploiting this framework is an aggregated presentation for use by individuals/groups to improve their understanding of the scenario at hand. Furthermore, as it has been proposed by the Layered Sensing Leadership Group (LSLG) at the Air Force Research Laboratory (AFRL). The focus of exploiting such a tool is termed “situational awareness.” Decision makers have access to a wealth of information germane to a scenario within a coherent setting which would otherwise be unavailable.
The novelty and utility of such a framework is clear when one considers the plethora of information sources at hand, e.g.: video, vibration sensors, satellite photography, processes such as object tracking, speech recognition and behavior analysis. A look at the implications and challenges of this framework can be found in [99].

Simplistically, layered sensing is the idea that all of the data pertinent to a chosen goal can be automatically amalgamated into a kind of omniscient data oracle, thus enabling better informed and timely decision-making. The output of a layered sensing framework, when applied to a particular problem, would enhance the decision-makers’ knowledge of the scenario by providing a coherent presentation of the whole relevant data; it is such an integration and the advantages thereof that we consider.

A key tenet of layered sensing is the integration of the different data sources with one goal (application) in mind. The goal of this effort grew out of a large set of problems known as remote sensing. One of the earliest definitions for remote sensing is attributed to D. A. Landgrebe, now professor emeritus at Purdue University:

“...[the] science of deriving information about an object from measurements made at a distance from the object”.

While this definition is simple and does succinctly capture the essence of the task, it does not elucidate on the many complexities and multifaceted approaches presently associated with remote sensing. A useful and directive approach is to abstract remote sensing as a system [19], whose process is indicated in Fig.1.1. This formulation facilitates the study of the concept in a structured manner and acts as a guide to related problems.

The particular application in mind for the effort in this thesis is human activity analysis. The physical object, in this sense, is the motion attributed to a person(s). The distinction of this work and the reason for the system approach utilized is the starting point: the sensor data.

When approaching human activity analysis, the data utilized is normally captured video from ground-level with many pixels on the subject. Also common to such efforts are the associated requirements of pre-processing on the data. If these requirements are slightly relaxed to allow for fewer pixels on the subject or greater distance between the sensor and the scene, and possibly the presence of multiple subjects, techniques change
Figure 1.1: Information flow in a Remote Sensing process. The four stages/requirements are: 1) the existence of an object or phenomenon, 2) the collection of data about the object/phenomenon, 3) information extraction through analysis of the data and 3) application of the extracted information.

accordingly. These techniques focus on either the environment within the scene or on the subjects. At this level, they are arguably inextricably linked [c.f. Chapter 4], and information is lost by focusing on only one aspect. If, again, the constraints are relaxed such that the sensor is now a great distance from the scene to the extent that the whole scene cannot be captured by a single sensor, the approach must be changed. The data from each individual sensor must be coalesced into a whole for analysis.

Our contribution is a complete approach for digesting such large-scale data and for analyzing it systematically to provide comprehensive descriptions at all of these varying scales. Starting at this most relaxed setting of multiple sensors located at distance from the scene, we combine the data from individual sensors to form a complete representation of the scene, described in Chapter 3. Once such a scene is composed, the scale is too great for detailed analysis to be tractable. The portion of our system to handle this stage, described in Chapter 4, breaks such a large scene into regions of interest based on the constituents of the environment as well as the motion occurring therein. The result is a description for each region via its environmental components and the motion/activity taking place. Again, common approaches to describing individual motion are not suited to handle so many subjects in so many diverse environments [c.f. Chapter 2]. Finally,
in Chapter 5, to address each region of interest, we present a novel technique to generate activity signatures. Furthermore, we leverage the motion signatures to discriminate between the materials present on the subject.

Descriptions of background material and existing techniques are left to their related chapter. However, common tools and methods utilized throughout the work are described in the following chapter.
Chapter 2

Prior Related Research

2.1 Image Mosaicking

The methods followed by many current efforts in image mosaicking such as [48], and others enumerated in [51] and [74], can be summarized as follows. The first step is to compute feature descriptors of the image tiles. A standardized feature point selection process is applied to each image. This step usually consists of computing some generic feature descriptor; many of which have been developed for such tasks and successfully employed repeatedly, e.g. SURF [9] described in §2.1.1. The techniques may vary in the choice of feature descriptor. Once each image is succinctly described, a search must then be performed across all possible pairings of these descriptors. The descriptors are compared using some measure, e.g. RANSAC [37] described in §2.1.2. The parsimony with which this comparison is performed, i.e. whether each descriptor set is exhaustively compared to every other descriptor set, is another facet by which the various techniques vary [44]. The scoring method produces a quantitative correspondence between each of the candidate pairings. These values are then compared to compute the best match between some subsets of those points.

A set of corresponding points \( \{x_j^1, \ldots, x_j^p\} \in U_j \) and \( \{x_k^1, \ldots, x_k^p\} \in U_k \) is then considered. This set of matching points is used to compute the homography [106] which is subsequently applied to the appropriate images. For example, to align the respective
images $I_j$ and $I_k$ via these point sets, a homography $H$ must be computed such that

$$x_i^j = Hx_i^k, \forall i.$$  \hspace{1cm} (2.1)

The result is computed by solving the system of equations in Equation (2.1) as described in [1, 106]. Note, the feature descriptor utilized and the scope within which the search for correspondences is performed, vary from application to application.

### 2.1.1 Feature-driven Analysis

**Scale-Invariant Feature Transform (SIFT)**

The scale-invariant feature transform (SIFT) was introduced by Lowe in [61]. This transform consists of computing, from data, a set of descriptors. In this context, the data are usually in the form of an image. The descriptors are vectors of gradient-based distributions about key-points in the image. These descriptors summarize the distinct components within the data. They are, additionally, invariant to common transforms in imagery: scaling, translation, rotation and, to some degree, shearing. The produced descriptors are thus frequently referred to as affine-invariant. Consequently, SIFT is still commonly utilized for applications such as panorama stitching and object detection.

The general process followed to compute these descriptors is presented next. Candidate features are detected across all scales of the input image. For this, a Gaussian scale-space is applied:

$$L(x, y; \sigma) = I(x, y) * g(x, y, \sigma),$$ \hspace{1cm} (2.2)

where $\sigma$ is the standard deviation of the Gaussian function, “*” the convolution operator, $I(x, y)$ is the input image, and $g(x, y)$ the Gaussian function

$$g(x, y, \sigma) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) \cdot e^{-\frac{x^2+y^2}{2\sigma^2}}.$$ \hspace{1cm} (2.3)

To extract the features in the scale-space, the local extrema in the difference-of-Gaussian function $D(x, t, \sigma)$ are computed. The difference-of-Gaussian (DoG) is defined
as the difference of two neighbouring images of the scale-space:

\[ D(x, y, \sigma) = L(x, y; 2\sigma) - L(x, y; \sigma) = (g(x, y, 2\sigma) - g(x, y, \sigma)) \ast I(x, y). \quad (2.4) \]

Each sample point of the DoG is compared to its eight neighbours in the current scale and nine neighbours on the levels below and above, i.e. points in the current and adjacent scales that are 8-connected to the candidate. In this manner, extrema within the scale-space are determined. These locations are taken as an initial set of candidate feature.

Generally, this set of candidate features is prohibitively large for computations and many such features may be the result of artifacts in the data. Features, or key-points, are considered unstable when they correspond to image noise or boundary effects. To detect such unstable points, a spline is locally fit to the extrema via a Taylor expansion of Equation (2.4). A ratio of the edge responses around the key-point is computed via the Hessian of \( D \):

\[ \mathbb{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}, \text{ where } D_{xy} = \frac{\partial^2 D}{\partial x \partial y}. \quad (2.5) \]

An extremum is considered to be unstable if there is great discrepancy between the principal curvatures of \( D \) in its neighborhood. The ratio of the trace of the Hessian, \( \text{Tr}(\mathbb{H}) \), to its determinant, \( \text{Det}(\mathbb{H}) \), is proportional to the difference between its eigenvalues.

\[ R = \frac{\text{Tr}(\mathbb{H})}{\text{Det}(\mathbb{H})} = \frac{D_{xx} + D_{yy}}{D_{xx}D_{yy} - D_{xy}^2}. \quad (2.6) \]

Thus, this ratio is also proportional to the difference in principal curvatures of \( D \). A threshold for this ratio is chosen and key-points with corresponding ratios below this threshold are kept as features.

Each retained feature is assigned an orientation and scale based on the prevailing gradient directions in the neighbourhood of its location in the image. This orientation is used to establish a reference frame for the construction of the descriptor. At the determined scale for each key-point, a set of histograms describing the distribution of gradient values within a \( 16 \times 16 \) pixel neighborhood of each key-point is produced. The resulting 128-dimensional vector is the SIFT feature descriptor. Since the computation of the descriptor is performed using the local reference frame (orientation, scale and neighbor-
hood), the resulting descriptor is invariant to transformations that affect this reference frame in a covariant manner but only partially invariant to affine transformations and illumination changes.

**Speeded-Up Robust Features (SURF)**

The Speeded-Up Robust Features (SURF) algorithm, introduced by Bay et al. [9], was influenced by the SIFT algorithm described above and is based on the Hessian-matrix approximation and the computation of Haar wavelet responses. Similarly to SIFT, features are detected in a scale-space to achieve invariance to scaling. In this approach, however, the scale-space is not constructed explicitly and integral images are instead employed to decrease the computation time as compared to the convolution required for SIFT. The SIFT algorithm estimates the Hessian matrix of the convolution of the image with the second derivative of the Gaussian. The trace and determinant of this matrix express the local change around the key-point. The points that are simultaneously local extrema of both the determinant and the trace of the Hessian matrix are chosen as candidate key-points.

Before the descriptor calculation, in order to achieve invariance with respect to image rotation, it is necessary to determine the local orientation of the feature in a manner similar to the SIFT algorithm. As in the case of the SIFT algorithm, the SURF descriptor is a vector describing the distribution of intensity values within the neighbourhood of the keypoint.

Despite the computational efficiency of the convolution operator in SIFT, a faster method utilizing integral images is employed in the SURF algorithm. In SIFT, the convolution is performed using a kernel that approximates the second derivative Gaussian kernel reasonably well, yet it is possible to evaluate the operation using a limited number of queries to the integral image structure via a Haar wavelet. Recall that the Haar wavelet responses can be calculated efficiently using integral images since they correspond to convolutions with kernels containing predominantly rectangular structures.

A SURF feature describes the surrounding area of the key-point. This area is divided into $4 \times 4$ subareas. In each of these subareas, the Haar wavelet responses are calculated.
in the $x$ and $y$ directions and then are described by the vector,

$$v = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|),$$

where $d_x$ and $d_y$ are the wavelet responses in $x$ and $y$ directions. Concatenating these for all $4 \times 4$ subregions gives a 64-dimensional vector descriptor.

In addition to this boost in performance, the SURF algorithm does not need to build the scale-space by explicitly blurring and down-sampling the images.

### 2.1.2 Random Sample and Consensus (RANSAC)

Random Sample Consensus (RANSAC) [37] is a non-deterministic iterative algorithm for estimating parameters of a mathematical model given a dataset. RANSAC provides a general technique for model fitting in the presence of outliers.

The RANSAC algorithm informally proceeds as follows:

1. Initially, a subset of the data (candidate inliers), is selected randomly and used to estimate the free parameters of the model.

2. The goodness of fit of the model is computed.

3. Additional data points are added to the subset.

4. The model parameters are re-estimated.

5. The model is evaluated by estimating the error relative to the model. If the error decreased, the additional points are considered inliers. Otherwise, the added points are removed from the inlier set and different points are considered.

