Summer 8-2015

A Visual Analysis of Articulated Motion Complexity Based on Optical Flow and Spatial-Temporal Features

Beau Michael Christ
University of Nebraska-Lincoln, beauredwolf@icloud.com

A VISUAL ANALYSIS OF ARTICULATED MOTION COMPLEXITY BASED ON OPTICAL FLOW AND SPATIAL-TEMPORAL FEATURES

by

Beau Michael Christ

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Doctor of Philosophy

Major: Computer Science

Under the Supervision of Professor Ashok Samal

Lincoln, Nebraska

August, 2015

A VISUAL ANALYSIS OF ARTICULATED MOTION COMPLEXITY BASED ON OPTICAL FLOW AND SPATIAL-TEMPORAL FEATURES

Beau Michael Christ, Ph.D.

University of Nebraska, 2015

Adviser: Ashok Samal

The understanding of motion is an important problem in computer vision with applications including crowd-flow analysis, video surveillance, and estimating three-dimensional structure. A less-explored problem is the visual characterization and quantification of motion complexity. An important motion class that is prevalent in living beings is articulated motion (segments connected by joints). At present, no known standardized measure for quantifying the complexity of articulated motion exists. Such a measure could facilitate advanced motion analysis with applications including video indexing, motion comparison, and advanced biological study of visual signals in organisms.

This dissertation presents an in-depth study of the development of several complexity measures for visual articulated motion. Optical flow is the basis of many motion estimation approaches and our first measure utilizes this as the starting point. Using optical flow, we develop a set of features to characterize different aspects of the motion and combine them to estimate the complexity of the movement.

The second measure also utilizes optical flow, but uses higher-order features as motion descriptors. Specifically, features that encode the periodic nature of movements, synchrony, and movement clusters are developed and used toward the design of a new and improved complexity measure. To validate the measure, a human study was conducted. Subjects were asked to (a) give motion complexity scores to a set of videos

and (b) rank features based on their importance to complexity. Using this study, we developed prediction models to estimate the motion complexity and also classification models to classify the videos.

We use an alternative approach for our third measure based on interesting motion points in the combined space-time domain. These spatial-temporal interest points integrate hidden complexity information in the movement sequence. High level features are proposed to capture different dimensions of movement complexity from these interest points and then combined to estimate the overall complexity of the movement.

All three approaches have been evaluated using two datasets: human movements and wolf spider movements. Extensive evaluation of the measures show the accuracy of estimating the complexity of articulated motion, and demonstrate the efficacy of their use toward classifying motion based on complexity.

DEDICATION

To my wife and family for their unwavering love and support. I love you all...

ACKNOWLEDGMENTS

I am extremely grateful to the numerous people I have encountered throughout my academic journey over the past few years that have helped me reach this finish line. But first and foremost, I want to thank God for the time and talents He has provided me. Through Him, all things truly are possible.

There are many people at UNL that deserve thanks for helping me along the way but I want to start by thanking my advisor, Dr. Ashok Samal, for the endless support, guidance, encouragement, and patience with which he has provided me over the last few years; I can only hope to mentor my future students as well as he has mentored me! I want to thank Dr. Eileen Hebets for her help and support in my work, but especially for planting the initial ideas that brought this very research to life; never in a million years would I have imagined myself studying wolf spiders! I want to thank the other members of my Supervisory Committee (Dr. David Marx, Dr. Carrick Detweiler, and Dr. Stephen Scott) for their guidance, support, suggestions, and other input toward this research and dissertation; every bit of feedback was very much appreciated. I am also extremely grateful to all the students and others who assisted me by participating in the various stages of my research for their help, advice, friendship, and encouragement. Even the smallest amount of participation or interaction contributed greatly to the work in this dissertation.

The idea of continuing on to graduate school did not spark until I was working on my undergraduate degree at Doane College, and so I have many people that inspired me there to give recognition toward as well. Specifically, I want to thank my computer science professors (Dr. Mark Meysenburg and Dr. Alec Engebretson) and mathematics professors (Dr. Christopher Masters, Dr. Jim Johnson, Peggy Hart, and J.L. Vertin) for providing me with the very foundational principles I have carried with me into my

graduate studies; the concepts and ideas I learned from them have helped me succeed every single day.

To the people that have been at my side my entire life, there is not enough thanks in the world I can give. Many thanks to my parents (Dallas and Vickie Christ), my brother (Chase Christ), and my grandmother (Betty Slagle) for the constant love and encouragement to pursue my extremely long list of dreams. To my grandparents who are no longer with me (Larry Slagle, and Loraine and Irene Christ), I will never forget the love and encouragement I received from you as well. I would like to thank my uncle, Tom Seevers, who has has been a role model to me in more ways than one. Without the support of my family, I am certain I would not have made it to where I am now.

I am also grateful for the companionship of my hamster, Ace, who would keep me company during the late hours of the night while writing this dissertation; he could always make me smile even in the most stressful times. And finally, I would like to thank my wonderful, supportive, loving wife (Christa) for her endless encouragement, advice, patience, and love. She was always there when I needed support of any kind, and did so much for me while I was busy finishing up this dissertation that I cannot even begin to express my gratitude. You have my love forever.

To everyone mentioned here and to anyone I might have missed, thank you...

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

This first chapter introduces the problem of characterizing and quantifying the complexity of visual articulated motion in video, in addition to listing the motivations toward pursuing further study in this domain. The approaches taken to address this problem that are detailed throughout the rest of this dissertation are briefly stated, along with the contributions and overall structure of this document. Related work and background material needed for a full understanding of the work presented throughout this dissertation are left for Chapter 2.

1.1 Overview

Motion is an important and powerful indicator used in many computer vision algorithms and applications, and can be estimated with remarkable accuracy by computationally examining a series of sequential images. Motion estimation is one of the oldest problems in the computer vision domain, and continues to receive a considerable amount of attention due to the abundant number of applications that rely on it. Some of these applications of motion estimation include foreground/background segmentation,

camera stabilization, crowd-flow analysis, health and rehabilitation, visual anomaly detection, and estimating three-dimensional structure. Motion estimation is also important in areas such as robotics, where robots can utilize it to navigate a complex environment.

A less-explored domain of motion estimation is its use as a description of how visually complex a given movement or series of movements appear over a given period of time in a video. One important motion class that is prevalent regarding living beings is *articulated motion*, where the observed movements involve a set of segments connected by flexible joints. This can also be thought of as limb-based movement (arms, legs, etc.) observed in various living beings. To the best of our knowledge, the current literature suggests that there is no existing standardized measure for quantitatively describing visual articulated motion complexity, and little work has been done toward the construction or the understanding of one. However, having such a measure available could greatly benefit a variety of research communities and allow for a more advanced analysis and understanding of motion. A few uses of having such a complexity measure available include video indexing (such as searching for videos of springboard dives more difficult to perform than some specified springboard dive), motion comparison (such as comparing and contrasting one dance routine from another), motion classification (such as identifying one species from another based on the complexity of observed motions), and advanced biological study of visual signals in organisms (such as the changes in visual communication of a spider that has been fed a large nutrient intake versus one that has been fed a low nutrient intake).

This dissertation presents an in-depth study of visual articulated motion complexity using algorithms from the computer vision domain. Throughout this work, we follow three assumptions about the observed video: (1) the camera is stationary, (2) there is only a single subject in the video, and (3) the observed motion is articulated (limb-

based). We begin with an approach that utilizes a popular optical-flow technique for estimating frame-by-frame motion, then use the computed estimation to extract various statistical flow-based features and propose a general complexity measure for motion. Second, we examine how to build upon these flow features for computing higher-order statistical features toward the design of a new, improved complexity measure. This includes using techniques such as Fourier analysis for determining repeating movement patterns and their relationship to complexity. The potential of using the measure for classification and predicting new complexity scores is demonstrated in conjunction with a user study of motion complexity, allowing for a comparison of pattern-recognition-based approaches against approaches based on human opinion. Finally, we abandon optical flow by proposing a third type of complexity measure based on the extraction of spatial-temporal features typically used in the action/activity recognition domain. The goal is to discover any complexity information in the space-time domain that may have otherwise been hidden by only examining the spatial or temporal domains separately. While classification and prediction abilities of the measures are observed throughout the dissertation, the goal of this study is to identify the motion signatures that contribute the most toward quantifying complexity, ultimately providing a better understanding of motion compelxity.

1.2 Motivation

In this dissertation, we aim to address the problem of how to quantitatively describe the complexity of visual articulated motion in video, while identifying the individual components that contribute the most toward complexity. Our efforts are specifically focused in two motion domains: (1) the motion displayed during the courtship routine of a pair of Schizocosa wolf spider species, and (2) the motion displayed by humans

performing basic actions (such as walking and waving a hand). The approaches utilized, however, are generalized in such a way that they could be applied to any general visual articulated motion using a stationary camera and a single subject. While much of the research in automated motion analysis has focused on human subjects due to the abundance of useful applications and readily accessible video data [1, 10, 18, 22, 24, 31, 37, 45, 51, 56, 61], an analysis of spider movements presents unique challenges both in their visual and auditory signals. For example, spiders have a vastly smaller size, a mostly uniform appearance among different specimens, and a differing variety of movements as compared to humans. Their movements tend to be very quick, enough so that they are difficult to fully understand by direct observation of the naked eye. In contrast, humans have a larger size, a variable appearance (different hair styles and colors, different clothing, etc.), and more complex movements that lead to diverse activities (such as brushing their teeth or playing tennis). With regards to sound, humans have different vocal sounds and speaking styles that make each person unique, while spiders rely on vibratory sounds, tapping, and scraping. These differences are summarized in Table 1.1.

To the best of our knowledge, he current literature suggests that there is no existing standardized measure for quantifying the complexity of visual articulated motion. Such a measure, however, could greatly benefit a variety of communities and allow for more advanced analysis of motion. It would also allow for advancing the study of biological species (such as the analysis of signals displayed during various types of communication). For example, the biology community has shown much interest in performing advanced analysis of the visual and auditory signals from various living organisms, such as Peters and Evans [50] with Jacky dragons, How et al. [27] with various species of fiddler crab, and Elias et al. [19] and Chiarle and Isaia [12] with spiders. The availability of a complexity measure would help bring unification to these

Feature	Human	Spider
Stillness	slight movement	perfectly stationary
Articulators	2 arms, 2 legs, head, torso	8 legs, 2 pedapalps, cephalothorax, abdomen
Overall size	large	small
Appearance	variable (clothing/hair/etc.)	mostly uniform
Movements	more variety	less variety
Movement speed	slower	more rapid
Audible output	mostly voice	mostly vibration

Table 1.1: A comparison of the visual and auditory traits of humans and spiders.

types of studies.

This work is also motivated by a desire to understand the differences between the human belief of complexity elements versus those identified by algorithms. There is no standard definition for what makes visual articulated motion complex, and one person may disagree from another as to whether one motion is more or less complex than another. The work in this dissertation provides sound evidence of the individual contributors toward complexity, and the degree to which they contribute. These identified domains and corresponding features are detailed in Section 3.2, Section 4.2, and Section 5.2 of this dissertation. Toward gaining insight into the human understanding of complexity, two user studies are presented in this dissertation for polling humans on their complexity beliefs. These studies allow for the identification of what humans agree versus disagree on regarding complexity. The first study (utilized in Chapter 3) provides insight into the understanding of a group of researchers that have experience studying spiders and their movements, providing an expert-based

opinion on the important contributing features. The second study (utilized in both Chapter 4 and Chapter 5) expands the complexity study to a general audience, providing a non-expert-based view of complexity beliefs. These user opinions are also used as ground truth information, as complexity has no standardized definition available to utilize for determining correctness.

1.3 Approaches

This dissertation details several novel approaches toward the creation of a complexity measure for visual articulated motion. Each one is based on algorithms for estimating motion in video, which in turn are used to generate higher-order sets of motion complexity features. These approaches are divided throughout this dissertation into three separate bodies of work as follows:

Approach 1 - The first approach focuses on computing the optical flow motion estimation of a series of video samples displaying wolf spider movements recorded with a high-frame-rate camera (250 FPS). From the estimated flow, a set of statistical-based features are computed for describing various identified complexity domains that are believed to have influence on the overall complexity values. A weighted-sum measure is constructed from the flow features that utilizes the opinion of a panel of experts (spider researchers) for determining the weights of each feature. The final computed complexity values are demonstrated in a classification experiment on the spider samples.

Approach 2 - The second approach expands on the optical flow technique of the first approach by creating a new set of features from the computed flow estimation that takes into account six identified motion-complexity domains believed to

address the various contributing aspects of complexity. Two new domains unexplored in the first approach include motion repetition (using techniques from Fourier analysis to extract the primary frequencies) and motion synchrony (measuring the degree that multiple areas of movement in a video are moving at the same time). The efficacy of using the new features for measuring complexity is demonstrated using a weighted-sum measure (as in the first approach), as well as trained linear-discriminant classifiers for distinguishing motion classes and predicting complexity scores for new motion samples. A sequential feature selection algorithm is utilized to identify the complexity features that contribute the most toward correctly predicting complexity scores and accurately classifying motion classes. In addition, a user study on motion complexity is presented for demonstrating participant belief of complexity domains and for use in training the classifiers as ground truth information.

Approach 3 - The third and final approach abandons the optical flow technique of the first two approaches for spatial-temporal features as the basis for motion complexity features. While the previous two approaches compute features based on local information (between two frames), spatial-temporal features integrate both space and time to determine where interesting and significant motion is happening within the video volume. Unlike the first two approaches, this approach completely disregards directional information in favor of only using the locations and characteristics of the space-time interest points in the space-time volume. The efficacy of using the new features for measuring complexity is demonstrated using trained linear-discriminant classifiers for distinguishing motion classes and predicting complexity scores for new motion samples. A sequential feature selection algorithm is utilized to identify the complexity

features that contribute the most toward correctly predicting complexity scores and accurately classifying motions. In addition, the user study on motion complexity from the second approach is also applied here for use in training the classifiers as ground truth information.

1.4 Contributions

This research presents a number of novel and interesting ideas, as well as identifies sets of motion complexity features. These contributions are aimed at providing not only a concrete understanding of what makes motion complex, but also a usable measure for more advanced understanding of organisms in other research domains. The overall contributions of this dissertation are as follows:

1. Identifies novel sets of motion complexity features based on both optical flow for encoding the various aspects of articulated motion complexity, as well as space-time interest points for integrating hidden complexity information

2. Defines a measure for quantifying general motion complexity by integrating the motion features as a weighted sum based on feature contribution

3. Demonstrates the performance of a pattern-recognition (linear discriminant analysis) model based on optical flow for predicting motion complexity scores and distinguishing motion classes

4. Conducts and presents the results of two user studies on visual motion complexity: (1) an expert poll on statistical feature importance for complexity, and (2) a user study where participants rate a dataset of videos for further analysis toward what a typical person believes contributes to complexity

5. Demonstrates the accuracy of a spatial-temporal feature approach for predicting motion complexity scores and distinguishing motion classes

6. Identifies the key contributing factors toward the quantification of visual motion complexity

7. Demonstrates the efficacy of the defined complexity measures in a real-world problem domain (the biological study of visual signals from spider movement)

8. Contributes a significant body of work toward several fields of study (planned for publication in [13–15])

1.5 Document Structure

Here, we provide a detailed outline of the entire dissertation by giving a brief summary of each chapter. The introductory chapters include Chapter 1 and Chapter 2, while the main body of work is found in Chapter 3, Chapter 4, and Chapter 5. Specifically, this dissertation is structured as follows:

Chapter 1 introduces the problem of visual motion complexity analysis for general articulated motion along with motivations, approaches, and contributions, as well as this detailed outline of the dissertation.