A model is sufficient if either enough points are considered inliers or some desired tolerance bound on the error is reached. In this way, model parameter estimation and outlier detection are simultaneous performed. The models where RANSAC is most commonly used are homography estimation and fundamental matrix estimation.

For homographies, the minimal number of required correspondences is 4 to account for transformations such as rotation, translation, scaling and shearing. The error is computed by applying the estimated homography to feature descriptors from one view/image
and considering how closely, i.e. within an “epsilon ball”, they are mapped to their corresponding matches in the second view/image.

2.1.3 Current Methods for Mosaic Comparison

Comparing mosaics is a difficult task because a reference for comparison is not usually available. Mosaicking algorithms produce new imagery, not reconstructions of existing imagery, so ground truth isn’t readily available. Additionally, a standard dataset hasn’t evolved in the image processing or machine vision communities on which to compare the performance of mosaicking algorithms. As indicated in [18], current efforts follow two paradigms for testing. The first takes a large-scale image of a scene and breaks it into tiles, from which the mosaicking algorithms reconstruct the large scene. The second takes both the large-scale image of the scene as well as many smaller images of portions of the scene. This second method experimentally captures the data that is synthesized by the first. Since the original data is available following either of these procedures, the most readily accepted means of comparison are those used in other signal processing applications. Some efforts utilize the mean-squared error (MSE) between pixel values of the ground-truth data and the produced mosaics. Another popular measure is the earth mover’s distance (EMD) [75], a measure of the difference between two distributions. In this application, the distributions are those of the intensity values of the mosaic and the ground truth. The basis of this measure is the idea of comparing mounds of dirt, i.e. distributions. The result of the EMD is the amount of energy required to convert one dirt pile into one the same size/shape as the other. Again, as shown in [75], this can be computed by exhaustively considering the myriad of ways in which elements from the histogram of one distribution can be permuted to produce the histogram of the other distribution. The EMD would be the minimal number of movements required to complete this task.
2.2 Scene Understanding

Current techniques focus on describing either the environment or the motion/action with probability density functions (mixture models, etc.). Some of the efforts that endeavor to describe the environment utilize segmentation techniques to separate the constituents, e.g. [54]. However, many utilize a multi-level approach to segment the constituents sequentially, e.g. [59] and [57], utilizing learned or defined distributions to classify different regions.

Works focusing on describing and/or classifying the actions in a scene are varied in approach. While some, e.g. [104], do utilize multi-level techniques for robustness, some, e.g. [86], also learn parameters to fit dynamical models while, still, others, e.g. [95], utilize statistical inference from graphical models.

While these techniques perform well and solve their intended problem, they do not address the problem of comparing scenes for both environment and activity. Only the description of one or the other (environment or motion/action) is available.

2.3 Activity Analysis

Human action analysis, as a specific problem in remote sensing and an application of computer vision, is an important focus of much current research [10, 60, 83]. While no standard definitions have been formulated, a common taxonomy utilized to distinguish between varying scopes of study is emerging.

1. An *action* is a simple, low-level motion such as “raising an arm” or “walking.”

2. *Activities* are combinations of actions, e.g. “waiting for a bus.”

3. An instant where multiple individuals perform a variety of activities is called an *event*, e.g. a football game.

4. A *behavior* is a pattern in an individual’s or group’s activities.

Many techniques and approaches have been implemented to tackle the problem of detecting, recognizing, identifying and then exploiting human action descriptions. A recent
review [2] separates the approaches into two main categories: “single-layer” and “hierarchical”. The latter category includes techniques that describe human activities by examining layered deconstructions of those activities. From those deconstructions, the classification for the activity is determined. Efforts in this direction can be more specifically described by the approaches they use to describe and combine the deconstructed pieces: statistics/Hidden Markov Models (HMM’s) [72], syntactic grammar organizations [50,98] or atomic-actions/subevents [3,42].

In contrast, the single-layer techniques directly utilize images or video to conclude the activities’ classifications. In [2], the single-layer approaches are further divided into space-time and sequential approaches. This further segregation is based on the method with which the efforts manage the temporal dimension of the data. Space-time approaches consider the temporal dimension as a third spatial dimension in that a video of an action is analyzed as a volume [14,52,80,89] or as local features [71,87,105] or trajectories within that volume [20, 77, 90]. Sequential methods abstract the input videos as sequences of measurements from which feature vectors are extracted. These features are used to classify action sequences by either updating a state model that is compared to a trained state sequence for each action [64,73,103] or by comparing them to templates of expected action sequences [31,41,97].

2.4 Statistical Tools

2.4.1 Statistical Moments

Consider a vector \( \mathbf{x} \in \mathbb{R}^n \) of \( n \) samples \( x_i, i \in [1,n] \), from a continuous random variable \( x \in \mathbb{R} \). One alternative approach to describing a probability distribution function \( f(x) \) of \( x \) is via its statistical moments. These moments quantitatively describe the shape of a distribution so as to facilitate its classification and identification. The \( k^{th} \) moment of the function \( f(x) \) centered about the value \( c \) is described by Eq.(5.16),

\[
\mu_k = \int_{-\infty}^{\infty} (x - c)^k f(x) dx.
\] (2.7)

The mean, or expected value \( \mu = \text{E}[x] \), of a random variable can be described as
the first moment of the distribution centered about zero. Moments centered about the mean of the variable are called “central moments”. This makes the central moments dependent only on the shape and scale of the distribution. Of particular focus in this effort are standardized, central moments that have been normalized by the $k^{th}$ power of the variance $\sigma$. Here, again, a level of invariance is added to the moments in that standardized moments are influenced strictly by the shape of the distribution and not by the location or scale. Commonly, in conjunction with the mean and the variance, the third and fourth standardized moments (the skewness and kurtosis, respectively) are also considered to describe a distribution,

$$\text{Skewness} = \gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mathbb{E}[(x - \mu)^3]}{(\mathbb{E}[(x - \mu)^2])^{3/2}}, \quad (2.8)$$

$$\text{Kurtosis} = \gamma_2 = \frac{\mu_4}{\sigma^4} = \frac{\mathbb{E}[(x - \mu)^4]}{(\mathbb{E}[(x - \mu)^2])^2}. \quad (2.9)$$

While these two descriptors seem similar, the difference in the exponent is key to the difference in describing the shape of the distribution. The skewness of a distribution is a measure of the asymmetry of that distribution, as shown in Figure 2.1. A normal or symmetric distribution would have a skewness of zero. The kurtosis of a distribution describes the peakedness of the distribution, as shown in Fig 2.2. So, for the random vector $x_i$, these moments describe the shape of the distribution of the random variable $x$. To calculate such moments, the above equations must be rewritten in terms of the sample vector, as in Eqs. (5.17) and (5.18):

$$\text{Skew}(x) = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^3}{\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2\right)^{3/2}}, \quad (2.10)$$
Figure 2.1: Illustration of the sign of the skewness and its relation to the asymmetry of the distribution. 2.1(a) - negative skewness, 2.1(b) - normal distribution, 2.1(c) - positive skewness.

Figure 2.2: Illustration of the kurtosis and its relation to the peakedness of the distribution. 2.2(a) - low kurtosis, 2.2(b) - normal distribution, 2.2(c) - high kurtosis.

\[
\text{Kurt}(x) = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^4}{\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2\right)^2}.
\] (2.11)

The combination of these higher order statistics with the location and spread information contained in the mean and variance, respectively, provide a good description of a distribution, and allow for a significant comparison between distributions.

### 2.4.2 Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test is a statistical test for comparing the mean rank of populations from two related samples. That is, two populations are sampled pair-wise, the
difference between the elements of each pair are computed and the resulting differences are ranked. The median of this ranking is determined. If this median is zero, the two populations are the same. However, if the median is non-zero, two samples are determined to have come from separate and distinct populations. The details of the test procedure follows.

The test is formalized by proposing two hypotheses:

\( H_0 \): the median difference between the pairs is zero,
\( H_1 \): the median difference between the pairs is non-zero.

One then proceeds to sample the two populations, in pairs, \( N \) times. This produces sample sets \( x_i^1 \) and \( x_i^2 \), for \( i = 1 \ldots N \). The difference and the sign of the difference between these pairs is computed, \( |x_i^2 - x_i^1| \) and \( sgn(x_i^2 - x_i^1) \), where \( |\cdot| \) is the absolute value and \( sgn(\cdot) \) is given by:

\[
sgn(x) := \begin{cases} 
-1 & \text{if } x < 0, \\
0 & \text{if } x = 0, \\
1 & \text{if } x > 0.
\end{cases}
\] (2.12)

The resulting differences are ranked smallest to largest (i.e. the smallest is ranked 1), and the ranks are assigned to corresponding variables \( R_i \in \mathbb{Z^+} \). Similarly, the result of the \( sgn \) function is assigned to variables \( S_i \). The test statistic, \( W \), is computed thus:

\[
W = \sum_{i=1}^{N} S_i R_i .
\] (2.13)

For large \( N \), the sampling of \( W \) converges to a normal distribution. Hence, the \( z \)-score can be computed:

\[
z = \frac{\sum_{i=1}^{N} S_i R_i}{\sqrt{\sum_{i=1}^{N} (S_i R_i)^2}},
\] (2.14)
For a desired significance level, this z-score can be utilized in the usual way to compute a \( p \)-value for the test. For this test, a \( p \)-value close to one (1), would indicate that the null-hypothesis, \( H_0 \), should be accepted, that the two samples come from the same distribution. However, a small \( p \)-value close to zero, would indicated that \( H_0 \) should be rejected in favor of \( H_1 \).

### 2.5 Robust Subspace Recovery via Bi-Sparsity Pursuit (RoSuRe-BSP)

Consider a data set \( L \in \mathbb{R}^d \) uniformly sampled from a union of subspaces

\[
S = \bigcup_{i=1}^{J} S^i,
\]

then assuming a sufficient sample density, each sample can be represented by the others from the same subspace with probability 1. Mathematically, we represent the data matrix by

\[
L = [l_1 | l_2 | \ldots | l_n], \quad \text{yielding} \quad L = LW,
\]

where \( W \) is \( n \times n \) block-diagonal matrix.

It is worth noting that, to recover the underlying data sampled from UoS, it is equivalent to finding a matrix \( L \) and \( W \) under the above constraints. Let us now define, mathematically, what the method seeks.

**Definition 1** (k-block-diagonal matrix) *We say that a \( n \times n \) matrix \( M \) is k-block-diagonal if and only if*

1. There exists a permutation matrix \( P \), such that the matrix \( \tilde{M} = PMP^{-1} \) is a block-diagonal matrix

2. the maximum dimension of each block of \( \tilde{M} \) is less-than or equal to \( k + 1 \).

*The set of all such matrices is denoted as \( BM_k \).*

We next define the set of self-representative matrices, based on the space \( BM_K \) as follows:
**Definition 2** (k-self-representative matrix). We say that a $d \times n$ matrix $X$ with no zero columns is k-self-representative if and only if

$$X = XW, W \in BM_k, w_{ii} = 0.$$  

The space of all such $d \times n$ matrices is denoted by $SR_k$.

The problem can then be formulated as

$$\begin{align*}
\min & \|W\|_0 \\
\text{s.t.} & \quad X \in SR_k.
\end{align*}$$

(2.17)

where $\|\cdot\|_0$ is the $l_0$ vector pseudo-norm. We have a fundamental difficulty in solving this problem on account of the combinatorial nature of $\|\cdot\|_0$ and the complex geometry of $SR_k$. For the former one, there are established results of using the $l_1$ norm to approximate the sparsity of $W$ [21] [101]. The real difficulty, however, is that $SR_k$ is a non-convex set and, thus the minimization process is often unattainable, in general.

To alleviate this problem, we opt to integrate this constraint into the objective function, and see the problem from a different angle. We hence have the following definition:

**Definition 3** ($W_0$-function on a matrix space). For any $d \times n$ matrix $X$, if there is $W \in BM_k$, such that $X = XW$, let

$$W_0(X) = \min_W \|W\|_0, \quad s.t. \quad X = XW, w_{ii} = 0, W \in BM_k.$$  

Otherwise, $W_0(X) = \infty$

Then instead of Eqn(2.17), we consider the following optimization problem:

$$\begin{align*}
\min_{L,E} & \quad W_0(L) \\
\text{s.t.} & \quad X = L \in SR_k.
\end{align*}$$

(2.18)

Next we will leverage the parsimonious property of the $l_1$ norm to approximate $\|\cdot\|_0$. First, the definition of $W_0(\cdot)$ is extended to an $l_1$ norm based function:
**Definition 4** (W₁-function on a matrix space). For any \( d \times n \) matrix \( X \), if there exists \( W \in BM_k \), such that \( X = XW \), let

\[
W_1(X) = \min_W \|W\|_1, \quad s.t. \ X = XW, w_{ii} = 0, W \in BM_k.
\]

Otherwise, \( W_1(X) = \infty \)

We then have the following problem,

\[
\min W_1(L) \quad \text{(2.19)}
\]

\[ s.t. X = L \in SR_k \]

The sparse support of \( L \) lies in the matrix \( W \) in the \( W_0 \) function, meaning that columns of \( L \) can be sparsely self-represented.
Chapter 3

Gigapixel Image Mosaicking via Parallel Sparse Tile Decomposition

The rate and quantity at which photographs are being produced and published has tremendously increased over the last decade. Frequently, many of these photos depict the same object or follow the same scenes, etc. The ability to combine such images to synthesize a single detailed representation would afford the viewer with a succinct and comprehensive understanding of the captured data. In this work, we compute such a mosaic for data captured by the prototype camera AWARE-2. This device produces many high-resolution images of a single scene. Our contribution is a theoretically sound and rigorous approach to combining these high-resolution images into a single consistent gigapixel image.

Image mosaicking is, like its arts and crafts namesake, the process of combining multiple small images (called tiles) that possible aren’t meaningful on their own, to produce a single larger and complete image. In computer vision, this refers more specifically to the combination of images of the same scene/subject to produce a more complete representation of that specific scene or subject. Particular instances of this application may be as local as the production of panoramas on an individual’s mobile phone or, as global as the combination of images collected from users all over the world and even as broad-reaching as combining astronomical images to create a picture of the cosmos.

Many current methods for mosaic formation follow the framework as described in §2.1.
Some preliminary output of a naive version of this process is shown in Figure 3.1. There are discontinuities, due to misalignments, visible in many portions of the constructed mosaic. This construction was performed without the aid of any additional information, i.e., only the images were used as input to the process. Here, we describe a method for computing more accurate mosaics by utilizing structure in the imaging array as well as in the images themselves. Unlike the feature descriptor-based techniques that must be tuned for each application, our global, data-driven technique requires no tuning, and is capable of producing the homography between images in parallel.

![Mosaic output from seven images using current state of the art tools.](image)

Figure 3.1: Mosaic output from seven images using current state of the art tools.
3.1 Background

3.1.1 Data

The apparatus by which the images utilized in this work were captured is a prototype camera array named AWARE-2. This is a product of the Duke Imaging and Spectroscopy Program led by Dr. David Brady at Duke University in Durham, NC, USA. This device consists of a semi-spherical array (shown in Figure 3.3(a)), of 98 micro-cameras, each with a 14-megapixel resolution [15]. Sample images from these micro-cameras are shown in Figure 3.2. The resulting imagery can be combined to form single images with resolutions nearing 1-gigapixel.

While the camera apparatus is constructed to hold approximately 200 micro-cameras, currently only 98 are populated. The configuration of these selected micro-cameras is illustrated in Figure 3.3(b), with a view of one looking into the imaging sensors.

3.1.2 Problem Formulation

Suppose that we are given an image $I$, and a set of sub-images (tiles):

$$\mathcal{I} = \{I_1, \ldots, I_s\},$$

where the $I_i$’s can be viewed as an intensity function from the region of the image-plane $U_i$ to a set of positive integers $[0, 255]$ for the case of gray-scale mosaicking. For simplicity, we will use the same symbol $I_i$ to denote the image as a collection of pixels with an intensity value associated to them and the corresponding function. With minimal modifications, we can extend this procedure to colored pictures by changing the function $I_i : U_i \to [0, 255]^3$, (RGB-values).

We say that two sub-images $I_1, I_2$ are related if $I_1 \cap I_2 \neq \emptyset$. We can then turn the set of sub-images into a graph where the nodes are the sub-images, and an edge exists between nodes $I_x$ and $I_y$ if they are related. We denote this graph also by $\mathcal{I}$, since it is completely determined by the set $\mathcal{I}$ for a given picture $I$.

*http://www.disp.duke.edu/projects/AWARE
Figure 3.2: Two example high-resolution images from the micro-camera array.

In the mosaicking problem we require the two following conditions to hold:

1. $\mathcal{I} = \bigcup_{i=1}^{s} I_i$ and

   \begin{equation}
   (3.2)
   \end{equation}

2. the graph of the sub-images is connected.

The second property implies that there is an overlap between each tile and some other neighboring ones, hence, a matching can be produced between them. This matching, or
mapping, between pairs of images is the crux of image mosaicking. Since our graph of sub-images is connected we can create a process which will gradually stitch neighboring images together to yield a global picture. The first property guarantees that the output picture will be achieved by this mosaicking.

### 3.2 Image Decomposition

From the geometry of the camera array [91], we can extract the relative positions of the sub-images. Using this information in conjunction with the extents of the sub-images, it was possible to compute the intersection, or overlap, between each sub-image pair and form our graph of sub-images.

The neighborhood of a sub-image $I_i$ is defined to be the set of sub-images related to $I_i$, or in other words, the set of sub-images that have some non-empty intersection with $I_i$:

$$\mathcal{N}_i = \{I_j | I_i \cap I_j \neq \emptyset\}. \quad (3.3)$$
Clearly, $I_i \in \mathcal{N}_i$. In the graph theoretic language this is the *vertex star* of the node $I_1$. An example graph formation from image overlaps is depicted in Figure 3.4.

Given a pair $(I_i, I_j), I_j \in \mathcal{N}_i$, their extent is used to compute the size and shape of the overlap between them. By performing this computation for all such pairs including $I_i$, the minimal overlap between $I_i$ and any of its neighbors is determined. We compute the rectangle with maximal area, entirely contained within this overlap and denote its dimensions by $p \times q$. We next partition the sub-image $I_i$ and all $I_j \in \mathcal{N}_i$ into $\frac{m}{p} \times \frac{n}{q}$ regions, to obtain

$$I_i = \bigcup_{k=1}^{\frac{m}{p} \times \frac{n}{q}} I^k_i.$$  \hfill (3.4)

This partitioning, although seemingly increases the complexity of the algorithm by increasing the number of tiles, it ultimately decreases the complexity of the problem by decreasing the size of the regions being compared. Further simplification is gained by allowing for the comparison of only the overlapping regions.
Figure 3.5: Partitioning of tile into minimal regions.
3.3 Sparse Subspace Recovery

In contrast to existing methods in image mosaicking, as described and referenced in §2.1, the proposed approach highlights similarity between portions of the image data itself. By computing a low-rank representation of the query image over the support of the column-space of its neighborhood, we are able to localize the overlap among the sub-images. This method follows closely that of Robust Subspace Recovery via Bi-Sparsity Pursuit (RoSuRe-BSP), described in [12] and recounted in §2.5.

For a particular sub-image $I_i$ with $|\mathcal{N}_i| = s$, we construct a dictionary matrix $X$ whose columns are the columns of the regions $I^k_j$ for all sub-images $I_j$ in the neighborhood $\mathcal{N}_i$ of $I_i$:

$$X = [I^1_1, \ldots, I^m_{1^n}, I^1_2, \ldots, I^s_{1^n}].$$  

Having constructed this dictionary whose atoms are columns representing portions of the neighborhood, we proceed to optimize for a low-rank representation, Equation (2.19).

$$\min_i \|W_i\|_1, \text{ s.t. } X = XW_i \text{ and } W_i \in BM_k.$$  

Here, $\| \cdot \|_1$ is the vector $l_1$ norm. For a matrix $X$ and an index set $J$, we let $X_J$ be the submatrix containing only the columns of indices in $J$.

As shown in [12] and described in §2.5, this procedure will result in a sparse and low-rank coefficient matrix $W$ that is also block diagonal. The blocks in $W$ will correspond to the different subspaces inherent in the data represented by the matrix $X_i$. In this particular application, these subspaces are the intersections between the regions. Additionally, the sparsity of the coefficients guarantees high values indicating correspondence between the similar columns in those regions. As a result, all the computations are performed on small, local neighborhoods without the need for global information about the mosaic. This is the parallel nature of the method.

In the following sections we demonstrate this method’s ability to define and localize similarity between images, while expounding on the structure of the coefficient matrix and how it can be leveraged to construct the desired mosaic.
3.4 The Structure of the Coefficients

3.4.1 Synthetic Data

To evaluate the proposed method for representation, increasingly realistic data sets were utilized as input. The first two tests focused on synthetically generated mosaic tile sets. The synthetic data was produced from a larger test image, shown in Figure 3.6, by spatial windowing.

Figure 3.6: Test image from which synthetic sub-images, or tiles, were generated. [23]

3.4.2 Tile Translation

To determine the proposed representation method’s functionality in regard to simple translation, the test image, Figure 3.6, referred to as $I_{orig}$, was partitioned into regular
blocks on a $3 \times 3$ grid:

$$I_{\text{orig}} = \begin{bmatrix} A & B & E \\ C & D & F \\ G & H & K \end{bmatrix}. \tag{3.7}$$

These blocks were then grouped into four overlapping sub-images as indicated in Equations (3.8)-(3.11).

$$I_1 = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{3.8} \quad I_2 = \begin{bmatrix} B & E \\ D & F \end{bmatrix} \tag{3.9}$$

$$I_3 = \begin{bmatrix} C & D \\ G & H \end{bmatrix} \tag{3.10} \quad I_4 = \begin{bmatrix} D & F \\ H & K \end{bmatrix} \tag{3.11}$$

Masks to indicate this subdivision of the test image can be seen in Figure 3.7.

In this experiment, we explicitly compute the similarity between the sub-images for both the column- and row-spaces. To compute this similarity amongst the columns, we follow the procedure exactly as outlined in §3.3 by constructing the dictionary and constraint as shown in Equation (3.12). For this test, we’ve decomposed the image such that there is minimal similarity between the columns of $I_1$ and $I_3$, as well as between the columns of $I_2$ and $I_4$.

$$X_c = X_c W_c, \quad X_c = [I_1, I_2, I_3, I_4] \tag{3.12}$$

For illustrative purposes, we have also selected the sub-images so that there is little similarity between the rows of $I_1$ and $I_2$, as well as between $I_3$ and $I_4$. To compute the row-space similarity, we simply need to transpose the sub-images, as shown in Equation (3.13).

$$X_r = X_r W_r, \quad X_r = [I^t_1, I^t_2, I^t_3, I^t_4] \tag{3.13}$$

Once the optimization from Equation (3.6) has been completed for both Equations (3.12-3.13), the similarity between the sub-images may be observed through the coefficient matrices $W_c$ and $W_r$, as shown in Figure 3.8. Red lines have been added to the coefficient matrices, artificially, to highlight the block structure. The lighter colors in the matrix represent higher-valued coefficients and the darker colors represent lower-valued coefficients.