Chapter 2 continues the introduction by presenting the background material needed for a fuller understanding of the remaining chapters in this dissertation, as well as a literature review of several previous studies and works regarding visual motion measures and the visual analysis of motion displayed in both human and non-human species.

Chapter 3 introduces the first approach: a weighted-sum motion complexity measure based on statistical features computed from optical flow. Its performance is demonstrated on a dataset of Schizocosa wolf spider movements displayed during their courtship routine, and the most important features are noted as identified by a feature selection process. This process includes a polling of experts in spider research, utilized to weight the features based on expert opinion.

Chapter 4 expands on the ideas of Chapter 3 by introducing a new set of optical flow-based features using higher order information such as frequency domain analysis for detecting repeating patterns of motion, and motion synchrony for measuring the degree to which multiple areas of movement are occurring. Motion classification and complexity prediction performance are demonstrated on two datasets: 1) a set of wolf spider movements and 2) a set of basic human actions. In addition, a user study on motion complexity is presented for identifying the features that are most important based on human belief. The user study also provides ground truth information for measuring prediction and classification accuracy.

Chapter 5 expands on the ideas of Chapter 3 and Chapter 4 by transitioning from features in the spatial domain (computed from the optical flow values between two image frames) to features in the space-time domain (space-time interest points). The aim is to reveal hidden complexity information not otherwise observed using a strictly optical flow-based technique. Motion classification and complexity prediction performance are demonstrated on the same two datasets from Chapter 4, and the same user study is utilized for ground-truth information during training and testing of the models toward complexity prediction and classification.

Chapter 6 concludes this dissertation with a discussion of the overall results along with some final closing remarks. Suggestions and ideas for further work in the domain of visual complexity analysis of articulated motion are also provided.

In addition to the previously mentioned chapters, several appendices have been included at the end of this dissertation for supplemental information and other results. These additional chapters are structured as follows:

Appendix A summarizes and details the expert-poll study that is presented and discussed in Chapter 3. The expert poll was performed to gain an understanding of the beliefs of complexity from a group a spider motion researchers (experts), which in turn is used to weight the features for the complexity measure.

Appendix B provides further information regarding the complexity rating study presented in Chapter 4. This in-depth study was performed to gain an understanding of what a typical (non-expert) person believes contributes toward the measuring of visual complexity. This information is utilized in both Chapter 4 and Chapter 5 as ground truth information for training and testing the prediction and classification models.

Appendix C details the two video datasets used throughout this work: 1) a dataset of wolf spider movements displayed during their courtship routine, and 2) a dataset of basic human actions (walking, waving, etc.).

1.6 Summary

This chapter introduced the problem of characterizing and quantifying the complexity of visual articulated motion in video, in addition to listing the motivations toward

pursuing further study in this domain. The approaches taken to address this problem that are detailed throughout the rest of this dissertation were briefly stated, along with the contributions and overall structure of this document. In the next chapter, we present the related work and background material needed for a full understanding of the topics and concepts presented throughout the rest of the dissertation.

Chapter 2

Background & Related Work

In this chapter, we begin with a discussion of the general concept behind optical flow, a detailed description of the optical flow algorithm that is utilized in both Chapter 3 and Chapter 4, and an overview of some of the strengths and weaknesses of using such a technique. Next, space-time interest points are mathematically defined and discussed, which are utilized for the visual motion complexity features in Chapter 5. Finally, a literature review is presented regarding the previous work toward visual motion measures as well as the visual motion analysis of the observed motion displayed by both human and non-human species.

2.1 Optical Flow

This section reviews the concept of optical flow for motion estimation, utilized in both Chapter 3 and Chapter 4 for computing the signatures for visual articulated motion complexity. A general overview is presented, along with the strengths, weaknesses, and alternative methods for motion estimation.

2.1.1 Overview

Optical flow (the apparent observed motion) is one of the oldest and most researched domains in computer vision, and continues to receive a considerable amount of attention. One of the main reasons for this continued interest is due to its use in a wide array of useful applications including three-dimensional reconstruction, object detecting/tracking, foreground/background segmentation, robotic navigation, and traffic analysis. A few examples of these application domains are examined in the work by O'Donovan [47].

While many optical flow algorithms exist, they all aim to solve the same problem: Where does each pixel in frame I^t move to in frame I^{t+1}? That is, optical flow is a motion-estimation algorithm for computing the displacement of the pixels between two sequential frames of a video. The goal at a higher level is to determine where the edges, corners, and other objects in a video frame move to in the next frame. The majority of the optical flow algorithms rely on any given moving pixel retaining the same intensity value in its displaced location from one frame to the next. This constraint is called the *brightness constancy constraint*. The output of any one of these optical-flow algorithms provides two key pieces of information about a pixel: (1) the distance the pixel moved (strength/speed of motion) and (2) the direction that the motion occurred. The set of all displacements of the pixels between two video frames is a set of two-dimensional vectors called the *motion field*. A recent survey of optical flow can be found by Fortun et al. [21], which organizes current approaches and practices.

In general, optical flow provides a way to estimate the raw motion of video. Alternative methods for motion estimation include block matching [5, 48] (matching neighborhoods of pixels to better correspond to the motion of real image artifacts)

and phase correlation [2, 57] (which utilizes a Fourier transform to determine the translation of an image). While optical flow shows remarkable accuracy at estimating motion in video, several weaknesses do exist. For example, if an object (such as a square) contains the same intensity value at every pixel, the best optical flow can do is estimate the edge motion. While this may be acceptable for some domains, others may rely on the movement of every single pixel to be accurate. Due to the assumption of moving pixels retaining their intensity values between frames, errors in the estimation arise with fluctuating lighting and prevalent shadows. One of the most cited issues with optical flow is the barber pole problem (which itself is an instance of the optical flow *aperture problem*. That is, assume you have a barber pole spinning on its cylindrical axis with a single stripe that wraps around from top to bottom. As the pole rotates, the stripe rotates horizontally with it. Even though the stripe rotates horizontally around the cylindrical axis, it visually appears as if the stripe is moving vertically upward or downward (depending on the direction of rotation). Optical flow will estimate the motion as moving vertically, while the actual motion field is moving horizontally. For many applications, however, these issues can be safely ignored.

2.1.2 Horn-Schunck Algorithm

Two of the most popular optical flow algorithms are the Horn-Schunck algorithm [26] and the Lucas-Kanade algorithm [42], with a comparison of the two given in [6]. In this dissertation, we specifically apply the Horn-Schunck algorithm, categorized in [21] as a regularization model that utilizes a spatial flow gradient constraint. As the general optical flow problem is under-constrained (an equation with two unknowns), optical flow algorithms need to utilize at least one more constraint toward the goal of motion estimation. With Horn-Schunck, this additional requirement is the constraint

of *motion smoothness*. That is, the algorithm chooses solutions in which the motion is smoothest (the rate of the change in velocity is nearest to zero). While any optical flow technique could be used to compute the features, the more traditional Horn-Schunck approach was chosen because (1) computation speed is not a concern to us at this time, (2) the majority of the species-analysis literature utilizes the algorithm and we desire to use an algorithm already understood by that community, (3) the flow of the interior parts of similar objects can be determined from the motion boundaries, and (4) empirical experimentation revealed sufficient accuracy for the utilized datasets in this work.

The algorithm computes a series of optical flow images $O = [O^1, O^2, \ldots, O^{t-1}]$ in which O^i is the optical-flow image between video frames F^i and F^{i+1}, $O^i_{x,y} = [u^i_{x,y}, v^i_{x,y}]$ is the set of optical flow vectors at time t, (x, y) is the spatial location of a pixel, and $[u, v]$ is a flow displacement vector (where u is the horizontal displacement and v is the vertical displacement). The brightness constancy constraint states that $F^i_{x,y} = F^{i+1}_{x+u,y+v}$. The optical flow problem using Horn-Schunck is an estimation of partial derivatives followed by a minimization of the sum of the errors generated by an iterative process. It ultimately favors smooth motion over non-smooth motion. The approach is defined as the minimization of a global energy functional

$$E = \int\int [(I_x u + I_y v + I_t)^2 + \alpha^2(|| \bigtriangledown u||^2 + || \bigtriangledown v||^2)]dxdy \tag{2.1}$$

where α is a regularization constant for motion smoothness, $\bigtriangledown$ is the gradient operator, and I_x, I_y, and I_t are the image intensity derivatives along the spatial (x, y) and temporal (t) dimensions. An iterative approach is used to minimize this functional

and solve for the displacement vector as follows:

$$u^{k+1} = \overline{u} - \frac{I_x(I_x\overline{u}^k + I_y\overline{v}^k + I_t)}{\alpha^2 + I_x^2 + I_y^2} \tag{2.2}$$

$$v^{k+1} = \overline{v} - \frac{I_y(I_x\overline{u}^k + I_y\overline{v}^k + I_t)}{\alpha^2 + I_x^2 + I_y^2} \tag{2.3}$$

where k represents the iteration number ($k+1$ denotes the next iteration) and $\overline{u}$ and $\overline{v}$ are a weighted average of u and v, respectively. The computation stops when the values of $[u, v]$ converge, or after a specified number of iterations. Thus, u is the estimated horizontal displacement of motion, while v is the estimated vertical displacement of motion.

2.2 Space-Time Interest Points

This section reviews several spatial-temporal feature detectors and descriptors, specifically emphasizing *selective space-time interest points (S-STIPS)*. These concepts are utilized in Chapter 5 for computing the signatures of visual articulated motion complexity. A general overview is presented, along with the strengths, weaknesses, and alternative methods for motion estimation.

2.2.1 Overview

The last decade has seen a surge in interest toward the field of visual activity recognition [46, 52, 58]. Due to the wide array of applications (such as health and security) and readily available datasets, the majority of the work toward action recognition has focused on human applications. Several types of strategies have been proposed toward modeling human actions such as those based on human models, trajectories, holistic

models, and local descriptors. As action recognition has evolved, more difficult challenges needed to be addressed. One such challenge is the complexity of scenes (moving cameras, noisy backgrounds, illumination changes, etc.). We next address a few of these STIP detectors.

Space-Time Interest Points To help overcome these challenges, a new concept called space-time interest points (STIPs) were introduced. STIPs were first proposed by Laptev and Lindeberg [35] for the purposes of action recognition by extending the popular Harris corner detector [25] from 2D to 3D. Regions having high intensity variation in both space and time are detected as spatial-temporal corners to indicate "interesting" movement in the spatial-temporal volume. A visualization of these STIP points can be seen in Figure 2.1.

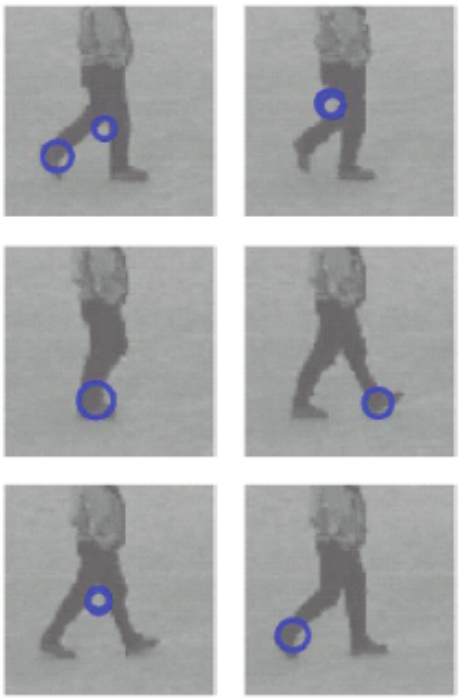


Figure 2.1: The detection of space-time interest points (image taken directly from [34]).

Space-Time Cuboids Many improvements to the initial STIP detector have been implemented in several works. Dollr et al. [17] apply temporal Gabor filters while selecting regions of high responses and applying a cuboid-based descriptor of the points. These cuboids encapsulate the motion happening around the interest points as histograms of optical flow directions and magnitudes, as well as the intensity values of the individual pixels. A visualization of cuboids is shown in Figure 2.2.

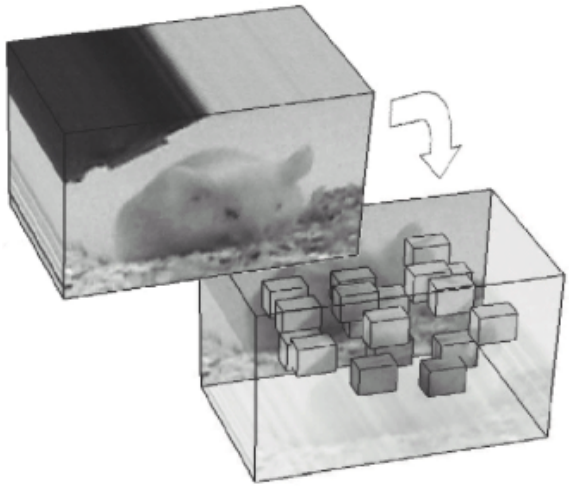


Figure 2.2: The detection of space-time cuboids (image taken directly from [17]).

Space-Time Velocity Histories Messing et al. [44] perform an alternative approach by tracking the detected STIPs over time to recognize actions using the velocity histories of the tracked points. Instead of only utilizing the information (such as magnitude, direction, and response) around each STIP point, the idea is that more useful information may be detected in the paths that the STIP points take over time. A visualization of velocity histories (displaying the tracked paths of the points) is shown in Figure 2.3.

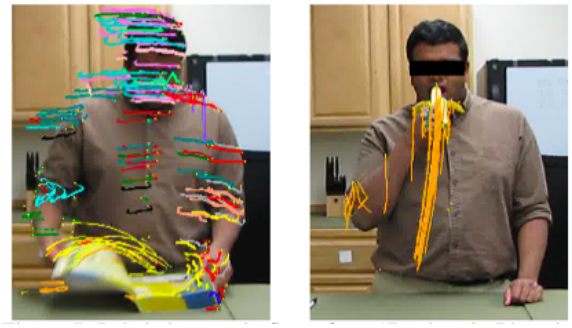


Figure 2.3: The detection of velocity histories (image taken directly from [44]).

Discussion Many other STIP-based approaches exist, and the activity recognition domain utilizing STIP points continues to receive considerable work. Wong and Cipolla [60] introduce an approach based on global information and select STIPs based on their probability of belonging to a relevant group of motion, while Willems et al. [59] propose dense STIPs that are an extension of the Hessian saliency measure. While not specifically used in our work, it is also worth mentioning space-time descriptors. While all of these STIP-point approaches exist to detect interesting locations in space and time, space-time descriptors exist as a way to describe the detected STIP points for training activity-recognition systems. The information included in a descriptor is generated from the shape or motion around the STIP point. A few works detailing some of the more popular descriptors for STIPs can be found in [17,32,33,35,36,54,59]. The descriptors are then used to form vocabularies of visual words, typically for use in a bag-of-video words video model for action recognition [9, 17, 38, 39, 62].

2.2.2 Selective Space-Time Interest Points

All of these approaches, however, tend to be vulnerable to moving cameras and noisy backgrounds. A recent technique by Chakraborty et al. [9] shows promise of overcoming the challenges encountered by other STIP-based techniques. Their approach detects a

set of spacial interest points (SIPs), suppresses unwanted background points, then imposes local and temporal constraints to obtain a more robust set of *selective STIPs*. In this chapter, we utilize the selective-STIP approach to build a set of visual motion complexity features in Chapter 5.

The selective-STIP approach is divided into several steps. Here, we describe the process in detail, as outlined in [9].

Detecting Spatial Interest Points The initial set of SIP points are first detected using a basic Harris corner detector [25]. The corner detector is initialized using corner strength C_σ, where σ is the spatial scale of the points. This initial set typically contains a large number of uninteresting "background" points, which are filtered in the remaining steps.

Suppressing Background Interest Points To suppress background points, a *surround suppression mask (SSM)* is used for every interest point, with the current point under consideration as the mask center. The influence of all surrounding points of the mask on the central point is estimated, and a suppression decision is made on whether the point is a background point or not. The idea behind SIP suppression is that the majority of corner points detected in the background follow a particular geometric pattern, while those that are on objects of interest are not.

Surround suppression is accomplished by computing an inhibition term for each point of C_σ. A gradient weighting factor is introduced and defined as:

$$\Delta_{\theta,\sigma}(x, y, x-u, y-v) = |\cos(\Theta_\sigma(x, y) - \Theta_\sigma(x-u, y-v))| \tag{2.4}$$

where $\Theta_\sigma(x, y)$ and $\Theta_\sigma(x-u, y-v)$ are the gradients at point (x, y) and $(x-u, y-v)$, respectively, and u and v define the horizontal and vertical range of the SSM mask.

For each point $C_\sigma(x, y)$, a suppression term $t_\sigma(x, y)$ is defined as the weighted sum of gradient weights in the suppression surround of the point:

$$t_\sigma(x, y) = \int\int_\Omega C_\sigma(x - u, y - v) \times \Delta_{\Theta,\sigma}(x, y, x - u, y - u)dudv \tag{2.5}$$

where Ω is the image coordinate domain. An operator $C_{\alpha,\sigma}(x, y)$ is defined as follows:

$$C_{\alpha,\sigma}(x, y) = H(C_\sigma(x, y) - \alpha t_\sigma(x, y)) \tag{2.6}$$

where $H(z) = z$ when $z \geq 0$, $H(z) = 0$ for $z < 0$, and α controls the surround suppression strength. The operator response will retain the original corner magnitude. If a larger number of interest points are detected in the background, the interest point will be suppressed.

Imposing Local Constraints A subset of the initial set of points is selected by applying non-maxima suppression as follows: for every position (x, y), the responses $C_{\alpha,\sigma}(x', y')$ and $C_{\alpha,\sigma}(x'', y'')$ in adjacent positions (x', y') and (x", y") are computed by linear interpolation. A point is kept only if the response $C_{\alpha,\sigma}(x, y)$ is greater than that of the two adjacent points, and discarded otherwise.

Scale Adaptive SIPs A multi-scale approach is used for scale selection. Suppressed SIPs are computed at five different scales $S_\sigma = \{\frac{\sigma}{4}, \frac{\sigma}{2}, \sigma, 2\sigma, 4\sigma\}$, where the best set of SIPs for each scale are kept based on the maximization of a normalized differential invariant.

Imposing Temporal Constraints To suppress the SIPs that might remain due to being static, temporal constraints are imposed. By considering two consecutive

frames at a time, the common interest points are removed as static points do not contribute any motion information. An interest point matching algorithm is used to adjust for camera motion. The entire selective-STIP process is divided into five algorithms, and are presented as Algorithm 1, Algorithm 2, Algorithm 3, Algorithm 4, and Algorithm 5 (which are adapted from [9]).

input : An image stack ($h \times w \times t$): iS

Array containing spatial scales: sA;

Alpha: α;

Mask: m;

output: Detected STIPs: $stip$

$sip = \{\}; stip = \{\}$;

$t = size(iS, 3)$;

for $i = 1 \rightarrow t$ **do**

  for $j = 1 \rightarrow size(sA)$ **do**

    $sip \leftarrow sip \cup \{SCD(iS(:, :, i), sA(j), \alpha, m), sA(j)\}$;

  end

  $stip \leftarrow stip \cup blobDetector(iS(:, :, i), sip)$;

end

$stip = temporalConstraint(iS, stip)$;

$Return(stip)$;

Algorithm 1: STIP detection (algorithm adapted from [9]).

input : An image $(h \times w)$: $image$;

Spatial scale: σ;

Alpha: α;

Mask: $mask$;

output : Detected selective spatial interest points: sip

$cp = harrisCorner(image, \sigma)$;

$cornerPoints = find(cp > 0)$;

$cp = cp(cornerPoints)$;

$\Theta = gradient(image)$;

$sip = \{\}$;

for *Each point* $(x, y, \sigma) \in cornerPoints$ **do**

$\Delta_{\Theta_{mask}} = |\cos(\Theta_{mask} - \Theta_{mask_{(x,y)}})|$;

$t(x, y) = cp_{mask} \otimes \Delta_{\Theta_{mask}}$;

$cp(x, y) = H(cp_{(x,y)} - \alpha t_{(x,y)})$;

$(x', y') = round(line(x, x + 1, y, \Theta(x, y)))$;

$(x'', y'') = round(line(x, x - 1, y, \Theta(x, y)))$;

if $(cp(x, y) > cp(x', y')) \wedge (cp(x, y) > cp(x'', y''))$ **then**

$sip \leftarrow sip \cup (x, y, \sigma)$;

end

end

Return(sip) ;

Algorithm 2: SCD: Selective STIP Detection (algorithm from [9]).

input : An image $(h \times w)$: im;
Corner points: $corners$;

output : Detected selective spatial interest points based on Gaussian blob strength: $cornerPoints$

$cornerPoints = \{\}$;

for *Each point* $(X, Y, \sigma) \in corners$ **do**

  $bS = \sigma^{1.75} * L_{y,im}(X, Y) * L_{xx,im}(X, Y)$;

  if $(bS > \tau)$ **then**

    $cornerPoints \leftarrow cornerPoints \cup (X, Y, \sigma)$;

  end

end

Return($cornerPoints$) ;

Algorithm 3: blobDetector: Corner strength detection using Gaussian blob (algorithm from [9]).

input : An image stack ($h \times w \times t$): iS;

Spatial corner points: cp;

output : Detected STIPs: $stip$

for $i = 1 \rightarrow h$ **do**

  for $j = 1 \rightarrow w$ **do**

    $gabor(i, j, :) = gaborFilter1D(iS(i, j, :))$;

  end

end

for $i = t \rightarrow 2$ **do**

  $f_1 = iS(:, :, i)$;

  $f_2 = iS(:, :, i - 1)$;

  $g_1 = gabor(:, :, i)$;

  $g_2 = gabor(:, :, i - 1)$;

  $im_1 = iS(:, :, i)$;

  $im_2 = iS(:, :, i - 1)$;

  $cp_{f_1} \leftarrow cp_{f_1} \setminus pointMatch(cp_{f_1}, cp_{f_2}, g_1, g_2, im_1, im_2)$;

end

Return(cp) ;

Algorithm 4: temporalConstraint: Imposed temporal constraint on the selected spatial corner points (algorithm from [9]).

input : Image frames: im_1, im_2;

Corner strengths: cp_1, cp_2;

Gabor strengths: g_1, g_2;

output : Detected matching STIPs: mS

$mP = \{\}$;

$cornerPoints_1 = find(cp_1 > 0)$;

$cornerPoints_2 = find(cp_2 > 0)$;

for *Each point* $(x_1, y_1, \sigma_1) \in cornerPoints_1$ **do**

  $h = \sigma_1$;

  for *Each point* $(x_2, y_2, \sigma_2) \in cornerPoints_2$ **do**

    $similarity = \frac{min(cp_1(x_1,y_1),cp_2(x_2,y_2))}{min(cp_1(x_1,y_1),cp_2(x_2,y_2))}$;

    $w = \sigma_2$;

    if $similarity > \tau_{sim}$ **then**

      $a_1 = cropRect(im_1, x_1, y_1, h, w)$;

      $a_2 = cropRect(im_2, x_2, y_2, h, w)$;

      $sC = crossCorrelation(a_1, a_2)$;

      if $(sC > \tau_{corr}) \wedge (g_1(x_1, y_1) > \tau_{gabor})$ **then**

        $mP \leftarrow mP \cup (x_1, y_1, \sigma_1)$;

      end

    end

  end

end

Return(mS) ;

Algorithm 5: pointMatch: Detect the set of matching corner points in two consecutive frames (algorithm from [9]).

2.3 Visual Motion Measures

Here, we present a literature review of visual motion measures for quantitatively describing motion. This section is used to illustrate the novelty and usefulness of the work presented in this dissertation. Specifically, it can be noted that the techniques here do not focus on visual articulated motion complexity.

The *perceived motion energy spectrum (PMES)* shot content representation proposed by Ma et al. [43] is based on angle distributions obtained by temporal energy and global motion filters. These filters are used to extract motion vectors from the MPEG stream. Specifically, a temporal energy filter is used to disregard object motion in a scene, and a global motion filter to shield object motions from camera motions. Their metric is tuned to closely match human perception of motion for the purposes of content-based video retrieval.

Liu et al. [40] propose a triangle model of *perceived motion energy (PME)* to model motion patterns for the purposes of extracting key frames from video sequences. PME is a combined metric of motion intensity and motion characteristics with more emphasis on dominant motion. It uses the percentage of dominant motion direction in an entire frame as an estimation of motion intensity. The goal is to identify the acceleration and deceleration points of motion over time, which can be used as a set of key frames where the most salient motion is occurring.

Chen et al. [11] develop *entropy motion value (EMV)*, a motion entropy metric to segment frames with high motion intensity from frames with low motion intensity in sports videos. Incorporating entropy into the metric allowed it to handle camera motion better than the PME metric. They introduce a time series change point detection algorithm that minimizes the homoscedastic error to approximate the motion entropy curve with a piece-wise linear model. The accumulated value is used to decide which

segment is a significant sport event.

Peker et al. [49] create a framework for the automatic measurement of motion activity in video sequences using the MPEG-7 motion activity descriptor [30]. They establish that the intensity of motion activity of a video is a direct indication of its ability to be summarized, and suggest the variance of the motion vector magnitudes is promising as a representative measure of visual motion. The framework is used to determine the highlights of sports videos. The work in [30] details how the MPEG-7 motion standard captures the unique aspects of motion. The goal is to provide descriptors that are easy to extract and match, where both motion activity and motion trajectory meet this objective.

Claypool [16] provides novel metrics for motion and scene complexity in video games, which are *percentage of forward/backward or intracoded macroblocks (PFIM)* for motion complexity and *average of intra-coded block size (IBS)* for scene complexity. The intuition behind the PFIM metric is that a video with visual changes from frame to frame will have these changes encoded while video without visual changes can skip much of the encoding.

Ali [3] quantifies the complexity of visual flows based on optical flow particle trajectories that measure the amount of interaction among objects. This approach is aimed at the application of crowd-flow analysis. Due to the interaction of individual particles in a flow, the two-dimensional trajectories become space-time braids. It is shown that the proposed approach is able to quantify the complexity of the flow, and at the same time provides useful insights about the sources of the flow complexity.

The majority of these proposed methods rely on motion magnitudes and/or directions. These previous works do not take into account a large number of motion features with the possibility of several of them contributing important information to the overall complexity from unrelated complexity domains. In addition, they do not

focus on articulated motion. To the best of our knowledge, little work has been done on finding a general complexity measure for visual articulated motion, and no known standardized measure currently exists. The work that does exist tends to generate various statistics from the optical flow vector directions and magnitudes (as seen in the cited literature). While the approaches presented in Chapter 3 and Chapter 4 rely as well on flow magnitudes and directions computed from optical flow, Chapter 5 utilizes space-time interest points while mostly ignoring direction and magnitude information.

2.4 Motion Analysis of Humans

Here, we provide a brief literature review of some of the interesting works in the domain of visual human motion analysis. While the work in Section 2.3 discussed measures for describing any general motion, this section reviews the work that has specifically been done regarding human movements.

Aggarwal & Cai [1] provide a review of the literature of human motion analysis. Their work is divided into three parts: (1) motion analysis involving human body parts, (2) tracking a moving human from a single view or multiple camera views, and (3) recognizing human activities from image sequences. Poppe [51] also provides a broad overview of human motion analysis, dividing the analysis into a modeling and an estimation phase. Sminchisescu [56] specifically gives an overview of the problem of reconstructing 3D human motion using sequences of images acquired with a single video camera. A more recent literature review can be found from Metaxas & Zhang [45], where they summarize motion analysis methods for nonverbal communication of humans. They summarize and group the methods based on face tracking, facial expression recognition, full body reconstruction, pose estimation, and

activity recognition.

The area of visual surveillance and security has received considerable work in recent years due to the vast number of useful applications for ensuring a safer environment, as well as to help address the issue of rising crime rates. Gowsikhaa et al. [24] provide a survey of methods in automated human behavior analysis with a focus on surveillance systems. They provide an overview of the state-of-the-art algorithms and techniques for abnormal human behavior detection. In addition, Gedikli & Ekinci [18, 22] present several works on human motion detection and analysis system focused on visual surveillance.

Yoo & Nixon [61] present a method for an automated markerless system for describing, analyzing, and classifying the motion observed in human gait. Their system consists of three parts: 1) the detection and extraction of the moving human body and its contour, 2) the extraction of gait figures by joint angles and body points, and 3) the analysis of motion parameters and feature extraction for classifying human gait. Chang & Huang [10] use Hidden Markov Models to describe the observed motion of human gait.

Kahol & Vankipuram [31] focus on the analysis of hand motion by predicting the expertise level of a user wearing a sensor glove and performing surgical movements. They present a novel algorithm that utilizes a dynamic hierarchical layered structure to represent the human anatomy, with low-level parameters to characterize the motion in the layers of this hierarchy (corresponding to different segments of the human body). Their approach achieved a near perfect recognition rate.

Lin & Kulic [37] focus on another key area of human motion detection that has received a considerable amount of attention: health and rehabilitation. They propose an approach for the automated segmentation and identification of movement segments from continuous human-movement time series data that is collected through

ambulatory sensors. Their approach uses a two stage identification and recognition process, based on velocity and stochastic modeling of each motion to be identified.

Due to the abundant applications and readily available datasets, the work in this dissertation utilizes a human collection of videos as one of two datasets. We specifically focus on quantifying the visual articulated motion complexity of humans, as well as identify the various components that make up the complexity. While not a focus of the chapter, the classification abilities of the complexity components toward distinguishing human motions are examined throughout this dissertation as well.

2.5 Motion Analysis of Non-Human Species

While the work in Section 2.3 discussed measures for describing any general motion, this section reviews the work that has been done regarding the visual analysis of communication in non-human organisms. Specifically, there has been previous work in the biological domain on understanding the visual and auditory signals of various species other than humans. Because of the strong interest in the field of biology toward having a measure for the advanced study of organisms, one of the utilized datasets in this dissertation is a dataset of wolf spider movements showing the visual signals displayed during their courtship routine.

Peters and Evans [50] examine the visual and auditory signals in Jacky dragons using optical flow for the purposes of classifying basic actions. They specifically use optical flow to generate *velocity signatures*, which are scatter plots representing the direction and speed of movement for each display component. The main idea is that quantitative analyses of movement-based signals can provide insights into the sensory processes of organisms, leading to a more detailed understanding of the species.