Considering Figure 3.8(a), the bright bands of coefficients in the off-diagonal blocks
correspond precisely to the duplicate columns in the sub-images. For example, the high-valued coefficients in the first block of the second row, $W_c^{2,1}$, indicate that the last few
Figure 3.8: Coefficient matrices $W_c$ and $W_r$ are computed in Equations (3.12) & (3.13). Lighter colors indicate high values. Blocks in the matrix, delimited by artificially inserted horizontal and vertical lines, are the representation of the sub-image corresponding to the column in the space of the sub-image corresponding to the row.

(right-most) columns of $I_1$ are quite similar to the first few (left-most) columns of $I_2$, as expected from our construction of the test imagery. Similarity between the columns of $I_3$ and $I_4$ can also be determined by the high-valued coefficients in the last block of the third row, $W_{c}^{3,4}$, and the third block of the last row, $W_{c}^{4,3}$. Similarly, considering Figure 3.8(b) allows for the determination of the similarity between the rows of the sub-images. Knowing the correspondence between rows and columns of the images allows for recovery of the possible translation that is necessary to align the images. It is important to note, however, that not only is a correspondence indicated by these coefficient values but, also the strength of the similarity.
3.4.3 Tile Rotation

The other portion of the problem presented by the real data for this application is that of rotation of the sub-images. To test the proposed method’s robustness to rotation, the same test image shown in Figure 3.6 was divided into two overlapping sub-images, $I_1$ and $I_2$. The test proceeded by applying a rotation to $I_2$ about its centroid to emulate the possible rotation in the AWARE-2 data. An example of $I_1$ and $I_2$ rotated is shown in Figure 3.9.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image-decomposition.png}
\caption{Portion of test image not rotated: $I_1$. Portion of test image rotated: $I_2$. Shown with arbitrary rotation.}
\end{figure}

Figure 3.9: Image decomposition used for rotation tests.

Once this rotation had been applied, the RoSuRe-BSP optimization was performed. This procedure was repeated for several rotations of $I_2$. Selected blocks of the resulting coefficient matrices are shown in Figure 3.10 for three such rotations.

Figure 3.10(a) illustrates the type of coefficient band to expect in the blocks of $W$ for perfect alignment. Here, a single coefficient value in each column represents a $1 - 1$
correspondence between the columns in $I_1$ and the columns in $I_2$. When $I_2$ was rotated, the local self-similarity in $I_{orig}$ allowed for matching to still occur between columns of the sub-images. A spreading in the coefficient band can, however, be observed in Figures 3.10(b)-3.10(c). As the rotation of $I_2$ is increased, the width of the band is increased. The case illustrated in Figure 3.10(a) is, indeed, a global minimum for the width of the coefficient band. The width of this band can, consequently, be exploited to compute the appropriate rotation of $I_2$ to best align with $I_1$. Besides the actual white bands that represent common features, one can see in Figure 3.8 other non-zero entries of the matrix $W$. The reason is two-fold:

1. The self-similarity within the natural scene $[65, 82, 96]$ and
2. artifacts coming from the RoSuRe-BSP implementation.

The prescribed decomposition was performed for 40 test images at 360 different rotations of $I_2$, covering a full circular rotation in $1^\circ$ increments. To illustrate the degradation of the detection of these bands, the signal-to-noise ratios (SNR) for the examples in Figure 3.10 were computed and are shown in Table 3.1.

The maximum rotation that allowed for detection of the strong coefficient band was $5^\circ$. This rotation tolerance will be utilized in future imaging as a calibration parameter for the insertion of the micro-cameras into the array.
Table 3.1: Table of SNR values for band of strong coefficients in $W$ corresponding to the listed rotations in $I_2$. SNR was computed in the standard way for the test images. [36].

<table>
<thead>
<tr>
<th>Rotation</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0°</td>
<td>14.84dB</td>
</tr>
<tr>
<td>0.5°</td>
<td>8.22dB</td>
</tr>
<tr>
<td>1°</td>
<td>5.55dB</td>
</tr>
</tbody>
</table>

The simple fact that RoSuRe-BSP is producing useful decompositions when the input images are rotated at all deserves some discussion. The rotations carried out in these experiments and on the real data, are not mathematically tractable. The physical rotation of a picture leads to a highly non-linear transformation of the corresponding pixel representation matrix. When working with matrices, it is theoretically unreasonable to expect similarity between a matrix and its rotated version. However, in this case we are targeting “natural” images where there is some smoothness of the features from column to column. Even when an image is rotated, local continuity of the intensity allows small patches to provide redundance for a linear combination of similar columns. The success of our proposed method hinges on the local homogeneity.

3.5 Results on Captured Data from AWARE-2

The proposed method was applied to all neighborhoods of each of the 98 images captured by the AWARE-2 camera array for a football scene. The resulting mosaic is shown in Figure 3.11. The final image had a resolution of 0.38-gigapixels and was 483.3MB in size.

The technique was also applied to six other scenes captured by the AWARE-2 camera array. The results were of similar size and resolution. The layout of the microcameras, however, was adjusted between collections on account of continued development of the apparatus. This change in the population of the array resulted in a slight variation of the aspect ratio and overall size/shape of the mosaics produced. Some of the mosaics computed for these scenes are depicted in Figures 3.12, 3.13 and 3.14. These figures also demonstrate a more obvious example of the differences due to the configuration of the
3.6 Evaluation

A small sample of Figure 3.11 is shown in Figure 3.15, in comparison to the same portion produced by the methods described in §2.1. In this specific instance, our method is shown to not suffer from a particular boundary effect and corrects the floating head. Since all methods suffer from some error-inducing effects, an objective measure of quality would prove invaluable. We do note in §2.1.3 two measures and the methods by which they are currently utilized for mosaic comparison. However, in spite of these techniques as noted in [18], there is no standard way to compare the results of image mosaicking algorithms except a visual comparison. This is partly due to the lack of a means of quantifying human subjective interpretation.

We propose a specific method, leveraging the subjective analysis currently employed, to quantify the quality of the mosaics constructed by various methods for comparison. Seven natural scenes were observed via the AWARE-2 gigapixel camera and seven sets of
micro-camera images were used. Each of these image sets was utilized as input to each of the three mosaicking algorithms. The result was seven sets of three mosaics. A poll was devised to quantize the difference in perceived quality of the output mosaics. The nature of the “quality” measure was intentionally deferred to a subjective decision by the participants. In this way, the opinion of the ideal end users, and hopefully people at large, was collected.

The poll consisted of a set of ranking tasks for all participants. Three mosaics of the same scene were simultaneously presented and each participant was prompted to “…rank these images as either ‘Good’, ‘Better’, or ‘Best’…” . The three mosaics were the three outputs from each of the algorithms, presented in a random order and without labels indicating such. Each time a participant submitted a ranking, a new scene and corresponding mosaics were immediately presented for evaluation. At the beginning of each participant’s session, the order in which the scenes were to be presented was chosen randomly. Following this schedule, each user was prompted to rank the output of each algorithm for all scenes such that no bias was induced by the polling method. Additionally, to prevent artificial bias/skew in the results, and possible frustration of the participants, ties were allowed between mosaics.

A total of 595 sets of mosaics were ranked, i.e. each scene was ranked 85 times. To numerically interpret the results of the poll, a score was assigned to each rank: ‘Good’ ⇒ 3,
Figure 3.13: Computed mosaic of graduation scene #2. The lack of a micro-camera in the array caused a blank space in the middle of the produced mosaic.

Figure 3.14: Computed mosaic of park scene.

‘Better’⇒2 and ‘Best’⇒1. The better algorithm would thus produce an average score closer to 1 than 3. The veracity of such a comparison required that the rankings for each algorithm indeed represent a different distribution. That is, since the difference between the ranks is at most 2, a test of the outcome was necessary to determine statistically significant distinctions between the algorithms. Consequently, a Wilcoxon signed-rank test was performed as described in §2.4.2. This test computes statistical significance of similarity between paired rankings. Each possible pairing of the algorithms’ rankings were tested for a 95% significance level, and the result was a p-value of, at most, 6.5e-11.
Figure 3.15: Comparison of other work to our computed mosaic via a sampled portion.

Hence, each algorithm's rankings were statistically significantly different from those of the other two, making the average scores comparison meaningful.

A pairwise comparison of rankings of the algorithms indicated that the proposed algorithm outperformed Graph-Cuts in 59.4% of the rankings and SURF+RANSAC in 72.5% of the rankings. Furthermore it outperformed both of the other algorithms in 48.5% of the rankings compared to 31.2% for Graph-Cuts and 14.3% for SURF+RANSAC. Only 10.5% of the time was the proposed algorithm ranked the worst of the three compared to 26.8% for Graph-Cuts and 51.2% for SURF+RANSAC. The average rank of our algorithm was 1.6, whereas the average rank of Graph-Cuts was 2.0 and SURF+RANSAC was 2.5. The variance in each rank was approximately 0.5. Since the Wilcoxon signed-rank test indicated a significant difference, these averages do indicate a higher subjective quality in the mosaics produced by the proposed algorithm in comparison to the other two tested algorithms.
3.7 Summary

We have proposed a novel utilization of sparse-decomposition techniques within a rigorous set-theoretical framework for image mosaicking. While the results cannot be quantitatively compared to other work due to the lack of a standard, the results were compared visually. Accordingly, we proposed a polling technique to exploit this subjective evaluation. The results confirmed the better performance of the proposed algorithm to other comparable existing techniques. We additionally note that the method does not rely on heuristically defined criteria. Furthermore, the local graphical structure utilized allows for parallelization of the computations for improved performance.
Chapter 4

Multi-Level Scene Understanding via Hierarchical Classification

Considering applications producing data of the type described in the previous chapter and, in general, applications where the use of video surveillance is necessary and/or beneficial, an understanding of the contents of the data. Information extraction, hence, is at least time-consuming, and possibly prohibitively expensive due to the size of the data. It is a reasonable goal to try to automatically identify the contents of the video. Of particular interest in such applications is the ability to recognize locations in the environment where events occur, and describe the events and/or elements of the scene common to those locations. This is one of the goals of scene understanding.

Scene understanding is traditionally addressed from one of two distinct points-of-view: the description of the underlying environment, or the action taking-place throughout the scene. Each of these facets is required to address the overarching goal but, is insufficient independently to address the problem entirely. These facets are, in fact, dependent and by considering both, a more complete description becomes possible. Here, we provide a novel, data-driven, scene understanding and classification technique that captures and utilizes information about both the environment, and activity within a scene.

It was our goal to describe large-scale scenes of several agents, performing different actions, by localizing regions of interest (ROI) where these actions were being performed, and produce a description of those activities. We then proceed to describe a framework
whereby one might use these descriptions in a supervised classification setting as training
data to compare/classify query/test scenes.

Our method addresses the problem of comparing scenes via both environment and
activity analysis, especially in examples where both are critical components for the appli-
cation. The layered-sensing setting provides a general class of problems wherein a goal is
to describe a situation at multiple scales simultaneously. For example, traffic monitoring,
crowd motion planning, air traffic control, etc..

We approach solving this problem by leveraging a particular scene decomposition
technique to describe the environment and the motion/action. We can then utilize the
information from each sub-problem as additional information to simplify the other. The
result is a two-fold description of the scene provided for quantized sub-regions. This
layered, or multi-level description, allows for a hierarchical classification of subsequent
query scenes for more thorough descriptions.