How et al. [27] analyze the dynamic visual signals of the claw-waving display of

various species of fiddler crab. They quantitatively measure features of seven species of fiddler crabs such as the claw path, the elevation of the claw over time, and motion intensity and direction. They find that the structure and timing of the features is species-specific, providing evidence of using motion features for species classification.

Riskin et al. [53] provide a kinematic study quantifying the complexity of bat wings. They assign importances to kinematic variables to address whether dimensional complexity of motion changes with speed, which body markers are optimal for capturing dimensional complexity, and which variables should a simplified reconstruction of bat flight include in order to maximally reconstruct actual dimensional complexity.

Work has also been done specifically on understanding dynamic signals in various species of spiders. Elias et al. [19] use optical flow to create features for investigating the courtship behaviors in jumping spiders. They use speed waveform, speed surface, and speed waterfall plots to demonstrate the ability to computationally differentiate various types of spiders, while pointing out that their technique could be used with any organism displaying dynamic visual signals (such as birds, insects, or mammals). Chiarle and Isaia [12] use optical flow to analyze various courtship elements in two species of wolf spiders aimed at understanding the evolution of the courtship and its role in species delimitation and speciation processes.

The work in this dissertation also commits a great deal of attention toward visual spider signal analysis, because the visual differences between human and spider movements pose interesting challenges. While the previous works in species analysis focus on optical flow features for classification, this work focuses on quantitatively measuring visual cues to describe complexity. There is, however, multiple experiments presented in this dissertation that apply the identified motion complexity features to classification-related tasks. Overall, the goal of our work is to identify the various motion signatures that contribute to complexity, the degree that each one contributes,

and the combination of those signatures to predict complexity values.

2.6 Summary

In this chapter, we presented the general concept behind optical flow, a detailed description of the optical flow algorithm that is utilized in both Chapter 3 and Chapter 4, and an overview of some of the strengths and weaknesses of using such a technique. Next, space-time interest points were formally defined and discussed, which are utilized for the visual motion complexity features in Chapter 5. Finally, an in-depth literature review was presented regarding the previous work toward visual motion measures as well as the visual motion analysis of the observed motion displayed by both human and non-human species. In the next chapter, we propose a first approach toward quantifying visual articulated motion complexity that utilizes optical flow and a feature-weighting technique.

Chapter 3

An Optical Flow Statistical-Based Metric for Motion Complexity

In this chapter, we introduce a statistical-based complexity measure for quantitatively describing visual motion in video. Its usefulness can be applied to tasks such as classification and video indexing, but is demonstrated here as a case study on a database of wolf spider movements displayed during their courtship routine. Objectively assessing the complexity of these action movements may inform a more thorough and detailed understanding of these species, as well as demonstrate the potential for using the measure for describing articulated motion in a general sense. An optical flow-based approach is used to derive interesting visual motion complexity features and demonstrate their utility in understanding motion complexity. The features are combined using a data fusion (weighted sum) approach as the measure. We compare and contrast the motion features of two different species of *Schizocosa* wolf spider, demonstrating the measure on a database of high frame rate (250 FPS) videos. It is shown that these features capture several unique movement traits of these spiders during courtship. This demonstrates the feasibility of our approach to use

motion signatures for representing the complexity of elements observed during spider courtship routines, the complexity of non-courtship spider movements, and ultimately the complexity of general articulated motion. Species classification, while not the aim of this work, is demonstrated on the dataset. A feature selection process is detailed for selecting the most relevant features from the initially large set. A user study from a team of spider motion researchers is also presented to demonstrate human belief of complexity versus that which the computer identifies as important complexity components. The work in this chapter is planned for publication in [14].

3.1 Introduction

Understanding motion is an important task in many application domains. In visual surveillance applications, for example, normal motion patterns can be learned in order to alert users when abnormal motion patterns are detected, possibly indicating a security threat in progress. However, motion can be characterized in several different ways. For example, there are short motion patterns (such as a person kicking their leg) as well as longer "tracked" motion patterns (such as the path a person walks in a surveillance video). Similarly, different motion patterns can have different levels of complexity. For example, the motion of a person waving a hand is less complex than a person performing a sophisticated dance routine. An important question in this context is: Is it possible to characterize the complexity of motion using a numerical metric? Such a measure could be useful in a number of applications such as video indexing (such as searching for videos of springboard dives based on difficulty), motion comparison (such as comparing and contrasting one dance routine from another), motion classification (such as identifying one species from another based on the complexity of observed motions), and advanced biological study of visual signals in

organisms (such as the changes in movement of a spider that has been fed a large diet versus one that has been fed a low diet).

In this chapter, we address the problem of quantifying motion complexity. Our efforts are focused in a single motion domain: a case study that includes an analysis of the motion displayed during the courtship routine of a pair of *Schizocosa* wolf spider species. Our approach, however, can be applied to any general articulated motion with a stationary camera. While much of the research in automated motion analysis has focused on human subjects due to the abundance of useful applications and readily accessible video data [1, 18, 22, 24, 31, 37, 45, 61], analysis of spider movements presents unique challenges both in their visual and auditory signals [12,19]. For example, spiders have a vastly smaller size, a mostly uniform appearance among different specimens, and a differing variety of movements as compared to humans. Their movements tend to be very quick, enough so that they are difficult to fully understand by direct observation of the naked eye. In contrast, humans have a larger size, a variable appearance (different hair styles and colors, different clothing, etc.), and more complex movements that lead to diverse activities (such as brushing their teeth or playing tennis). With regards to sound, humans have different vocal sounds and speaking styles that make each person unique, while spiders rely on vibratory sounds and tapping. Table 3.1 summarizes these key differences between the movements shown during human activities versus those shown during spider courtship routines. We specifically focus on two species of *Schizocosa* wolf spider: *S. bilineata* and *S. crassipalpata*.

3.1.1 Problem Definition

The problem addressed in this chapter is the creation of a measure for quantifying visual motion complexity with a focus on wolf spiders. We formally define the problem

Feature	Human	Spider
Stillness	slight movement	perfectly stationary
Articulators	2 arms, 2 legs, head, torso	8 legs, 2 pedapalps, cephalothorax, abdomen
Overall size	large	small
Appearance	variable (clothing/hair/etc.)	mostly uniform
Movements	more variety	less variety
Movement speed	slower	more rapid
Audible output	mostly voice	mostly vibration

Table 3.1: A comparison of human and spider traits.

of finding a complexity measure as follows: Given a video $V = [F^1, F^2, \ldots, F^n]$ where F^i is a video frame and n is the total number of frames in the video, the goal is to define a complexity function $C : V \rightarrow [0, 1]$, where $[0, 1]$ is the set of real numbers between 0 and 1, inclusive. Here, $F^i_{x,y}$ indicates the pixel in the x^{th} row and y^{th} column of the i^{th} image frame of the video. Thus, we aim to find a function C that takes a motion sequence of images (video) as input and generates a value between 0 (lowest possible complexity) and 1 (highest possible complexity).

3.1.2 Approach

An optical flow-based approach is used to estimate the basic elements of visual motion. A discussion and overview of the optical flow technique used in this chapter is detailed in Chapter 2.1. Specifically, we utilize the Horn-Schunck algorithm [26] for the optical flow computation using the default parameters as specified in MATLAB 2015a[1]. While

[1]www.mathworks.com

any optical flow technique could be used to compute the features, the Horn-Schunck approach was chosen because 1) the speed of computation is not a concern to us, and 2) the majority of the species-analysis literature utilizes that algorithm and we desire to use an algorithm already understood by that community. These optical flow motion elements are used to derive vectors of *local temporality features* (one value per feature per frame), which are then further refined into *global temporality features* (one value per feature per video). These global temporality features are reduced to a more manageable and useful quantity using a feature selection process, then used as building blocks for the final measure of complexity. This feature selection process includes computing correlation to detect similar (and redundant) features, as well as utilizing the results of a user study on a group of experts in the spider-motion domain. An overview of this approach is shown in Figure 3.1.

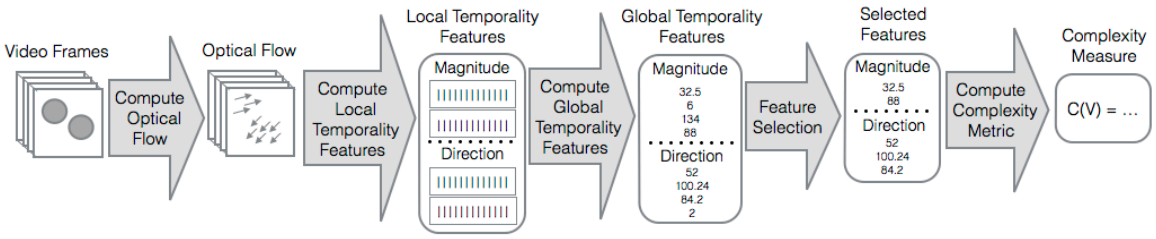


Figure 3.1: Our approach for computing the complexity measure.

The optical flow is computed using the displacement of fixed points between two successive frames of video. We define a number of optical-flow-based features (or *motion complexity features*) which describe different aspects of the observed motion, then present a measure for the complexity of the motion that utilizes these features by integrating them together with a data-fusion approach (weighted sum). The feature weights signify the degree to which each feature contributes to the final complexity value. The efficacy of using these motion complexity features for detecting the important motion signatures is demonstrated using a dataset of spider actions

recorded during the courtship routines of two species of wolf spider, but is applicable to any scenario involving a stationary camera and a single subject remaining mostly still except for articulator-based movements. While previous attempts have typically focused on using few features to determine a measurement of complexity, our approach takes into account a multitude of various features based on optical flow with the goal of integrating several important factors into the overall complexity measure.

3.1.3 Contributions

This chapter makes a number of useful contributions toward a variety of domains including motion understanding, optical-flow-based analysis, motion complexity, and the biological understanding of spider movements. These contributions are listed as follows:

1. Identifies a novel set of motion complexity features based on optical flow that encodes the various aspects of articulated motion complexity

2. Defines a measure for quantifying general motion complexity by integrating the motion features as a weighted sum based on feature contribution

3. Presents the results of a user study: an expert polling on statistical feature importance for visual motion complexity in spiders

4. Demonstrates the efficacy of the defined complexity measures in a real-world problem domain (the biological study of spider movement)

3.2 Temporality Features

In this section, we define the local (per frame) and global (per video) temporality features for articulated motion complexity (shown as step 2 and step 3 in Figure 3.1). The conversion of the local temporality features to global temporality features is also addressed. All of the local features and global features were chosen in order to address what we believe to be the foundational building blocks of complexity. These complexity building blocks (complexity domains) are listed in Table 3.2, along with the utilized corresponding complexity scale that states our belief in how the domain affects the final complexity value. For example, we believe that quicker movements, more areas of movement, more changes in direction, and periodic (repeating) movements are more complex than slower movements, fewer areas of movement, fewer changes in direction, and non-periodic movement. It is important to note that these measures are not provided as fact, but are what our beliefs and opinions indicate are important. The same applies to what constitutes more complexity versus less complexity for each measure.

Complexity Domain	**Corresponding Scale**
Movement coverage	More movement = more complex
Movement speed	Quicker movement = more complex
Movement coverage clusters	More clusters = more complex
Movement periodicity	Periodic movements = more complex
Movement entropy	More random = more complex
Directional smoothness	Sharper transitions = more complex
Directional changes	More changes in direction = more complex
Directional change frequency	Quicker changes = more complex

Table 3.2: Building blocks for defining motion complexity.

As described in formal detail in Chapter 2.1, an optical-flow algorithm is used to compute a series of flow images $O = [O^1, O^2, \ldots, O^{t-1}]$ in which $O^i_{x,y} = [u^i_{x,y}, v^i_{x,y}]$ is the set of optical flow vectors, O^i is the motion flow image between video frames F^i

and F^{i+1}, (x, y) is the spatial location of a pixel, and $[u, v]$ is a flow displacement vector (where u is the horizontal displacement and v is the vertical displacement). Two key components can be computed from each optical flow vector $[u, v]$: the flow *direction* and the flow *magnitude*. That is, the computed displacement vectors indicate both the direction of motion at each point of a frame of video in addition to the strength of motion (speed) in that direction. All features proposed in this chapter are computed from the magnitude-based and direction-based images derived from the optical flow values.

3.2.1 Local Temporality Features

We first discuss the *local temporality (per frame)* features built from the optical flow directions and magnitudes. For a given video, this set of features aims to capture the unique motion signatures for use in quantitatively describing the complexity of motion. As these features are measured as one value per feature per video, each local temporality feature is represented as one vector for each video. We distinguish the local temporality features as either being based on optical flow magnitudes or directions. The features are based on statistical measures that aim to numerically quantify each motion complexity domain listed in Table 3.2.

Magnitude-Based Features Motion strength, or *magnitude*, is a useful measure for determining the intensity of motion over time. It also allows for an estimation of motion speed, where larger magnitude values indicate faster motion. The set of flow magnitude images $M = \{M^1, M^2, \ldots, M^{t-1}\}$, where $M^i_{x,y}$ represents the magnitude at spatial location (x, y) between video frames F^i and F^{i+1}, is computed on the set of optical flow vectors O using the Euclidean distance formula:

$$M^i_{x,y} = \sqrt{(u^i_{x,y})^2 + (v^i_{x,y})^2} \tag{3.1}$$

where $[u^i_{x,y}, v^i_{x,y}]$ is the motion displacement vector at $O^i_{x,y}$. We will use the notation m^i to represent the set of magnitudes in flow magnitude frame M^i as a flattened (one-dimensional) vector with length l. This set of magnitude values m^i at a given point in time is used to compute several statistical features based on motion strength/speed. These statistical features are summarized as follows (providing a total of 48 magnitude-based local temporality features):

- Overall motion strength/speed in a single frame i:

  Mean – The average speed of motion: $mean(m^i)$

  Median – The median speed of motion: $median(m^i)$

  Max – The maximum speed of motion: $max(m^i)$

  Min – The minimum speed of motion: $min(m^i)$

- Motion strength/speed variability in a single frame i;

  Range – The range of motion speed: $max(m^i)$ - $min(m^i)$

  Variance – The degree of motion speed variance: $variance(m^i)$

  Skewness – The degree of motion speed skewness: $skewness(m^i)$

  Kurtosis – The degree of motion speed kurtosis: $kurtosis(m^i)$

  Entropy – A measure of randomness in motion speed: $-\sum_{x=1}^{l} P(m^i_x) log P(m^i_x)$

Each of these features was computed five times: once for all motion vectors, and once only for motion in the leftward, rightward, upward, and downward directions. If a vector was pointing both up and to the right, it was included in both the collection

of upward vectors and the collection of rightward vectors. The features were computed for each direction out of our belief in motion complexity information being hidden among a subset of the magnitude values. For example, it is possible for the average magnitude of all motion vectors in a frame to be small, while at the same time having the average magnitude of only the upward motion vectors measuring at a strong magnitude. In addition to these features, three other local temporality features were defined and computed to provide greater coverage for the various complexity domains listed in Table 3.2. These features are listed as follows:

Cluster Count – The number of areas showing movement in a frame

Cluster Size – The average size of the areas of movement in a frame

Movement Percentage – The percentage of a frame showing movement

For the two cluster-based features, a flow magnitude frame M^i is converted to binary frame B^i, where $B^i_{x,y} = 1$ if $M^i_{x,y} > \alpha$, and 0 otherwise. Here, α is a threshold value, where anything below the threshold is considered to be noise or non-movement. For our experiments, $\alpha = 0.02$. By increasing α, only the strongest motions are kept (possibly missing out on important smaller movements). By decreasing α, noisy and nonmoving areas are taken into consideration for the feature computations, giving inaccurate results. The binary frame B^i is then input into a connected-components labeling algorithm, where any pixels in B^i that have been assigned a '1' and border another pixel that has been assigned a '1' are given the same label. Each group of pixels with the same label is considered to be its own *cluster*. Thus, Cluster Count summarizes the number of areas showing movement in a given frame, while Cluster Size sums up the size (number of pixels) of each cluster in a frame, and divides by the

number of clusters. Movement Percentage is considered to be the number of '1' pixels in a binary frame B^i divided by the total number of pixels in B^i.