4.1 Methods

We start with a video sequence represented by \( X(x, y, t) \) where \((x, y) \in \mathcal{I} \) are the pixel
coordinates in the image plane \( \mathcal{I} = [1, w] \times [1, h] \subset \mathbb{N}^+ \times \mathbb{N}^+ \) with height and width
indicated by \( h \) and \( w \), respectively, and the frame index \( t \in [t_i, t_f] \subset \mathbb{N}^+ \). We partition it
along its three dimensions, two spatial and one temporal. The temporal partitioning is
achieved via a windowing in the time dimension to produce \( X(x, y, t) \), with \( t \) constrained
to the set \( T = [t_1, t_n] \subset [t_i, t_f] \).

The spatial partitioning remains constant throughout the time-window. We assume
the video sequence is produced via a stationary camera. We first segment the field-of-
view (FOV), the image plane, into regions that correspond to different activities within
the scene. Then we provide two descriptions: one for the activity in each region, and one
for the regions themselves.

We formulate the problem as one of matrix partitioning. To that end, the video-
cube \( X \) is transformed into a matrix. This is accomplished by vectorizing each frame
\( X_j = X(x, y, t_j), \forall t_j \in T \) in the usual way: concatenating the columns to produce a
where \( \text{"tr"} \) indicates the transpose operation. These vectors are then arranged into a matrix

\[
Y = \begin{bmatrix} y(1), y(2), \ldots, y(n-1), y(n) \end{bmatrix},
\]

(4.2)

to represent the data. We subsequently decompose this matrix using robust subspace recovery - bi-sparsity pursuit (RoSuRe-BSP) [12] as described in §2.5. The decomposition takes the form

\[
Y = BW + F,
\]

(4.3)

where the columns of \( B(s, t) \) are the low-rank basis vectors for the data, the background of the time-windowed video sequence, \( W \) is a coefficient matrix, and the columns of \( F(s, t) \) correspond to the sparse foreground components for each frame, the human silhouettes. Here, \( s \in S \) is the index of the pixels in each frame. Making use of the correspondence between values of \( s \) and pairs \((x, y)\), we construct the functions \( B(x, y, t) \) and \( F(x, y, t) \), both in the range \([0, 255]\), that are the background (low-rank) and foreground (sparse) components over the image plane \( I \).

### 4.1.1 The partitioning of the FOV

To ensure that the partitioning of the FOV corresponds to different activities, we consider all the motion contained in \( X \) to segment the regions. Since the decomposition has already separated the movement \((F)\) from the stationary \((B)\), we can perform our computations on just the moving part.

Define a summary function, \( P(x, y) \), of all motion in \( X \) thus:

\[
P(x, y) = \sum_{t \in T} F(x, y, t).
\]

(4.4)

This image represents the sum of all foreground components throughout the time-windowed video sequence. It is possible to localize regions of activity by decomposing \( P(x, y) \) into partitions \( U_l \) where \( l \in \Lambda \) is an index set for the partitions, i.e. regions of interest (ROIs).
To help simplify in the segmentation of different regions of activity, we follow a procedure similar to [34], and define a threshold set

$$Q = \{q \in \mathcal{I}|P(q) > \tau\},$$

(4.5)

and an associated indicator function:

$$\mathcal{I}(q) = \begin{cases} 0, & q \in Q, \\ \infty, & \text{otherwise}. \end{cases}$$

(4.6)

The value for this threshold $\tau$, in this effort, was chosen experimentally. This threshold allows the tuning of the amount of activity required to determine a ROI. Such additional formalism facilitates the computation of the traditional distance transform of the image $P(x, y), \forall p \in \mathcal{I}$ thus:

$$D(p) = \min_{q \in \mathcal{I}} (d(p, q) + \mathcal{I}(q)),$$

(4.7)

where $d(p, q)$ is the $l_1$ distance in the image plane. Local maxima of $D(p)$ provide the boundaries of a partitioning of the set into subsets denoted $U_l$, our ROIs. These subsets can be projected onto the summary function $P(x, y)$, and thus the time-windowed set $X(x, y, t)$ as ROIs:

$$V_l = \{(x, y) \in \mathcal{I}|D(x, y) \in U_l\}, \forall t \in T.$$  

(4.8)

4.1.2 Description of the ROIs

To describe the environment of the scene observed in the time-windowed video sequence $X(x, y, t)$, we first separate the ROIs based on the decomposition described in §4.1.1. This is accomplished by considering the background of the video sequence over each region $l$:

$$B_l = \{B(x, y, t)|(x, y) \in V_l\}.$$  

(4.9)

A feature set, $f_l$, for each background subset $B_l$ is constructed following the SURF [9] algorithm instead of the method described in §4.1. The use of key-points, as in SURF, enable a succinct description of discrete and possible disparate objects in the
scene whereas our method in §4.1 describes the entire frame. In this application, the similarity sought between ROIs is not that of visibly identical regions but, rather similar scene constituents.

The SURF descriptors are invariant under affine transformations, and capture details about the constituent objects in a scene. A comparison between regions can be performed by finding and quantifying the number and strength of matches between respective descriptor feature sets.

The comparison is performed using the nearest neighbor ratio matching strategy [8]. We proceed by finding the nearest neighbors (c.f. k-nearest neighbors, k=2) to the features in the query set \( f_i \) from the training sets of descriptors. The distance (e.g. Euclidean distance), between these descriptors serves as a quantifier of the strength of the potential match. The region described by the candidate set of descriptors with the greatest number of matches to the test set, is considered the most similar region.

### 4.1.3 Region Activity Description

The action description herein is one of activity density. For each region \( V_i \), we define the activity density:

\[
    A_l(t) = \sum_{(x,y) \in V_i} F(x, y, t). \tag{4.10}
\]

This results in the time-series \( A_l(t) \) for each ROI, \( V_i \), which depicts the density of the activity in that region for the time-window \( T \).

To eliminate the necessity of aligning these time-series for comparison, we consider the Fourier representation of these series,

\[
    A_l(t) = \mathcal{F} (A_l(t)), \tag{4.11}
\]

normalized for unit power, as the signature for the activity density in ROI \( l \) for time-window \( T \). To discern/categorize these representations, the Fourier coefficients can be directly compared.
4.1.4 Training and Classification Framework

With these comparison rules in hand, it is possible to proceed according to a supervised classification setting.

A training set is produced by performing the prescribed analysis on a large database of labelled scenes. The result will be descriptor sets and activity density signatures for each region segmented in each scene. The scenes are summarized by their constituent regions.

By supplying a sufficient database of scenes with associated descriptor feature sets and activity density signatures, it is possible to classify any incoming test/query scene by decomposing it into regions, according to this technique, and subsequently comparing them to templates in the database. Once similar regions in the training set are identified, the query scene may be described as either a match to a training scene (all similar regions in one scene class), or a combination of regions from several training scenes. Due to the multiple levels of description and the fine granularity at the details, a comprehensive description for the query scene is generated on the basis of the provided training set.

4.2 Experimental Testing

The proposed framework was tested on many video sequences of a picnic scene, taking place in a courtyard, containing several different regions of activity including: regions where tents are being constructed, regions for playing games, regions containing picnic tables at which participants can eat, and areas wherein no activity took place. The video sequence was captured from an overhead view at 10fps and a resolution of 640 × 480 pixels. An example image is shown in Figure 4.1.

A time-window length of 1.25min was chosen for these experiments to highlight some instructive results. The frames captured in the first such window were used for the test set $X(x, y, t)$. 
4.2.1 Background/Foreground Separation

Applying RoSuRe-BSP to optimize per Equation (4.3), the foreground $F(x,y,t)$ and background $B(x,y,t)$ of the video sequence were separated. An example is shown in Figure 4.2. The result of the decomposition illustrates a complete separation of the foreground and background contents: the dynamics in the scene, e.g. people and their shadows, are extracted and even the “stationary” object, e.g. trees, that move slightly over throughout $X(x,y,t)$ are separated as background components. This accuracy in decomposition allows for the two-pronged approach of considering the motion elements and the background elements separately. However, such a breakdown also affords the possibility of leveraging the analysis of one component to address the other. Utilizing the dynamic elements to segment the background into ROIs is the next step.
4.2.2 The partitioning of the FOV

Proceeding with the partitioning, the summary function \( P(x, y) \) was computed from the foreground component \( F(x, y, t) \) as shown in Equation (4.4). The threshold for this data was set at \( \tau = 150 \) to produce the set \( Q \) described in Equation (4.5). It was this set that was utilized for the definition of the boundaries of the regions \( V_i \), following Equations (4.7) and (4.8), which are enumerated in Figure 4.3.

4.2.3 Description of the ROIs

Once the regions \( V_i \) were segmented via their defined boundaries, the associated descriptor sets were produced for each via the SURF-128 [9] algorithm. As described in §4.1.2, utilizing these features, it’s possible to quantitatively compare the regions by considering the number of matching descriptor pairs among all of the regions. An image representa-
tion of the resulting similarity matrix is shown in Figure 4.4. For clarity, the number of matches have been normalized by the largest number of matching feature pairs.

The region $V_7$ is quite similar to several regions due to the diversity of its contents: it contains part of a building and a large grassy area in addition to the sidewalk. These contents are common with regions $V_2$, $V_3$, and $V_4$. It is less similar to regions $V_5$, $V_6$, and $V_8$ since these regions do not contain any portion of the building surrounding the courtyard and additionally contain picnic tables. Regions $V_9$, $V_{10}$, and $V_{11}$ are not significantly similar to other regions, as could be inferred from inspection.

4.2.4 Region Activity Description

The last part of the algorithm includes the computation and comparison of the activity density signatures $A_l$. This was completed following the method described in §4.1.3. First, the activity densities $A_l$ were computed from $F(x, y, t)$ for each ROI $V_l$ as given by Equation 4.10. Examples of these time-series are shown in Figure 4.5.

These time-series accurately represent the intuitive understanding of the activity tak-
Figure 4.4: Similarity between regions indicated by the number of matches, normalized.

...ing place in the corresponding regions. At the beginning of the scene, region $V_2$ is populated with individuals moving throughout. Soon, within the first 10s (100 frames), the crowd exits $V_2$ and the region remains mostly unpopulated for the majority of the remainder of the sequence with the exception of a few people passing-through near frame 300.

Region $V_{10}$ is one of the more active regions in this data set in that several agents begin the scene by constructing a tent therein, and exit the region upon task completion. The activity density time-series indicated in Figure 4.5, again, accurately represents an intuitive understanding of the activity. $A_{10}$ increases from the beginning of the scene to about frame 180, when the actors are entering the region and starting construction. The magnitude of $A_{10}$ remains relatively high for many frames, corresponding to the ongoing construction of the tent. Once the construction nears completion, the activity density decreases as the agents exit $V_{10}$.

Since both facets of the scene understanding problem have been addressed, we can produce a similarity measure that combines both. The result captures the similarity between region descriptor sets but, is tempered by the disparity between activity density signatures, c.f. region $V_2$. 
Figure 4.5: Activity density time-series for regions $V_2$ and $V_{10}$, as labeled in Figure 4.3.

For comparison, we proceed to compute the activity density time signatures $A_l$ as indicated in Equation (4.11). A simple correlation comparison of these signatures results in the similarity matrix shown in Figure 4.6. While there are some consistent similarity results between those of the region descriptors and the time-series, there are also dissimilarities, e.g. region $V_1$ is similar in descriptor feature set to region $V_7$ but, not so in time-series.

### 4.3 Summary

We have described and demonstrated a novel data-driven scene understanding and classification technique. Current scene understanding approaches focus on either describing the environment captured within the scene, or the action taking place during the scene. Our technique addresses scene understanding as a single problem by leveraging the data available for each of these facets against the other, and then combining these intertwined results into a single result.
Figure 4.6: Similarity between activity density signatures, normalized.

This tool allows the practical use of the large-scale, high-resolution mosaics presented in Chapter 3 for automatic scene decomposition and understanding; another integral part in our system-focused approach to layered sensing.
Chapter 5

Activity Analysis and Material Discrimination via LWIR Polarization

5.1 Introduction

Usage of video capture devices in areas such as security, has afforded an additional temporal facet to traditional remote sensing. Separately, the advent of multi and hyperspectral imaging devices have further improved the diversity, and thus the utility of the resulting data for decision making. The two ideas have recently been merged via the conjunction of video data collection in multiple wavelengths with polarimetric sensors. The resulting data is video of the traditionally collected intensity information, as well as the polarized status of the incoming radiation. An example of a current application that could possibly benefit from this extension, is human activity analysis. Approaches to this problem can be characterized by the scale of operation. In this effort, we focus on the actions performed by humans, e.g. walking. This is a more coarse scale than that of atomic-actions, e.g. raising a foot, and finer than that of activities, e.g. walking to the car. In infrared video, thermal diversity may be limited to the extent that objects of interest may be indistinguishable from other elements of the scene. In such a situation, current human action classification techniques are unable to perform on account of the
lack of a correct segmentation. Information about the polarization of a scene provides sufficient diversity to perform such a detection/segmentation.

The first effort here described utilizes a physics-based model for emission polarization in the long-wave infrared band to segment the pixels in each frame into classes representing different materials in the scene. While this is a computationally demanding process, the segmentation is useful as an input to other algorithms for further processing. The additional polarization diversity is, however, sufficient to perform the action classification without the detection/segmentation process.

In the second effort, we distinguish between, and classify three actions: walking, walking while carrying a heavy load, and running. By exploiting higher order statistics of the polarization diversity as features we are able to classify the actions without segmenting the human actors from the remainder of the scene.

The data used for testing was a polarimetric infrared video dataset consisting of videos of 20 individuals performing two actions (walking and walking while carrying a heavy load) in three separate instances and an eight individuals running in three separate instances (a total of 148 sample actions). Each scene was comprised of between five and ten materials. The data was captured from instrumentation mounted on a simulated low-flying platform.

5.2 Related Work

Polarization has been of interest for many years [7,93]. It has been utilized for modelling [28,49,63], communications [62] and many RADAR-related applications [26,29,38,70,76]. In this work, we consider the polarization of a different wavelength signal and novel applications thereof: polarimetric long-wave infrared (LWIR) video.

While much of the current work in polarimetric LWIR video focuses on the development of the sensors and post-processing methods, some application efforts have surfaced in the last few years. The detection of manmade objects [27,78,84] is of great interest due to the strong polarizing nature of manmade surfaces. Some efforts have also been made in the 3D reconstruction [39,79] and discrimination of such objects [25,32]. Much work has been done toward automated target detection and tracking [4,6,13,58,66–68,102].
While some efforts have utilized LWIR for human detection [5] and human action classification [43], no such efforts have, however, been made using polarimetric LWIR video.

An entire field of study, ellipsometry, centers around observing the polarization of reflections and emissions of electromagnetic waves from target materials. For reflections, the subject materials are illuminated whereas for emissions they are heated. Both the intensity and polarization states of the reflected/emitted waves are observed such that the response of the material is sampled in a hemispherical pattern centered on the subject material. By studying these observations, one can discern properties of the materials, e.g. the index of refraction. Numerous physics-based parametric models [24, 30, 35, 45] for electromagnetic wave emissions have been proposed and applied [85] to match the data collected through such studies. In the first work described here, we propose a means to distinguish between materials in a scene of the micro-polarimeter video. This is accomplished by fitting one of the more recent [39] models to the observed data via the parameters. Thus, the parameters of the material that produced the data were attained.

Segmentation of images in modalities operating in parts of the spectrum outside the visible range, and the subsequent study of the corresponding response of materials in those ranges, are not new (e.g. hyperspectral remote sensing). Existing techniques, however, focus on the problem strictly as an image processing problem. We propose a novel contribution of combining the physics-based emission polarization model, and the collected video data for similar applications. The difficulty of this task lies in the unknown number of materials present in the scene. Not only is the number of materials unknown, so are their parameters. This difficulty is addressed via statistical analysis techniques in the space of parameters. With the assumption that the materials are unchanged in a single scene, it is possible to utilize all the data collected over the entire length of a single video to characterize the constituents. Once this parameter estimation is completed for the different materials in the scene, the frames are segmented by material.

The second contribution of this effort focuses on the phenomenon of human activities. We are able to distinguish between actions in different video sequences, each showing an individual performing one of several possible activities, and subsequently classify the actions. Statistical calculations for activity recognition solely based on the data have been difficult on account of the variation in texture, illumination, occlusions, etc.. While we do
not address occlusion in this work, we do consider texture and illumination. As in Section 5.4.2, data in the LWIR spectrum is generally better due to the lack of shadows and slow changes in intensity. This is equivalent to circumventing the illumination problem with traditional visible spectrum observation. The issue of textural diversity is, however, still present. The polarization information of the scene becomes advantageous for this obstacle. In the data collection utilized for this work, the background is a gravel lot that is stationary with respect to the observer.

The significance of this analysis also lies in exploiting a novel sensing modality: long-wave infrared (LWIR) polarimetric video. The novelty of the LWIR micro-polarimeter array is its real-time video capture of the polarization diversity of the electromagnetic waves observed in the thermal sub-band. In the past, such polarization data has only been available for still scenes captured at separate instances in time: a polarization filter was placed between the scene and the sensor, an image was collected and the filter was rotated in preparation for the next image collection. The added sequential nature afforded by the advent of video to this modality allows time-series analysis techniques to be applied. This approach proved viable for leveraging the discriminatory properties of polarization diversity for the problem, with respect to material composition and surface structure, in contrast to the traditionally complex preprocessing techniques common to video analysis.

Human action analysis, as a specific problem in remote sensing and an application of computer vision, is an important focus of much current research [10, 60, 83]. While no standard definitions have been formulated, a common taxonomy utilized to distinguish between varying scopes of study is emerging.

1. An action is a simple, low-level motion such as “raising an arm” or “walking.”

2. Activities are combinations of actions, e.g. “waiting for a bus.”

3. An instant where multiple individuals perform a variety of activities is called an event, e.g. a football game.

4. A behavior is a pattern in an individual’s or group’s activities.
Many techniques and approaches have been implemented to tackle the problem of detecting, recognizing, identifying and then exploiting human action descriptions. A recent review [2] separates the approaches into two main categories: “single-layer” and “hierarchical”. The latter category contains techniques that describe human activities by examining layered deconstructions of those activities. From those deconstructions, the classification for the activity is determined. Efforts in this category can be more specifically described by the approaches they use to describe and combine the deconstructed pieces: statistics/Hidden Markov Models (HMM’s) [72], syntactic grammar organizations [50,98] or atomic-actions/subevents [3,42].

In contrast, the single-layer techniques directly utilize images or video to conclude the activities’ classifications. In [2], the single-layer approaches are further divided into space-time and sequential approaches. This segregation is performed via the method with which the efforts manage the temporal dimension of the data. Space-time approaches treat the temporal dimension as a third spatial dimension in that a video of an action is analyzed as a volume [14,52,80,89] or as local features [71,87,105] or trajectories within that volume [20,77,90]. Sequential methods abstract the input videos as sequences of measurements from which feature vectors are extracted. These features are used to classify action sequences by either updating a state model that is compared to a trained state sequence for each action [64,73,103] or by comparing them to templates of expected action sequences [31,41,97].

Due to the discrete nature of the action categories, a signature-based approach is explored to distinguish the set of all action primitives. Since the goal of this effort is to classify actions, a single-layer space-time approach is chosen over the added complexity of the hierarchical methods. Our approach, in addition to being a single-layer approach, intuitively appealing in using statistical analysis afforded by the data-type. Many current techniques require considerable preprocessing to provide suitable data on which they may operate. These preprocessing stages consist of a multitude of tasks ranging from user input such as the initial location of a subject’s extremities which in turn, depend on “perfect” segmentation of the subject in each frame of an image sequence.
5.3 Background

5.3.1 Polarization of EM Radiation

Electromagnetic Wave Equation

In general, an electromagnetic (EM) wave traveling in a direction ‘z’ is represented by the superposition of two plane-waves also traveling in the z-direction as described by the following equation and shown in Figure 5.1,

\[ \vec{E}(z, t) = \vec{E}_X(z, t) + \vec{E}_Y(z, t). \] (5.1)

Following the derivation in [46], we can rewrite the constituent waves \( \vec{E}_X(z, t) \) and \( \vec{E}_Y(z, t) \) in terms of amplitude \( E_{0x} \), frequency \( \omega \) and phase difference \( \delta \), as in Eqs. (5.2) and (5.3), whose ratio yields an equation of an ellipse, as shown in Eq. (5.4), the equation of an ellipse:

\[ \vec{E}_X(z, t) = i E_{0x} \cos(kz - \omega t), \] (5.2)

\[ \vec{E}_Y(z, t) = j E_{0y} \cos(kz - \omega t + \delta), \] (5.3)

\[ \left( \frac{E_Y}{E_{0y}} \right)^2 + \left( \frac{E_X}{E_{0x}} \right)^2 - 2 \left( \frac{E_X}{E_{0x}} \right) \left( \frac{E_Y}{E_{0y}} \right) \cos(\delta) = \sin^2(\delta). \] (5.4)

As indicated by the above equations, a propagating electromagnetic wave will have some rotation about the z-axis, and will thus trace out an ellipse (Fig.5.2) in the plane of rotation, as shown in Fig.5.1. For different values of \( E_{0x}, E_{0y}, \text{ and } \delta \) (the phase difference between the two orthogonal waves), the ellipse will have different shapes. There are certain states, called degenerate states, that are produced for a certain parameter value of interest, which act as a basis by which any generally polarized waveform may be completely described (Fig.(5.3)). That is precisely what the Stokes’ vector representation exploits.

Stokes’ Vector Relations

Stokes [92] proposed a convenient means to describe incoherent polarized light in the form of a vector (or a set of parameters), \( S \). This idea was particularly novel for interpreting
Figure 5.1: The E-field component of an EM wave rotates about the transverse axis and thus traces an ellipse in the plane of rotation. This wave, $E$, can be decomposed into orthogonal components $E_X$ and $E_Y$.

Figure 5.2: The polarization ellipse with the labelled parameters (From [22]). $E_{0x}$ and $E_{0y}$ are the amplitudes of the corresponding orthogonal constituent waves, $\xi$ and $\eta$ are the major and minor axes of the ellipse, $\psi$ is the angle of polarization and $\chi$ is the angle of ellipticity.
Figure 5.3: The degenerate states of the polarization ellipse with conditions (From [22]). $E_{0x}$ and $E_{0y}$ are the amplitudes of the corresponding orthogonal constituent waves, and $\delta$ is the phase difference between those waves. From top-left, proceeding clockwise: linear horizontal polarized (LHP), linear vertical polarized (LVP), linear $-45^\circ$ polarized (L-45P), left circular polarized (LCP), right circular polarized (RCP), linear $+45^\circ$ polarized (L+45P).
each parameter as a measure of the “preference” of the observed light in a certain direction of polarization. The vector comprises four entities:

- $S_0$ - intensity,
- $S_1$ - horizontal preference,
- $S_2$ - $+45^\circ$ preference,
- $S_3$ - right circular preference.