Directional Features In addition to magnitude, direction is also a useful feature that can indicate the orientation in which the most motion is made. Similar to magnitude, direction can be obtained from an optical flow vector $[u, v]$. A set of flow direction images $D = \{D^1, D^2, \ldots, D^{t-1}\}$, where $D^i_{x,y}$ represents the direction at spatial location (x, y) between video frames F^i and F^{i+1}, is computed on the set of optical flow vectors O using the following equation:

$$D^i_{x,y} = tan^{-1}(\frac{v^i_{x,y}}{u^i_{x,y}}) \tag{3.2}$$

where $D^i_{x,y}$ is expressed in radians. We will use the notation d^i to represent the set of directions in flow direction frame D^i as a flattened (one-dimensional) vector with length l. Simply using the radian values for computing statistics gives misleading results, since the directional values loop around in a circle from positive to negative. Because of this, all features obtained from the directional values are computed using techniques from circular statistics by representing the directional values as vectors. These techniques are described in mathematical detail in [8]. A more in-depth coverage can be found in Jammalamadaka & Sengupta [29].

We propose, in addition to the set of magnitude-based local temporality features, a corresponding set of direction-based local temporality features to quantify complexity information hidden in the directional values. The aim of this set of features is to again address the needed descriptors for the complexity domains listed in Table 3.2. These features are summarized as follows (for a total of 60 local temporality features when combined with the previously defined 48 features):

- Overall motion direction in a single frame i:

  Mean – The average direction of motion: $mean(d^i)$

  RVL – The resultant vector length of average direction: $magnitude(mean(d^i))$

  Up Movement – The total percentage of upward movement

  Down Movement – The total percentage of downward movement

  Left Movement – The total percentage of leftward movement

  Right Movement – The total percentage of rightward movement

- Motion direction variability in a single frame i:

  Variance – The motion direction variance

  Skewness – The motion direction skewness (the degree of motion being pulled toward a single direction)

  Kurtosis – The motion direction kurtosis (the degree of motion not being pulled equally in all directions)

- Motion direction distribution tests in a single frame i:

  Rayleigh Test Score – A test of how large the resultant vector length (RVL) must be to indicate a non-uniform distribution [20]

  Omnibus Test Score – An alternative to the Rayleigh test that works well for unimodal, bimodal and multimodal distributions [63]

  Rao Test Score – A spacing test for circular uniformity [7]

By utilizing a statistical description of the directional values, we aim to capture the motion signatures that contribute directly to the overall complexity of the observed motion. As with the magnitude-based features, these statistics are only computed on

the directional values that are not considered noise or non-movement (using threshold value $\alpha = 0.02$). The majority of these features specifically capture information about the distribution of the directional values. It is possible that several of these statistics (in addition to the magnitude-based measures) are quantifying similar information, and thus are redundant. During this stage, however, we focus only on proposing a large list of statistical-based features, while delaying the feature selection process (step 4 in Figure 3.1).

3.2.2 Global Temporality Features

We now turn to the generation of the global temporality features (step 3 in Figure 3.1). As we ultimately want to have each vector of local temporality features transformed into a single global value for use as a descriptor in the final measure, the local temporality features are converted to global temporality features by computing the means of each vector. This provides a general quantified measure of how each feature performed on an entire video clip. As we previously defined 60 local temporality features, this direct transformation also computes 60 global temporality features. In addition to this newly generated feature set, an additional three features are defined in order to provide additional statistical descriptions of the video as a whole and address more of the complexity domains in Table 3.2 not yet covered by other features for a stronger measure. The three additional global temporality features are defined as follows:

Run Count – The number of periods of movement

Run Average – The average length of the periods of movement

Directional Changes – The number of significant changes in direction over the duration of the video sequence

These new global features specifically quantify information regarding the periodicity and directional change domains in Table 3.2. Run Count is computed by taking the Average Magnitude local temporality feature vector, performing a one-dimensional connected component operation, and counting the number of connected components. This provides a measure of the number of movement "bursts". Run Average is also computed using the connected components, except the average number of frames involved in the bursts (cluster size) is used. Directional Changes is computed by taking the average direction vector v for each frame (the Mean Motion Direction local temporality vector), and counting the number of times v_x for frame F_x is greater than a 90 degree change from both v_{x-1} and v_{x-2} (the two frames appearing sequentially before it).

3.3 Feature Selection

In this section, we address the issue of feature selection (step 4 in Figure 3.1). Computing all 63 global temporality features based on magnitude and direction results in a high-dimensional feature space. As several features may be redundant by quantifying and contributing similar information, we utilize a multi-step feature selection process to identify any redundant features, as well as features that may not actually be useful toward quantifying articulated motion complexity. The steps taken in our feature selection process are as follows:

- Principal feature analysis
- Statistical t-test

- Correlation
- User study utilizing expert opinion

The first step utilizes *principal feature analysis (PFA)* [41] as a tool to reduce the large feature space to a smaller one. PFA is based on *principal component analysis (PCA)* [55], a dimensionality reduction technique that computes a new set of features (components) that are linear combinations of the original features. PFA expands on this idea by exploiting the structure of the principal components to choose the original features that retain most of the information. These principal features both contain maximum variability of the features and minimize the reconstruction error.

Several other techniques are used in conjunction with PFA to find the most important and non-redundant features. A statistical *t-test* is applied to reveal which of the original features are able to be used to determine if two datasets are significantly different from each other based on their means. This provides a measure of the classification abilities for each feature, a topic that is explored later in this chapter. Additionally, the statistical *correlations* of the features are computed to determine which features may be contributing similar information, allowing for the removal of any redundant features.

Finally, a *panel of experts* are polled to determine what they feel to be the most important contributors to complexity, as well as provide the confidence of their answers. The expert panel is important in that they have domain knowledge of what they feel contributes to movement complexity, thus allowing us to observe how their responses match up with what the previously described feature selection techniques reveal. Performing all four feature-selection steps computes a new, smaller set of more useful global temporality features.

3.4 Complexity Metric

With the final selected subset of global temporality features, we now turn to the creation of a measure for articulated motion complexity (the final step in Figure 3.1). To the best of our knowledge, there exists no standardized motion complexity measure for measuring articulated motion complexity, so a new measure is presented. We use the complexity domains listed and summarized in Table 3.2 as a foundation for defining complexity, where each measure contributes to the overall complexity measure. We specifically utilize a data fusion (weighted sum) approach. After the most important and non-redundant features have been selected, a weighted combination of the features is used as basis for the complexity measure. The weights for each feature are determined by the mean response of the expert panel during feature selection, but can be adjusted as desired depending on the application domain. In a case where a group of similar features exist (such as both the leftward movement percentage and rightward movement percentage), a single weight can be used for the group as a whole and divided equally among them.

The weighted sum allows for integrating all of the individual selected global temporality features into a single measure, while giving more weight to features that are believed to contribute more to motion complexity and less weight to those that may not have as large of a degree of influence. In order to adjust the complexity measure output to fall on a $0-1$ scale, the final weighted sum is divided by the total sum of the weights, and each feature is divided by the maximum value of the feature. On this scale, a '0' indicates the lowest possible complexity while a '1' indicates the highest possible complexity. Therefore, the final motion complexity measure is defined as

$$\frac{\sum_{i=1}^{n} weight_i \times \frac{feature_i}{max_i}}{\sum_{j=1}^{n} weight_j} \tag{3.3}$$

where n is the total number of selected features, i is the current feature, $weight_i$ is the weight value assigned to feature i, $feature_i$ is the value of feature i, and max_i is the maximum value of feature i observed over all of the videos (used to scale between $0 - 1$).

3.5 Implementation & Results

This section discusses the details of the dataset that was used in this work, as well as the implementation details of the proposed method in Figure 3.1. The results of using the proposed measure on the dataset for both predicting complexity values and classifying the video class are also observed. While we summarize the utilized dataset here, a more detailed description and overview of the spider dataset is provided in Appendix C.1.

3.5.1 Dataset

Our work utilizes a dataset of high frame rate videos containing samples of two species of *Schizocosa* wolf spider: *S. bilineata* and *S. crassipalpata*. There are 52 total grayscale videos in the dataset, where each video is roughly six seconds in length. The dataset is divided into two halves (one half for each species), while each of those halves is further divided into two (high diet and low diet). The separation of high diet from low diet comes from the biological expectation that nutrient intake could influence the degree to which spiders can engage in complex courtship displays. Thus, by varying the diet of individuals, we can assess whether there is a link between nutrient intake

and courtship complexity.

Each video has a temporal resolution of 250 FPS for capturing the extremely quick movements of the spiders, and a varying spatial resolution due to cropping out the areas of interest in each clip. When computing optical flow vectors for each frame, any vector with a very small magnitude ($\alpha < 0.02$) is discarded before computing the local statistical features to eliminate noise and ignore areas with no movement. Adjusting this threshold could cause significant changes in the final results, and should be selected carefully depending on the application domain.

Using an optical-flow coloring technique [4], a color can be assigned to each optical flow vector corresponding to the magnitude and direction. The color mapping chart for accomplishing this technique is shown in Figure 3.2. A sample frame from both species of spider along with corresponding colored movement frame using the technique from [4] is shown in Figure 3.3 and Figure 3.4. More detailed examples of the types of movements being displayed by these spiders can be seen in Appendix C.1 in Table C.2.

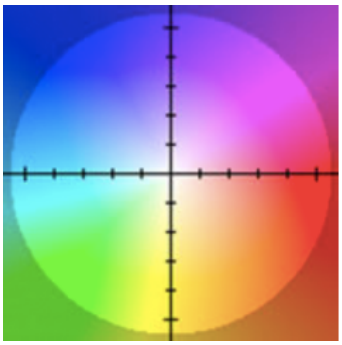


Figure 3.2: Mapping chart for colorizing optical flow vectors (image taken directly from [4]).

3.5.2 Feature Selection

After computing the 63 global temporality features (as discussed in Section 3.2), we perform feature selection on the spider dataset. Due to the completely different nature

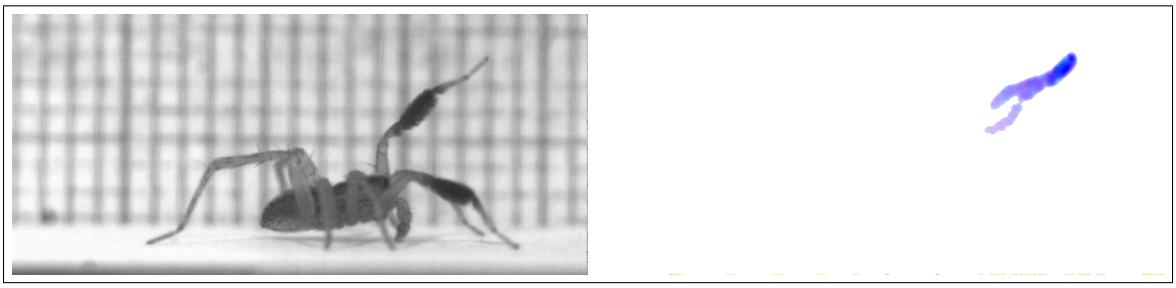


Figure 3.3: A sample of *S. bilineata* from the spider dataset (left) with corresponding colored optical flow (right) using the coloring technique described in [4].

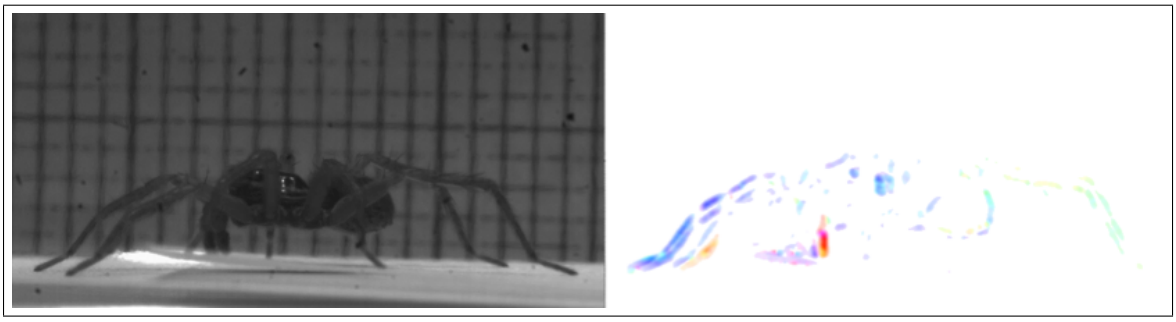


Figure 3.4: A sample of *S. crassipalpata* from the spider dataset (left) with corresponding colored optical flow (right) using the coloring technique described in [4].

of magnitude and direction, PFA was computed on the two sets of features separately. As PFA utilizes k-means clustering as part of the algorithm, the clustering can result in different clusterings based on its random initialization. To remedy this, PFA was computed 1000 times on each set to minimize the impact of different chosen features as a result of the random seeding of k-means.

Running PCA on the entire set of global directional features revealed that the data can be represented with over 99% accuracy using only 5 out of the 12 components. For this reason, the PFA algorithm was set to choose the 5 best features from the original 12 features (and thus k-means clustering was set to 5 clusters by the PFA algorithm). The 5 selected directional features were 1) *mean*, 2) *RVL*, 3) *skewness*, 4) *Rayleigh test score*, and 5) *right movement percentage*. To adjust for spiders facing opposite directions (symmetry), the *left movement percentage* was kept as well to give

the best 6 features.

Similarly, running PCA on the entire set of global magnitude-based features revealed that the data can be represented with over 99% accuracy using only 4 out of the 45 components. For this reason, the PFA algorithm was set to choose the 4 best features from the original set (and thus k-means clustering was set to 4 clusters by the PFA algorithm). The 4 selected magnitude-based features were 1) *kurtosis of rightward movement*, 2) *kurtosis of leftward movement*, 3) *kurtosis of upward movement*, and 4) *maximum*. Note that *kurtosis of downward movement* was not included with the other three directions, but increasing the desired number of magnitude-based features to 5 then selects this feature as well. By including this feature, this gives the best 10 selected features from the original 63, and reveals the importance of kurtosis as a descriptor.

From these features selected using PFA (in addition to a few other features we still hold belief in them contributing significantly to complexity), a panel of spider researchers were asked to provide a rating on how important they believed each of the features to be toward contributing to complexity (0=No effect, 1=Small effect, 2=Medium effect, 3=Large effect). They were also asked to provide for each feature whether they were confident of their response or not (0=Not confident, 1=Confident). If they indicated a high confidence value that a feature was important, we included the feature. Likewise, if they indicated a high confidence value that a feature was not important, we disregarded the feature. In addition, a t-test was performed on each of the features to test data set separability for three cases: a) species 1 vs. species 2, b) species 1 (high diet) vs. species 1 (low diet), and c) species 2 (high diet) vs. species 2 (low diet) where '0' indicated *not significant* and '1' indicated *significant*. A detailed summary of this expert complexity study is provided in Appendix A.

Finally, a correlation test between the features was observed to see if any selected

features were highly correlated (indicating that several features might be redundant). After performing PFA, polling spider researchers with domain knowledge, and observing the results of the t-test and correlations, the final set of selected features were chosen. These are presented in Table 3.3 along with the results of the expert polling and t-test experiments. A value of *N/A* indicates that the expert panel was not polled about that feature, as they were only polled on features chosen by the PFA algorithm. It is observed that skewness, kurtosis, and entropy are all useful measures for complexity. Human expert understanding stated that directional changes, number of clusters, movement runs, and Rayleigh test are expected to contribute greatly to the final measure, although their confidence was low for both movement runs and Rayleigh test. If a feature was shown to have a t-test value of '1' for at least two of the three scenarios (species 1 versus species 2, species 1 high diet versus species 1 low diet, and species 2 high diet versus species 2 low diet), then it was kept in the final feature set.

3.5.3 Complexity Metric

After the final subset of global temporality features were selected, a data fusion (weighted combination) of the features was used as the complexity measure function as described in Section 3.4. The weights for each feature were determined by the mean response of the expert panel, and rounded to an integer in the set $\{1, 2, 3\}$. In a case where a group of similar features existed (such as kurtosis of the magnitude values in each of the four directions), a single weight was given to the group and divided equally among them. Thus, instead of assigning a 1 to each of the directional kurtosis values, 0.25 was assigned. This prevented a large group of similar features from overshadowing the weights of non-grouped, individual features in the final calculation. A weighted sum approach allows for the contribution of a large set of features, while weighting

Feature	Expert mean (0-3)	Expert confidence (0-1)	T-test (1 vs. 2)	T-test (H vs. H)	T-test (L vs. L)
directionRVL	1.45	0.09	1	1	1
directionRTest	2.27	0.45	1	1	1
directionRightMagnitudeEntropy	N/A	N/A	0	1	1
directionDownMagnitudeEntropy	N/A	N/A	0	1	1
directionLeftMagnitudeEntropy	N/A	N/A	0	1	1
directionUpMagnitudeEntropy	N/A	N/A	0	1	0
directionRightMagnitudeSkewness	N/A	N/A	1	1	1
directionDownMagnitudeSkewness	N/A	N/A	1	1	1
directionLeftMagnitudeSkewness	N/A	N/A	1	1	1
directionUpMagnitudeSkewness	2	0.45	1	0	1
directionRightMagnitudeKurtosis	1.73	0.27	1	1	1
directionDownMagnitudeKurtosis	2	0.45	1	1	1
directionLeftMagnitudeKurtosis	1.73	0.27	1	1	1
directionUpMagnitudeKurtosis	N/A	N/A	1	0	1
magnitudeSkewness	N/A	N/A	1	1	1
magnitudeKurtosis	N/A	N/A	1	1	1
numberOfClusters	2.91	0.82	1	0	0
movementRuns	2.18	0.36	1	1	1
directionalChanges	2.45	0.82	1	0	1

Table 3.3: The final selected features with the t-test results and the expert panel's belief.

specific ones higher or lower in terms of its contribution to the final complexity value.

3.5.4 Complexity Results

After determining the feature values and weights, the complexity measure was used to compute a complexity value for each spider video (as shown in Figure 3.1). For each video, a motion complexity value between '0' (not complex) and '1' (complex) was computed. We observed how the complexity values compared between the two species as a whole, as well as between the high and low diets of the two species separately. Figure 3.5 shows the computed complexity values plotted for each of the four cases, split over four lines for easier visualization. Note that classifying the species using the measure is not the goal of this chapter, but is ultimately a desired effect of having

the measure. It can be seen in Figure 3.5 that *S. bilineata* (high diet) forms a tight cluster, with the exception of a few outliers. *S. bilineata* (low diet) follows a similar pattern. These outliers in the dataset do correspond to spiders that were more visually active, while the videos with the tightly clustered complexity values were mostly still except for an occasional leg movement. Thus, for these cases where outliers do occur, the spiders were indeed displaying more complex movements in the corresponding videos (such as displaying multiple movements while walking forward) that caused the complexity to spike. Therefore, this demonstrates that the measure is producing accurate values.

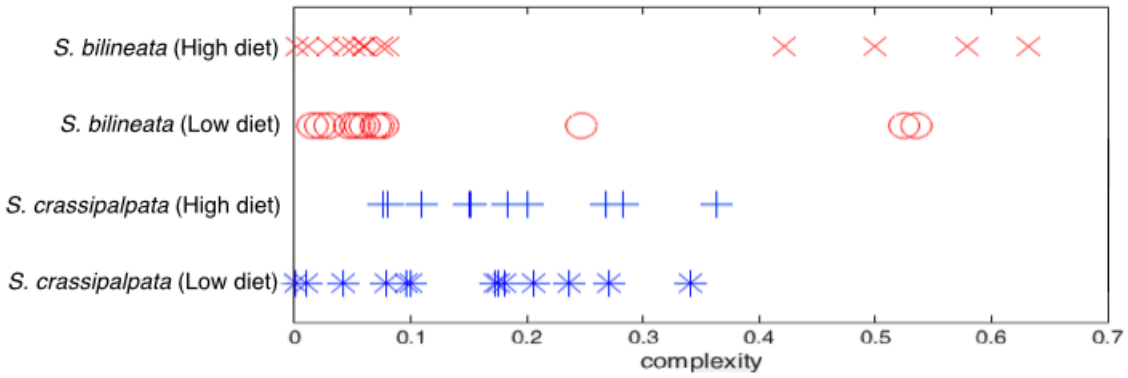


Figure 3.5: The final complexity values for the dataset (each data point corresponds to a single video in the dataset).

As outliers and overlapping values can make the data difficult to interpret in Figure 3.5, Figure 3.6 displays the same data points as normal distributions for the three comparative cases. As can be seen in each case, and as confirmed by a statistical t-test, the complexity values are not significantly different (that is, significantly separable by their means for classification purposes), which is mostly due to the outliers of *S. bilineata*. As a human observer, it is also not obvious that visual differences exist between the motion of high diet samples and low diet samples within a species. We do believe, however, that the complexity values accurately correspond to what is

displayed in each video from the dataset, providing an accurate measure.

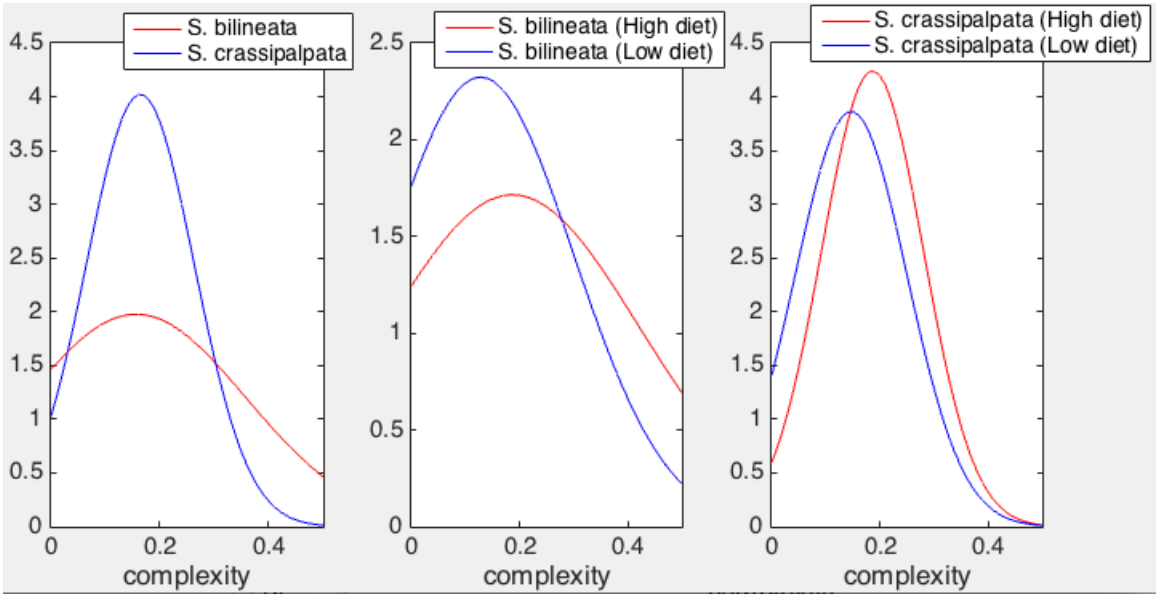


Figure 3.6: The normal distributions of the complexity values for a) *S. bilineata* vs. *S. crassipalpata*, b) *S. bilineata* (High diet) vs. *S. bilineata* (Low diet), and c) *S. crassipalpata* (High diet) vs. *S. crassipalpata* (Low diet).

Some interesting observations revealed in Figure 3.6 are 1) *S. bilineata* complexities deviate more greatly from the mean than *S. crassipalpata*, and 2) the high diets in each case have (on average) greater complexities than the low diets. A significantly larger dataset may give a better image of the separability (classification) abilities of the measure, but is again not the focus of this chapter. In addition, it is not completely understood if spider motion can be accurately differentiated by an automated numerical measure, and more work needs to be done in this area. In Chapter 4 and Chapter 5, we present two alternative approaches toward quantifying articulated motion complexity, and demonstrate the abilities of both in terms of both prediction of complexity scores and classification.

3.6 Summary

In this chapter, we have shown that optical flow-based features can form the basis of a measure for quantifying the complexity of visual articulated motion. The potential of these features was shown using a case study on *Schizocosa* wolf spider movement during courtship, and demonstrated that the measure can be used for describing general articulated motion. The computed complexity values not only provide a descriptive measure of motion, but demonstrate some capability for being used in the classification of species by their movement complexity. These results can further contribute to a better understanding of the behavior and communication of wolf spiders during their courtship routine, as well as other non-spider species. In addition, a user study on spider researchers was presented to obtain expert belief of which features contribute to motion complexity, and to what degree do the features contribute. Most importantly, this work provides the foundation for a general motion complexity measure that can benefit the community.

In Chapter 4, we improve on this optical flow technique by creating a new set of features from the computed flow estimation that take into account six identified motion-complexity domains believed to cover the various aspects of complexity. Two new domains that are explored include motion repetition (using techniques from Fourier analysis to extract the primary frequencies) and motion synchrony (measuring multiple areas of movement in a video frame that are moving at the same time). The efficacy of using the new features for measuring complexity is again demonstrated using a weighted-sum measure, but also trained linear-discriminant classifiers for distinguishing motion classes (classification) and predicting complexity scores for new motion samples. A sequential feature selection algorithm is utilized to identify the complexity features that contribute the most toward correctly predicting complexity scores and accurately

classifying motions. In addition, a user study on motion complexity is presented for demonstrating participant belief of complexity domains and for use in training the classifiers.

In Chapter 5, we abandon optical flow for spatial-temporal measures as the basis for motion complexity features. Spatial-temporal features integrate both space and time to determine where interesting and significant motion is happening within the video volume. The efficacy of using the new features for measuring complexity is demonstrated using trained linear-discriminant classifiers for distinguishing motion classes and predicting complexity scores for new motion samples. A sequential feature selection algorithm is utilized to identify the complexity features that contribute the most toward correctly predicting complexity scores and accurately classifying motions.

Chapter 4

Prediction and Classification Using an Optical Flow-Based Complexity Metric

In this chapter, we present two scenarios for analyzing visual articulated motion complexity: motion complexity score prediction and motion classification. The problem of predicting motion complexity scores is addressed using a data-fusion (weighted-sum) approach based on both human belief and empirical evaluation, as well as using a pattern-recognition (linear discriminant analysis) approach. The problem of classifying motion classes is addressed using a pattern-recognition (linear discriminant analysis) approach. Other than prediction and classification, uses for such a complexity measure include video indexing, motion comparison, and the biological study of species. A user study of motion complexity is presented, as well as a novel set of motion complexity features that are computed from optical flow vectors and used in the training of the complexity models. The accuracy of these models is demonstrated on both a dataset of human actions and a dataset of high-frame-rate wolf spider movements. We show

that this proposed set of motion complexity features is useful toward the creation of a complexity measure for general articulated motion in video, and has significant classification abilities. The work in this chapter is planned for publication in [15].

4.1 Introduction

Motion analysis is an important component of many computer vision systems and application domains. In medical systems, motion signatures can be used to track rehabilitation progress in order to assist with a quicker recovery. In visual surveillance systems, motion patterns can be learned in order to signal when abnormal motion signatures are detected, possibly indicating a current or upcoming security threat. Motion, however, can be characterized in several different ways. For example, there are shorter motion patterns (such as kicking or waving) as well as longer "tracked" motion patterns (such as following a person through a crowd in a surveillance video). Similarly, different motion patterns can have different levels of complexity. For example, the motion of a person walking is less complex than a person performing a challenging juggling routine. There has been, however, little work done on finding a general measure for quantifying the complexity of visual motion. Having an established measure would be useful for many tasks such as motion classification, motion comparison, video indexing based off of complexity values, and advanced biological study of the visual communication of species.

The goal of this chapter is to present a detailed study on the quantification of visual motion complexity. Specifically, we propose a novel set of motion complexity features that are used for both predicting complexity scores for videos and classifying the motion from a video. While classification is not the goal of having a complexity measure, it is a desired effect. One possible issue with using a complexity measure for

classification is that two separate motions could have the same complexity value, thus making the process of distinguishing between them in this way impossible. However, as explored in this chapter, classification abilities are desired for instances such as the biological study of spider species. For example, we explore the potential of using the complexity features for the classification of spiders within a species with high nutrient intake versus those with low nutrient intake.

The problem of predicting motion complexity scores is addressed using both a data-fusion (weighted-sum) approach and a pattern-recognition approach based on linear discriminant analysis. The problem of predicting motion classes is addressed using a pattern-recognition (linear discriminant analysis) approach. Both the prediction model and the classification model are trained by utilizing a user study on visual motion complexity. The results of the user study are presented in this chapter, and in further detail in Appendix B.

We specifically focus our efforts on articulated movement from a still camera and a single subject. Our belief is that features should be chosen from several different complexity domains that each contribute to the overall concept of visual motion complexity instead of a single domain (such as motion intensity). These domains are summarized in Table 4.1.

Motion Domain	**Description**
Movement amount	Spatial coverage of movement
Movement speed	Speed of movement
Movement periodicity	Repetition of movement
Movement synchronization	Multiple parts moving simultaneously
Directional changes	Degree of changes in direction
Number of moving parts	Number of moving areas

Table 4.1: Set of motion complexity domains.

Several concepts from Chapter 3 are used again in this chapter. The same

Horn-Schunck algorithm is used with no change in parameters. In addition, some of the features remain similar (statistical measures of the optical flow). The main differences in the features are the addition of statistical measures of both the horizontal displacements and vertical displacements of the optical flow separately, and the addition of higher-order features (motion synchrony, motion frequency analysis, and additional clustering statistics). In addition, the motion complexity domains have been adjusted to reflect the lessons learned from Chapter 3.

4.1.1 Problem Definitions

This work presents an analysis of visual motion complexity by exploring the prediction of motion complexity scores, as well as distinguishing motion classes (classification). Here we provide formal definitions for each of these problems scenarios.