In this formalism the intensity parameter $S_0$ is, for example in imaging, the gray-scale picture one would normally capture with a camera. The preference images (parameters, elements, etc.) are relations between pairs of degenerate states representing the observed light, e.g. the horizontal preference ($S_1$) is the difference in intensities of the horizontal (LHP) and vertical (LVP) degenerate states which characterize the observed light.

Another advantage of the Stokes’ vector formalism is the ease of calculation from readily measurable phenomena. By observing light through a linear polarizing filter at various orientations, one can observe the representative degenerate states, as shown in Fig. (5.3), from which the Stokes’ vector may be determined. These images are denoted by $I_\phi$, where $\phi$ is the angle corresponding to the orientation of the linear polarizer’s transmission, measured clockwise from the horizontal. Eqn. (5.5) relates the Stokes’ vector representation ($S$) to the collected image representation ($I_\phi$) and also to the orthogonal linear ($E_X, E_Y$) basis where $\delta$ is the phase difference invoked in Fig.(5.3)):

$$
S = \begin{bmatrix}
S_0 \\
S_1 \\
S_2 \\
S_3
\end{bmatrix} = \begin{bmatrix}
E_{0x}^2 + E_{0y}^2 \\
E_{0x}^2 - E_{0y}^2 \\
2E_{0x}E_{0y}\cos(\delta) \\
2E_{0x}E_{0y}\sin(\delta)
\end{bmatrix} = \begin{bmatrix}
\frac{1}{2}(I_0 + I_{45} + I_{90} + I_{135}) \\
I_0 - I_{90} \\
I_{45} - I_{135} \\
I_R - I_L
\end{bmatrix}.
$$

(5.5)

The above equations only hold true for completely polarized EM waves. Naturally occurring EM waves are, however, not completely polarized, but are partially polarized. This is tantamount to saying that they can be represented as the summation ($S'$) of
polarized \((S_p)\) and unpolarized \((S_u)\) components weighted by a value \(\mathcal{P} \in [0, 1]\), the degree of polarization (DoP).

\[
S' = (1 - \mathcal{P})S_u + \mathcal{P}S_p = (1 - \mathcal{P}) \begin{bmatrix} S_{u0} \\ 0 \\ 0 \\ 0 \end{bmatrix} + \mathcal{P} \begin{bmatrix} S_{p0} \\ S_{p1} \\ S_{p2} \\ S_{p3} \end{bmatrix}.
\]

Classically, two variables/features have been used to characterize and to explore the polarization phenomenon: \(\psi\) and \(\mathcal{P}\). On one hand, the angle of polarization, \(\psi\) (Eqn. (5.7)), is directly observed on the polarization ellipse, on the other, the degree of polarization (DoP, \(\mathcal{P}\)) is a derived feature. In conditions lacking circular polarization, which is most often the case in passive sensing settings [88], the DoP is well approximated by the degree of linear polarization (DoLP), as shown in Eq. (5.8) also denoted by \(\mathcal{P}\).

\[
AoP = \psi = \frac{1}{2} \tan^{-1} \left( \frac{S_2}{S_1} \right), \quad (5.7)
\]

\[
DoLP = \mathcal{P} = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}. \quad (5.8)
\]

To simplify the development of our algorithm, we employ a model of the above described physical phenomenon.

### 5.3.2 Physics-based Model of Emission Polarization

Observed radiance can be represented as a Stokes vector \(\mathbb{L}\) and thus related as a function of the irradiance of the object, Stokes vector \(\mathbb{E}\), and the \(4 \times 4\) Meuller matrix \(f\), through which the irradiance is transformed:

\[
\mathbb{L} = f \mathbb{E}. \quad (5.9)
\]

This \(f\) matrix is a general representation of any optical element but, in our setting, represents how the material of an object transforms the irradiance before it is observed as radiance. Such a matrix can be modeled via a bi-directional reflection distribution
function (BRDF), a statistical means of describing the material of the object and its response at different wavelengths ($\lambda$), producing $f_{pBRDF}(\lambda, \theta, \phi, n, k)$. Here, the $\theta$ and $\phi$ are the angles representing the three-dimensional orientation of the observation with respect to the normal of the object’s surface.

In [35], a polarimetric BRDF (pBRDF) model, based on the micro-facet model of Torrance and Sparrow [94], for LWIR emissions is presented. It has since been generalized in [40] to a combination of probabilistic distributions with tunable parameters.

Since the natural occurrence of circular polarization is commonly known to be negligible [88], it is only necessary to consider the unpolarized and linearly polarized components of the Stokes vectors. Additionally, in the setting of emission polarization, the irradiance from within an object is typically considered unpolarized and can be modeled as a black body irradiator. These simplifications lead to the following form:

$$L = fE = \begin{bmatrix} S' \end{bmatrix} = \begin{bmatrix} f_{00} & f_{01} & f_{02} \\ f_{10} & f_{11} & f_{12} \\ f_{20} & f_{21} & f_{22} \end{bmatrix} \begin{bmatrix} S_0 \\ 0 \\ 0 \end{bmatrix}, \quad (5.10)$$

where $S_0$ is the irradiance of the object (e.g. a function of $\lambda$ and temperature for black body models) and $S'$ is the observed radiance. From the pBRDF models, the components of $f$ can be manipulated to produce a model for $P$, denoted $\hat{P}$, mapping the parameters of the component distributions to an estimate of $P$. It is this model estimate, $\hat{P}(\theta_v(\theta, \phi), n, k)$, that we utilize (Eqn. 5.11), as derived in [53].

$$\hat{P} = \frac{2Asin^2(\theta_v)cos(\theta_v)}{A^2cos^2(\theta_v) + sin^4(\theta_v) + B^2cos^2(\theta_v)}, \quad (5.11)$$

where the functions $A$ and $B$ are defined in terms of the function $\theta_v(\theta, \phi)$ as follows:

$$A(n, k, \theta_v) = \sqrt{\frac{\sqrt{C} + D}{2}}, \quad (5.12)$$

$$B(n, k, \theta_v) = \sqrt{\frac{\sqrt{C} - D}{2}}, \quad (5.13)$$
\[ C(n, k, \theta_v) = 4n^2k^2 + D^2, \quad \text{and} \]
\[ D(n, k, \theta_v) = n^2 - k^2 - \sin^2(\theta_v). \]

From this representation of \( \hat{P} \), given a sufficient sampling of \( P \), we can compute candidate values for \( n \) and \( k \) to distinguish between materials in a scene.

### 5.3.3 Statistical Analysis Tools

Here we briefly cover some rudimentary statistics utilized in a portion of this work.

Consider a vector \( x \in \mathbb{R}^n \) of \( n \) samples \( x_i, i \in [1, n] \), from a continuous random variable \( x \in \mathbb{R} \). One alternative approach to describing a probability distribution function \( f(x) \) of \( x \) is via its statistical moments. These moments quantitatively describe the shape of a distribution so as to facilitate its classification and identification. The \( k^{th} \) moment of the function \( f(x) \) centered about the value \( c \) is described by Eq. (5.16),

\[ \mu_k = \int_{-\infty}^{\infty} (x - c)^k f(x) dx. \tag{5.16} \]

The mean, or expected value \( \mu = \mathrm{E}[x] \), of a random variable can be described as the first moment of the distribution centered about zero. Moments centered about the mean of the variable are called “central moments”. This makes the central moments dependent only on the shape and scale of the distribution. Of particular interest in this effort are standardized, central moments that have been normalized by the \( k^{th} \) power of the variance \( \sigma \). Here, again, a level of invariance is added to the moments, in that standardized moments are strictly influenced by the shape of the distribution, and not on the location or scale. Commonly, in conjunction with the mean and the variance, the third and fourth standardized moments (the skewness and kurtosis, respectively) are also considered to describe a distribution (shown as calculated),

\[ \text{Skew}(x) = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^3}{\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2\right)^{3/2}}, \tag{5.17} \]
\[ \text{Kurt}(x) = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^4}{\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2\right)^2}. \] (5.18)

While these two descriptors appear similar, the difference in the exponent is key to the difference in describing the shape of the distribution. The skewness of a distribution is a measure of the asymmetry of that distribution and the kurtosis of a distribution describes its peakedness. So, for the random vector \( \mathbf{x}_i \), these moments describe the shape of the distribution of the random variable \( x \).

The combination of these higher order statistics with the location and spread information contained in the mean and variance, respectively, provide a good description of a distribution, and allow for a significant comparison between distributions.

### 5.4 Data Description

#### 5.4.1 Data Collection and Characteristics

The instrument by which the data used in this work was captured, produces video sequences of the Stokes [92] vector of a scene. The video is in three channels, much like the three color channels of a CCD camera:

- \( S_0 \) - intensity,
- \( PS_1 \) - horizontal preference,
- \( PS_2 \) - \(+45^\circ\) preference.

The data utilized in this work is the result of a post-processing method applied to an output from a cooled micro-polarimeter focal plane array (FPA). Each cluster of four pixels on the FPA forms a “super-pixel,” as depicted in Fig.(5.4). The patterns indicate the orientation of a linear polarizing filter placed over each sensor in a pattern reminiscent of the Bayer filter pattern used for color CCD cameras.
Figure 5.4: Superpixel polarizing filter pattern of micro-polarimeter focal-plane array. Starting at the top left and proceeding clockwise, the orientations are $0^\circ$, $+45^\circ$, $90^\circ$ and $-45^\circ$.

The raw output from these sensors is a four-channel (one for each $I_\phi$ from Section 5.3.1) image sequence. For each image in the sequence, these channels are combined and processed, as described in [56], to produce a video sequence with three channels, one for each Stokes component, and hence the data analyzed in this work. Each frame is $471 \times 641$ pixels in size, and the video is captured at twenty-four frames per second (24 fps), resulting in each sequence spanning approximately 1500 frames. The target wavelength of the sensor is the LWIR spectrum (8-14 μm), i.e. the thermal spectrum. Subjects in the dataset performed three activities along a predefined course: walking, walking while carrying a heavy bag, and running. The resulting scenes consist of a gravel lot across which each action is performed, progressing from the right to the left extents of the field of view (fov) of the sensor.

5.4.2 Infrared Band Implications

In Sect.5.3.1 our discussion did not focus on any specific band of the EM spectrum, though some intuition from visible light was called upon in the given examples. The formalism is indeed applicable to all frequency ranges of the EM spectrum. In remote sensing, diversity of modalities is a necessity, and here we focus on an imaging type that caters to such a
Table 5.1: Ranges of the defined sub-bands within the IR spectrum band.

<table>
<thead>
<tr>
<th>Sub-bands of the Infrared (IR) Spectrum</th>
<th>Wavelength ($\lambda$) Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Near infrared (NIR)</td>
<td>0.7 – 1µm</td>
</tr>
<tr>
<td>Short-wave infrared (SWIR)</td>
<td>1 – 3µm</td>
</tr>
<tr>
<td>Mid-wave infrared (MWIR)</td>
<td>3 – 5µm</td>
</tr>
<tr>
<td>Long-wave infrared (LWIR)</td>
<td>8 – 14µm</td>
</tr>
</tbody>
</table>

requirement. Imaging in the infrared band is useful for observation in low-light conditions since IR sensors, in spite of their relatively small band and collection of relatively few photons, are very accurate. With proper scaling, they afford object discernment under a greater range of operating conditions than does visible light imaging [81]. As described in Table 5.1, there are several sub-bands defined within the infrared band ($\lambda \in [0.7, 14]µm$).