Complexity Score Prediction The problem of predicting motion complexity scores is defined as follows: Given a set of videos $\mathbf{V}$ where each video $V = [F^1, F^2, \ldots, F^t]$, $F^i_{x,y}$ is the pixel in the x^{th} row and y^{th} column of the i^{th} image frame, and t is the total number of frames in the video, our goal is to create a complexity model C that takes a motion sequence of images (video) as input and generates a value between 1 (lowest possible complexity) and 10 (highest possible complexity) as output. Thus, we aim to find or train a function $C : V \rightarrow [1, 10]$, where $[1, 10]$ is the set of integers between 1 and 10, inclusive. The scale was changed from the $[0, 1]$ scale used in Chapter 3 to compensate for the user study presented in Section 4.3. The values, however, can be scaled to any other range of values depending on the application domain.

Motion Class Classification The problem of classifying the motion class of a video is defined as follows: Given a set of videos $\mathbf{V}$ defined as above, our goal is to

train a classification model C using feature set M and assigned motion classes taken from the set L for the purposes of classifying unknown motion instances. Formally, we aim to train a function $C : V \rightarrow \{L_1, L_2, \ldots, L_n\}$ using a set of motion complexity features and videos from a dataset that, given a new video as input, predicts a motion class from L. Here, n is the number of motion classes that could be assigned.

4.1.2 Approaches

The proposed set of motion complexity features is computed from optical flow. A recent survey of optical flow can be found in [21], which organizes current approaches and practices. Two of the most popular optical flow algorithms are the Horn-Schunck algorithm [26] and the Lucas-Kanade algorithm [42], with a comparison of the two given in [6]. We specifically apply the Horn-Schunck algorithm, categorized in [21] as a regularization model that utilizes a spatial flow gradient constraint. While any optical flow technique could be used to compute the features, the Horn-Schunck approach was chosen because 1) the speed of computation is not a concern to us, and 2) the majority of the species-analysis literature utilizes that algorithm and we desire to use an algorithm already understood by that community. Thus, the same optical flow algorithm applied in Chapter 3 is used here with no changes.

As visual motion complexity has no standard definition, our approach relies on a user study where the complexity values for videos are obtained from a group of participants based on human opinion. This user study on visual motion complexity is introduced in Section 4.3 and detailed further in Appendix C. A set of motion complexity features are then computed for each video, which are defined in Section 4.2. We utilize a sequential feature-selection algorithm to choose the subset of features that give the best accuracy in terms of prediction and classification, and compare

a scenario using the best features against a scenario using all of the features. The selected motion complexity features are 1) integrated using a data fusion technique to create a weighted-sum complexity prediction model, 2) combined with the user supplied complexity ratings to train a discriminant analysis classifier for prediction, and 3) combined with the motion classes to train a discriminant analysis classifier for classification. A visual overview of our approach is shown in Figure 4.1.

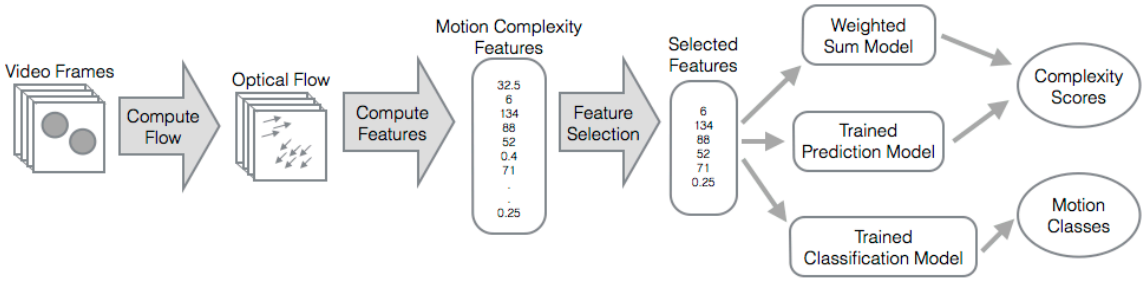


Figure 4.1: Overview of the three approaches.

Complexity Score Prediction (Data Fusion) We propose an approach for motion complexity prediction that utilizes a fusion of the motion complexity features. Specifically, we create a weighted-combination model C defined as

$$C = \frac{\sum_{i=1}^{n} weight_i \times \frac{F_i}{max_i}}{\sum_{j=1}^{n} weight_j} \tag{4.1}$$

where n is the total number of selected features, i is the current feature, $weight_i$ is the weight value assigned to feature i, F_i is the value of feature i, and max_i is the maximum value of feature i observed over all of the videos (used to scale between 0-1). The set of weight values are generated for each feature in F based on human belief from a user study on motion complexity. The performance of this model is compared against the same model where the weights are determined by a computer. The accuracy is determined by comparing the model against human-assigned complexity scores for

each video in **V**. We also investigate an alternate scenario where the weights are assigned by a computer algorithm that attempts to find a set of weights maximizing the accuracy. The accuracy of this empirical-based model is compared against the human-belief model.

Complexity Score Prediction (Linear Discriminant Analysis) We propose an alternative approach for motion complexity prediction that utilizes a pattern-recognition-based technique. We train a classification model using the human-assigned complexity scores from the user study. Specifically, we use linear discriminant analysis as the learning algorithm for the classifier. This technique was chosen using empirical testing among several classification algorithms including decision trees, clustering techniques, and support vector machines (SVMs). This trained classifier attempts to correctly predict the complexity score of an unseen motion class.

It is worth noting that the prediction problem is being treated as a classification problem. That is, instead of training a regression-based function, the prediction values are rounded to the nearest integer and used in a trained classification model. While the ultimate goal that this work progresses toward is a specific, real-number-based score, many application domains only require a higher-level categorization of the complexity scores. For example, many applications may only require the knowledge of whether the computed score is low complexity, medium complexity, or high complexity. Other domains may only need the complexity score on a scale of one to ten. This work presents the categorized version of the problem that can ultimately lead to a regression-based analysis in future work.

Motion Class Classification (Linear Discriminant Analysis) Instead of predicting complexity scores, our third approach toward visual motion analysis attempts

to correctly classify the motion class of the video. For example, it will attempt to distinguish between a walking motion video and a running motion video. Similar to our approach for complexity score prediction using a pattern-recognition-based approach, we use the same approach here. That is, we train a linear discriminant classifier using the motion complexity features and the motion classes. The classifier attempts to correctly determine the motion class of an unseen video.

4.1.3 Contributions

The goal of this work is to investigate both a data-fusion and a pattern-recognition approach towards the creation of a model for predicting the complexity scores of articulated motion in video as well as classifying the motion class using a novel set of *motion complexity features*. Such a model could be useful as a measure in a number of applications such as providing a numerical value that can be integrated in the understanding of various species (such as wolf spiders), indexing motion/activity videos, or classifying a wide range of movements (such as one dance routine from another). The overall contributions of this chapter are as follows:

1. Summarizes the results of a user study on visual motion complexity
2. Presents a novel set of motion complexity features based on optical flow for use in analyzing complexity in general articulated motion
3. Demonstrates the performance of a data-fusion (weighted sum) model for predicting motion complexity scores
4. Demonstrates the performance of a pattern-recognition (linear discriminant analysis) model for predicting articulated motion complexity scores and classifying video motion classes

4.2 Motion Complexity Features

In this section, the motion complexity features are defined that are the foundation of the methods used in the rest of the chapter. These features are computed directly from the optical flow output, where only the moving pixels (with sufficient magnitude) in each frame are considered in the computation. Optical flow computes a flow frame O_i between every two pairs of sequential video frames F_i and F_{i+1}. Each pixel in a flow frame is a vector $[u\ v]$ where u is the horizontal displacement and v is the vertical displacement of the pixel under consideration.

We specifically compute the motion complexity features using the flow magnitudes and the flow directions. Motion strength, or *magnitude*, is a useful measure for determining the intensity of motion over time. It also allows for an estimation of motion speed, where larger magnitude values indicate faster motion. The set of flow magnitude images $M = \{M^1, M^2, \ldots, M^{t-1}\}$, where $M^i_{x,y}$ represents the magnitude at spatial location (x, y) between video frames F^i and F^{i+1}, is computed on the set of optical flow vectors O using the Euclidean distance formula:

$$M^i_{x,y} = \sqrt{(u^i_{x,y})^2 + (v^i_{x,y})^2} \tag{4.2}$$

where $[u^i_{x,y}, v^i_{x,y}]$ is the motion displacement vector at $O^i_{x,y}$. We will use the notation m^i to represent the set of magnitudes in flow magnitude frame M^i as a flattened (one-dimensional) vector with length l. Direction indicates the orientation in which the most motion is made. Similar to magnitude, direction can be obtained from an optical flow vector $[u, v]$. A set of flow direction images $D = \{D^1, D^2, \ldots, D^{t-1}\}$, where $D^i_{x,y}$ represents the direction at spatial location (x, y) between video frames F^i and F^{i+1}, is computed on the set of optical flow vectors O using the following equation:

$$D^i_{x,y} = tan^{-1}(\frac{v^i_{x,y}}{u^i_{x,y}}) \tag{4.3}$$

where $D^i_{x,y}$ is expressed in radians. We will use the notation d^i to represent the set of directions in flow direction frame D^i as a flattened (one-dimensional) vector with length l. For features that are computed as one value per video frame, these were averaged over all frames. Only optical flow vectors over a given threshold were used in the calculations, while the rest were disregarded as areas of non-movement. The motion complexity features are defined as follows for a given frame i:

AverageMagnitude – Average flow magnitude per frame: $mean(m^i)$

AverageU – Average horizontal flow magnitude per frame: $mean(u^i)$

AverageV – Average vertical flow magnitude per frame: $mean(v^i)$

MaximumMagnitude – Maximum flow magnitude per frame: $max(m^i)$

MaximumU – Maximum horizontal flow magnitude per frame: $max(u^i)$

MaximumV – Maximum vertical flow magnitude per frame: $max(v^i)$

MedianMagnitude – Median flow magnitude per frame: $median(m^i)$

MedianU – Median horizontal flow magnitude per frame: $median(u^i)$

MedianV – Median vertical flow magnitude per frame: $median(v^i)$

EntropyMagnitude – A measure of randomness in motion speed: $-\sum_{x=1}^{l} P(m^i_x)logP(m^i_x)$

EntropyU – A measure of randomness in motion speed: $-\sum_{x=1}^{l} P(u^i_x)logP(u^i_x)$

EntropyV – A measure of randomness in motion speed: $-\sum_{x=1}^{l} P(v^i_x)logP(v^i_x)$

KurtosisMagnitude – A measure of the peakedness of the magnitude distribution: $kurtosis(m^i)$

KurtosisU – A measure of the peakedness of the magnitude distribution: $kurtosis(u^i)$

KurtosisV – A measure of the peakedness of the magnitude distribution: $kurtosis(v^i)$

NumberOfClusters – Number of areas of movement (connected component clusters) in a frame

DirectionalChanges – Number of times the average direction of the moving pixels ($mean(d^i)$ in a frame significantly changes (more than 45 degrees) from one frame to the next over the course of the video, normalized by frame count t

DominantFrequency – Dominant frequency (in hertz) of the average magnitudes ($AverageMagnitude$) for an entire video, computed by taking the discrete Fourier transform and selecting the largest frequency response

DominantFrequencyStrength – The response value of $DominantFrequency$

MovementSynchrony – Number of frames with more than one area of movement divided by the number of frames with at least one area of movement, where the areas of movement are considered to be connected component clusters

AverageClusterSize – Average size of the movement areas (connected component clusters) per frame (in pixels)

The features are chosen to represent all six motion complexity domains, as presented in Table 4.1. Each feature is mapped into exactly one of the six motion complexity domains. This helps to identify which features work together to determine what makes a motion complex. In addition, a weight is assigned to each motion complexity

domain, determined during the user complexity study (as summarized in Section 4.3) by summing the rating scores in each domain, and normalizing by dividing by the total rating sum. The weights for individual features are assigned by dividing the associated domain weight equally among them. The mappings and assigned weights are shown in Table 4.2.

Motion Domain	**Weight**	**Included Features**	**Weight**
Movement amount	0.19	AverageClusterSize	0.19
Movement speed	0.15	AverageMagnitude	0.01
		AverageU	0.01
		AverageV	0.01
		MaximumMagnitude	0.01
		MaximumU	0.01
		MaximumV	0.01
		MedianMagnitude	0.01
		MedianU	0.01
		MedianV	0.01
		EntropyMagnitude	0.01
		EntropyU	0.01
		EntropyV	0.01
		KurtosisMagnitude	0.01
		KurtosisU	0.01
		KurtosisV	0.01
Movement periodicity	0.15	DominantFrequency	0.075
		DominantFrequencyStrength	0.075
Movement synchronization	0.16	MovementSynchrony	0.16
Directional changes	0.21	DirectionalChanges	0.21
Number of moving parts	0.14	NumberOfClusters	0.14

Table 4.2: Motion complexity features mapped into their respective domains with associated weights.

4.3 User Study On Complexity

In this section, a user study on motion complexity is presented to determine what makes a motion complex in terms of human belief. This study is further detailed

in Appendix B. In addition, the datasets that the user study utilizes are detailed, which are also used in the rest of this chapter. The datasets are further described in Appendix C.

4.3.1 Datasets

Basic human actions The human action dataset used in this chapter is the Weizmann dataset [23], a widely used collection of basic human motions for comparing action classification systems. It contains 81 low-resolution (180×144) video sequences, recorded at 25 FPS, displaying nine different people performing nine basic actions. The displayed action classes are "running (run)", "walking (walk)", "jumping jack (jack)", "jumping forward on one leg (skip)", "jumping in place on two legs (jump)", "galloping sideways (side)", "waving one hand (1wave)", "waving two hands (2wave)", and "bending (bend)". An example of the "jumping jack" action can be be seen in Figure 4.2 with significant motion colorized according to the technique used by Baker et al. [4].

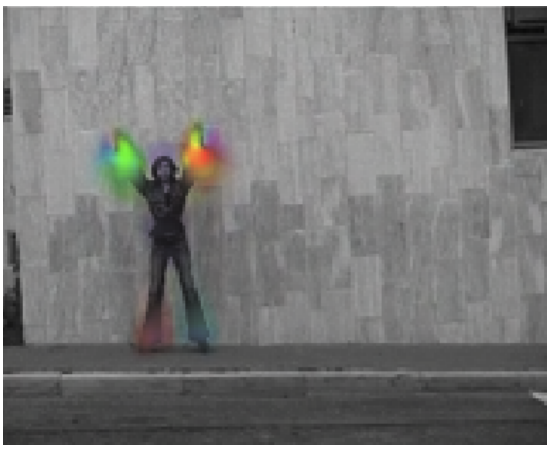


Figure 4.2: Example of a jumping-jack action in the human database, with optical flow field colorized [4] to indicate motion.