These different bands are defined by the bandwidth that different sensor types cover, e.g. NIR is from the end of perception by the human eye through the response of silicon based detectors [69]. Our work focuses on the LWIR band. As can be seen in Fig.5.5, reflections dominate all the IR bands, as in the visible light spectrum, except for the LWIR band. In the LWIR band emitted radiation is dominant, i.e. thermal emissions. The advantage of imaging in this band is that heat/radiation sources are themselves observed, and small sources are not so obstructed by the reflections/shadows from larger/stronger sources, e.g. the sun, as they would be in other bands. While this may cause some problems, e.g. there are multiple sources instead of one strong one (e.g. the sun in the visible sub-band), it simplifies many calculations since it is not necessary to take into account reflections and shadows in the presence of moving objects in a scene. There is not a total absence of shadows and reflections, and these can cause ambiguities in thermal equilibrium conditions, but because heat transfer is a relatively slow process compared with visible light transmittance, and with video capture rates, they are negligible over short time intervals.
5.5 Experiments and Results

5.5.1 Material Segmentation

It is well known that the DoLP of a given object of a given material can be represented as a function of the view angle with respect to the normal of the surface [100]. Assuming a sensor that is stationary with respect to the background of the scene, the motion of the objects in the scene will induce changes in the observed DoLP peculiar to those objects’ structures and materials. We can represent this relationship accurately as follows:

$$\frac{\partial P}{\partial t} = \frac{\partial P}{\partial \theta} \frac{\partial \theta}{\partial t}.$$  \hspace{1cm} (5.19)

This relation captures the well-established correspondence between the DoLP of a material and the view angle. It also includes the dependence of the change in view angle on the motion observed in the video by describing the view angle as a function of time, \(\theta(t)\). Abstracting a video sequence as a space-time volume, and considering the values of all its voxels simultaneously allow one to derive a relative distinction between moving
Figure 5.6: $\mathcal{P}(\theta)$ curves for Uranium glass (shown in red) and Tungsten (shown in blue) objects in the scene from their material properties. This is due to the video capture resulting in, effectively, a sampling of multiple $\mathcal{P}(\theta)$ curves, one for each material, at different view angles, $\theta$. Figure 5.8 is a plot of each pixel measurement in such a space-time volume as observed in the data. While the peak at zero for $\frac{\partial \mathcal{P}}{\partial \theta}$ is dominant due to the small size of movers in the scene, the other lobes correspond to different materials in the scene, indicating that distinction is viable. To demonstrate this principle, we exploit the previously-described physics-based model, Eqn 5.11 from Section 5.3.2, to simpler but real measurements.

The data utilized for this part of our experiments are samples from two measured $\mathcal{P}(\theta)$ curves, one for Uranium glass and one for Tungsten (Fig. 5.6). A separate monotonic sampling in $\theta$ was devised for each material to mimic a smooth rotation in view. Gaussian noise was added to those sampled points to produce data (Fig. 5.7) similar to that observed from the collected video (Fig. 5.8).
Figure 5.7: Scatter plot of data resulting from sampling $\mathcal{P}(\theta)$ curves for Uranium glass and Tungsten and the addition of Gaussian noise.
Parameter Estimation

For each voxel $x_i$ in the data, we fit the model given in Eqn. (5.11) (sweeping through $\theta$), to the points in its surrounding neighborhood $\mathcal{N}(x_i) = \{x_j | d(x_i, x_j) < \epsilon\}$ where $d(x_i, x_j) = \|x_i - x_j\|_2$ is the Euclidean distance between the points $x_i$ and $x_j$ in the space $\mathcal{P} \times \frac{\partial \mathcal{P}}{\partial t}$. The Nelder-Mead simplex direct search method for solving nonlinear optimization problems was employed for this fitting [55]. The value for $\epsilon$ was chosen so that in the most dense regions, few points populated each neighborhood, so as to improve the performance (speed) of the optimization. This procedure was iterated over many neighborhoods until all the voxels had been considered. The result was a set of points, one for each neighborhood, in the $n \times k$ parameter space, as shown in Fig. 5.9.

A hierarchical clustering was then performed in the “$n \times k$” space to determine the number of clusters and their centers. The corresponding parameter values for these centers were used as the representative parameters for each voxel within a cluster. In this fashion, each voxel was categorized via material composition. Figure 5.10 depicts the
Figure 5.9: Points in the $n \times k$ parameter space with two clusters evident.
Figure 5.10: Material curves resulting from using cluster centers as parameter values. Data points colored to indicate true material-type (Uranium glass as red and Tungsten as blue).

result of such a process. The dashed lines indicate the material curves from the model $\hat{P}$ when the cluster centers were utilized as parameter values. The data points are colored to indicate the material from which they were sampled.

**Turbo Estimation Error Correction**

Clearly, from Fig. 5.10, if one were to classify the data points via which model material curve they were nearest, some errors would be incurred. However, some information present in the data, omitted from the discussion in Section 5.5.1 but, included in Eqn. 5.19 is yet to be utilized: the time-ordering of the samples, i.e. $\theta$’s dependence on time.

Each voxel represented in the scatter plots has a corresponding time at which it was observed. Drawing inspiration from the Turbo principle, a philosophy used in turbo
coding [11] and back-propagation [17], we utilize information available that hasn’t been employed, to improve the estimate (classification result) already computed. In this case, for points classified as the same material, a simple evaluation of the smoothness in the corresponding $\theta$ trajectories is sufficient to improve the classification result.

For each set of points $C^m = \{c^m_i\}$ in the data classified as the same material $\omega_m$, we consider the time at which each point was sampled, $t_i$. Using the model curve produced from the earlier classification and the corresponding parameters ($n$, $k$ and measured $\mathcal{P}$), we perform the inverse calculation of $\theta$ corresponding to each point. Observing the trajectory of this ordered sweep through $\theta$, a smoothness criteria was implemented to indicate errors in the classification:

$$\left| \frac{\partial \theta}{\partial t} \right| < \delta$$

(5.20)

The current method for choosing the parameter $\delta$ is empirically based on the observed data. However, more sophisticated methods are being developed based on models of the actions to be observed in practice.

For the values of $\frac{\partial \theta}{\partial t}$ that do not meet this criteria, they are considered errors in classification, and are compared to the $\theta$-trajectories of the other material classes for proper alternative classification.

### 5.5.2 Action Classification

For the experiments described below, the DoLP portion of the collected data was utilized. This data was produced following the computation given in Eqn. (5.8).

We represent each image sequence as a matrix $X = \{x^j_i\}$ where $i \in [1, 471 \times 641]$ is the pixel index for each frame, and $j \in [1, 250]$ is the frame index. For each time instant $j$, we calculate the mean, variance, skewness and kurtosis of the sample vector $x^j_i$ to produce a time series of each statistic. The result is a collection of four corresponding time series for each scenario performed by each subject.
Figure 5.11: Resulting $\theta$ trajectory from similarly classified data points. Arrows indicate errors in the classification indicated by marked deviations from the otherwise “smooth” trajectory.
Simulations

To validate the idea that the statistics of the frames would follow some periodicity induced by the subject action in the video, the experiment was performed with simulated data. This surrogate data consisted of a video of the same frame size and rate as the collected data, and also contained sequences of single individuals performing simple actions. The notable difference between the simulated data and the collected data is the binary nature of the simulated data. Whereas the collected data pixel values are measurements from the scene, the pixel values of the simulated data are either one (1) or zero (0), corresponding to pixels that belong to the foreground and background of the scene, respectively.
Figure 5.13: Correspondence between frames of the simulated walking video and the kurtosis value in each frame.
Figure 5.14: Correspondence between frames of the simulated running video and the kurtosis value in each frame.
The computed time-series of the statistics, shown in Fig. 5.12, indicates that the statistical moments of each frame vary periodically, just as the action in the videos are periodic in nature. To further explore this correlation, a qualitative frame-by-frame matching was performed (Fig.s 5.13 & 5.14), and it was subsequently determined that the periodicity of the statistics was induced by the motion in the video, as was expected from a binary video sequence.

In the binary setting, the statistics are determined by the variation in the number of foreground/background pixels. The frequencies of these values in a single frame are the direct results of the pose of the subject with respect to the viewing orientation of the camera. The performed action performed dictates the statistics of the frame.

**Data Exploitation**

The same computations as those performed with the simulated data were performed with the collected data. Figure 5.15 is representative of the statistics calculated for an action sequence from the collected dataset. The noteworthy difference between the statistics of the collected data and the simulated data is that the mean and variance of the frames of the collected data do not only not follow the periodicity of the actions, but the magnitude and variations are so slight as they can be attributed to noise. Also evident from Fig. 5.15 is the close correlation between the kurtosis and the skewness time-series of the representative action. Due to this extended similarity and the added variation/magnitude afforded, the kurtosis time-series was chosen as a characteristic measure of the data. Using the kurtosis time-series as a description of the action for each trial (one individual, one action), the classification was performed in a supervised learning fashion. Figure 5.16 shows the kurtosis time-series for the three different actions as performed by a single individual, indicating a different periodicity for each action.

The spectrum of each time-series was computed, and used as the feature vector for classifying the actions. To focus on the utility of these features in classifying the actions, a standard classifier was used: K Nearest Neighbors (KNN). Due to the relatively few trials available in the dataset, the leave-one-out (LOO) method was used for training/testing. The classifier results are displayed as a confusion matrix in Table 5.2. The errors of classifying the “carry” action as “running” is due to the use of the spectra of the time-series.
Figure 5.15: Representative higher order statistical moments for each frame of a section of the collected data.
Figure 5.16: Kurtosis time-series sections for all actions as performed by the same individual.

series as a feature for classification. The additional motion of the load created higher frequency components in the spectra, thus increasing the similarity between the spectra for “carry” and “running.”

5.6 Summary

By utilizing the LWIR polarimetric video, we were able to capitalize on physical phenomena not previously exploited for material distinction and human action classification. While this is a first step in a new direction for applications of this modality, the results appear promising. Further effort must be applied to considering the performance of these techniques as compared with currently employed methods. For instance, a comparison of the material distinction power of polarimetric LWIR (PolarLWIR) imagery versus hyperspectral imagery (HSI) should be performed. It is expected that, while the HSI should prove more accurate in material identification, the PolarLWIR result could be sufficient in many settings and with a decreased sensor/processing cost. Also the simplicity of the
Table 5.2: Confusion matrix representing the accuracy of the classification performed (in %) via the spectra of the kurtosis time-series characteristics.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Walk</td>
</tr>
<tr>
<td>Walk</td>
<td>97</td>
</tr>
<tr>
<td>Carry</td>
<td>0</td>
</tr>
<tr>
<td>Run</td>
<td>0</td>
</tr>
</tbody>
</table>

computations required for the action classification to produce results indicated in Table 5.2, merits further study of similar applications.
In this thesis, we proposed a top-down parallelizable system to combine imagery from many different sensors and produce a rigorously quantifiable description of human activity and its relation to the scene at several scales. This system consisted of three main components: a mosaicking algorithm utilizing a sparse image decomposition for fast parallel tile matching and a testing method thereof, a scene segmentation algorithm that delineates large scenes and produces descriptions of possible regions of interest and the activities occurring therein based on observed crowd motion, and a subsystem for simultaneously describing the motion of an object and its material composition.

Each of these components were mathematically developed from first principles and were, additionally, tested on real collected data. Mosaics with gigapixel resolutions were produced from large camera arrays. A polling procedure was proposed and implemented to evaluate the mosaics’ apparent quality. Video of a picnic scene was de-constructed into relevant regions of interest and descriptions were produced for both the regions and the activities observed in them. Finally, the actions of the individuals were represented by statistical signatures and utilized to leverage a new sensor for material discernment.

Future related work may include fusion of other sensing modalities at the various system components. Additionally, the system should be implemented on a distributed computing platform to realize its advantages.
BIBLIOGRAPHY


[33] Steven Feller. Aware2 multiscale gigapixel camera.