Spider courtship movements The second dataset contains high frame rate videos of samples of two species of *Schizocosa* wolf spider: *S. bilineata* and *S. crassipalpata*. There are 52 total grayscale videos in the dataset, where each video is roughly six seconds in length. The dataset is divided into two halves (one half for each species), while each of those halves is further divided into two (high diet and low diet). The separation of high diet from low diet comes from the expectation that nutrient intake could influence the degree to which spiders can engage in complex courtship displays. Thus, by varying the diet of individuals, we can assess whether there is a link between nutrient intake and courtship complexity. Each video has a temporal resolution of 250 FPS for capturing the quick movements of the spiders, and a varying spatial resolution due to cropping out the areas of interest in each clip. When computing optical flow vectors for each frame, any vector with a very small magnitude (< 0.02) is discarded before computing the motion complexity features to eliminate noise and areas of non-movement. Adjusting this threshold could cause significant changes in the final results, and should be carefully selected depending on the application domain. An example of *S. crassipalpata (high diet)* can be be seen in Figure 4.3 with significant motion colorized according to the technique used by Baker et al. [4].

Figure 4.3: Example of *S. Crassipalpata* in the spider database, with optical flow field colorized [4] to indicate motion.

4.3.2 User Study On Complexity

A user study was conducted on 24 participants from varying backgrounds to investigate what (based on user belief) makes a motion complex. For each participant, a series of videos was shown from the datasets described in Section 4.3.1. Each clip was played repeatedly while waiting for the user to rate the displayed motion on a scale of one (low complexity) to ten (high complexity). 25% of the displayed videos were randomly chosen to be shown twice to measure a rater's consistency. A one-way ANOVA model was used to determine which raters were being sufficiently consistent, which lead to one user being thrown out from use in future computations. A range test was also performed to ensure that the range of ratings a user gave were larger than four. One user gave all ratings between one and three, and was also thrown out from future computations. A summary of the ratings given to humans for the nine motion classes is shown in Figure 4.4, while the summary of the ratings given to spiders for the four cases is shown in Figure 4.5. It is interesting to note that while the spider scores were nearly identical for each of the two species, the low diet samples received a few more votes towards being more complex than the high diet samples.

In addition, each user was asked to rate on a scale of one (not important) to five (important) the degree of influence they believed each of the six motion complexity domains presented in Table 4.1 to have in terms of contributing to the overall complexity value. These ratings are used as the basis for determining the feature weights in the weighted-sum prediction model. A summary of the ratings given to the six motion complexity domains is shown in Figure 4.6. It is shown that users strongly believe that the number of moving parts and the amount of movement were the most important contributors to complexity, while the least important were repeating movement (periodicity) and synchronized movement. The complexity domain with

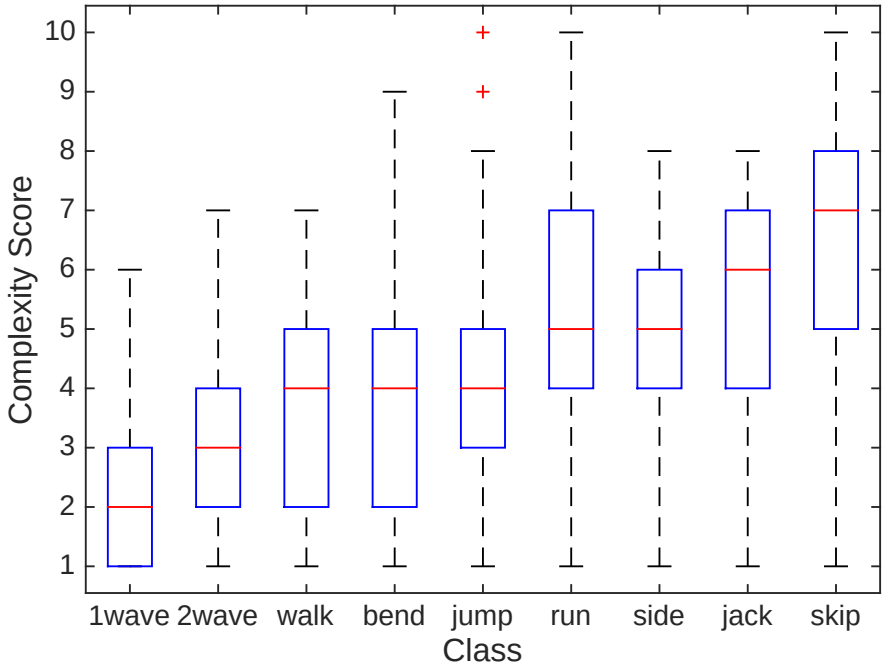


Figure 4.4: Overview of user ratings for the nine human motions.

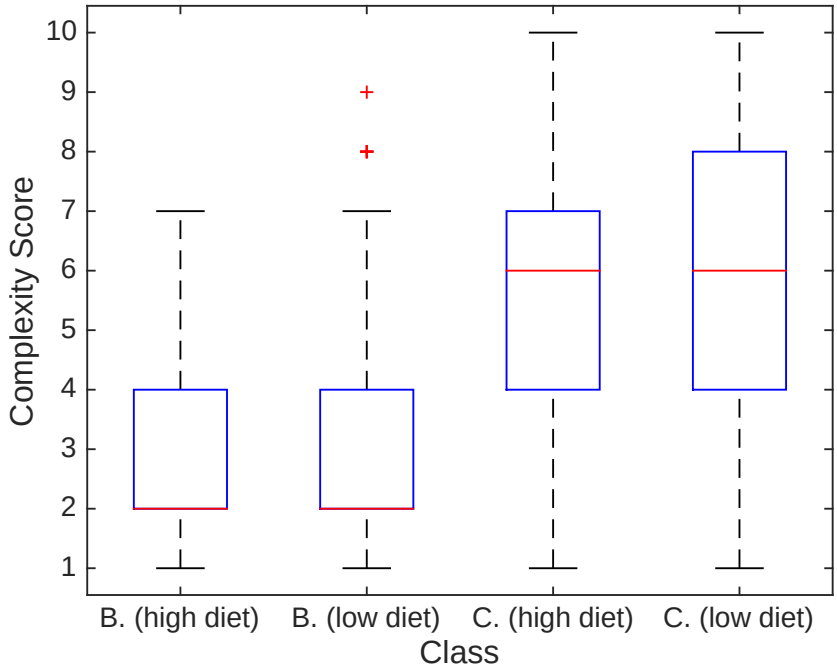


Figure 4.5: Overview of user ratings for the spider movements of *S. bilineata* (B) and *S. crassipalpata* (C).

the most disagreement among the users was the directional changes domain, while the domain with the most agreement was revealed to be the number of moving parts.

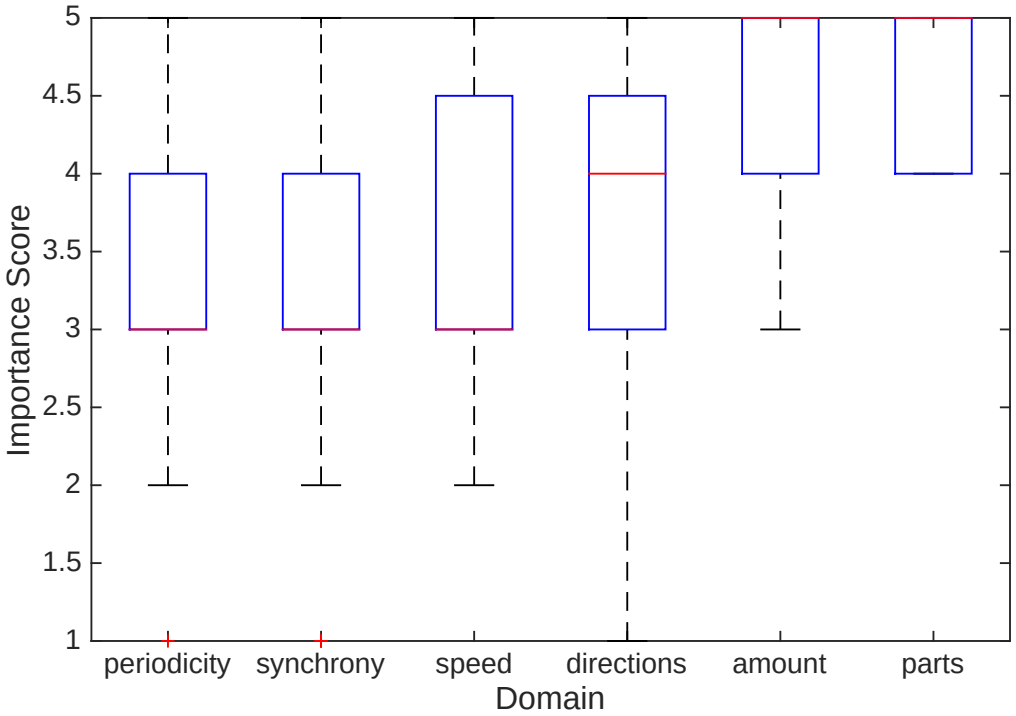


Figure 4.6: Overview of user scores for complexity domain influence.

4.4 Implementation & Results

Here we present the implementation of the three problem scenarios, along with performance accuracy results. All three approaches rely on the motion complexity features defined in Section 4.2, as well as the ratings and beliefs from the user study in Section 4.3.

4.4.1 Data Fusion Approach

Using the assigned feature weights in Table 4.2 (obtained from the user ratings of the complexity domains), a data-fusion technique was used to determine the accuracy of

a model based on human belief of complexity importance. The motion complexity features were computed for both the human and spider datasets, then used with the associated weights as a weighted-sum model. The predicted scores were then computed using the weighted model, and compared to the average ratings that the user-study participants assigned to each video. The "allowed range" was varied to demonstrate the difference between only allowing an exact complexity prediction (range 0), allowing the complexity score to be off by one in either direciton (range 1), etc. The results (accuracy and correlation of the predicted scores against the user provided scores) of this experiment are presented in Table 4.3.

To compare against the "participant belief" weighting approach, a computer program was executed to continually randomly generate (over a period of ten minutes) a set of weights that summed to one, where the set of weights with the best accuracy was used. That is, the program locates the best possible set of weights in the allotted amount of time. This is reported in Table 4.3 as "empirical weighting". To compare against the participant belief weighting set of {0.19, 0.15, 0.15, 0.16, 0.21, 0.14} from Table 4.3 used for both humans and spiders, the randomization technique recorded the best combination of human weights as {0.47, 0.02, 0.18, 0.09, 0.18, 0.07} and spider weights as {0.24, 0.21, 0.05, 0.12, 0.33, 0.04}. The top three domains for influencing complexity in humans are movement amount, movement periodicity, and directional changes, while the top three for spiders are movement amount, movement speed, and directional changes. Thus, human belief matches the randomization technique on movement amount and directional changes when choosing the top three, but misses on the importance of movement synchronization.

As shown in Table 4.3, the empirical-based weighting scheme chose feature weights that performed significantly better, indicating that human belief may not be as reliable in terms of weighting the complexity domains and, correspondingly, the features.

Dataset	Weighting	Allowed Range	Correlation	Accuracy
Human	Participants	0	0.06	3%
		1	0.34	28%
		2	0.34	67%
		3	0.59	83%
	Empirical	0	0.21	47%
		1	0.63	89%
		2	1.00	100%
Spider	Participants	0	0.46	38%
		1	0.57	80%
		2	0.85	96%
		3	1.00	100%
	Empirical	0	0.39	52%
		1	0.77	92%
		2	1.00	100%

Table 4.3: Accuracy of the data fusion approach.

Allowing the complexity prediction to be off by one (allowed range = 1), which may be acceptable depending on the application domain, showed accuracy improvements by up to 42%. Setting the allowed range to two again revealed a significant increase in accuracy, with the empirical-based weights showing 100% accuracy.

4.4.2 Pattern Recognition Approach (Predicting Complexity Scores)

Instead of relying on human belief to determine which areas of complexity are most important, a pattern recognition approach was used to learn the important features for each dataset. Specifically, a linear discriminant classifier was trained on the motion complexity features using the average human complexity scores (rounded to the nearest integer) as the training labels. The training/testing split used was $\frac{2}{3}/\frac{1}{3}$. To compensate for the random selection of the training and testing sets, each classifier was trained and tested 1000 times, with the average classification score used for the classifier accuracy.

Several tests were performed to observe how the accuracy changes. For both the human and the spider datasets, the training score was first set to be the average human complexity score for each individual video. The second case set the training score to be average human complexity score for the given video class. For example, if the average "bend" action for the human videos was a score of 0.3, then every "bend" video was assigned a 0.3. The number of features used was also varied from using all of the features to only using the "best" (top) features determined by a sequential feature selection algorithm. The top features for the human dataset were MaximumU, EntropyU, KurtosisMagnitude, NumberOfClusters, MovementSynchrony, and MeanClusterSize, while the best features for the spider dataset were EntropyMagnitude, EntropyU, KurtosisMagnitude, KurtosisU, NumberOfClusters, and DominantFrequency. Thus, NumberOfClusters matches the belief of the user study participants as being important for contributing to complexity. These results are shown in Table 4.4. As can be seen, using only the top features increases the accuracy in most cases. Increasing the allowed range to 1 or 2, which can be acceptable in some domains, reveals significantly greater prediction accuracy. In addition, using the mean class scores instead of the individual video scores yielded significantly better accuracy.

4.4.3 Pattern Recognition Approach (Classifying Motion Classes)

The previous classifier-based approach was used to predict complexity scores. Here, another classifier using linear discriminant analysis is trained, but instead used to learn and classify video motion classes. That is, instead of learning and predicting scores for a bend video using motion complexity features, it attempts to learn and classify a video motion as "bend". We reiterate that classification is not a goal of the metric, but

Dataset	Training Score	Features	Allowed Range	Accuracy
Human	Video Score	All	0	30%
			1	62%
			2	86%
		Top	0	46%
			1	77%
			2	97%
	Mean Class Score	All	0	70%
			1	80%
			2	93%
		Top	0	66%
			1	81%
			2	91%
Spider	Video Score	All	0	30%
			1	76%
			2	93%
		Top	0	43%
			1	88%
			2	99%
	Mean Class Score	All	0	66%
			1	87%
			2	92%
		Top	0	73%
			1	93%
			2	96%

Table 4.4: Accuracy of the discriminant analysis approach for predicting complexity scores.

a desired effect. Here, the classifier is trained for five different scenarios: 1) classifying human actions, 2) classifying spiders into the original four classes, 3) classifying as either spider species one or species two, 4) classifying between high diet and low diet for species one, and 5) classifying between high diet and low diet for species two. Using sequential feature selection, the top features for each of the five scenarios, respectively, are 1) KurtosisV, NumberOfClusters, DominantFrequencyStrength, and MeanClusterSize, 2) EntropyU, KurtosisV, NumberOfClusters, and DominantFrequency, 3) EntropyMagnitude and NumberOfClusters, 4) MedianV, DirectionalChanges, and

MovementSynchrony, and 5) AverageMagnitude, KurtosisMagnitude, and DominantFrequency. This reveals that important motion complexity features for classification includes NumberOfClusters, as well those regarding the dominant frequency. The results are shown in Table 4.5. For classification, it can be seen that using only the top features produces mixed results, and actually causes a significant drop in accuracy for several cases. While the classifier does well at classifying one spider species from another, it struggles to correctly classify between high diet and low diet. This matches the human complexity belief between high and low diet (that is, humans cannot distinguish between the two cases), which is shown in Table 4.5.

Dataset	**Label Domain**	**Features**	**Accuracy**
Human	Nine human actions	All	66%
		Top	61%
Spider	$\{B_H, B_L, C_H, C_L\}$	All	44%
		Top	42%
Spider	$\{B, C\}$	All	87%
		Top	92%
Spider	$\{B_H, B_L\}$	All	61%
		Top	42%
Spider	$\{C_H, C_L\}$	All	47%
		Top	50%

Table 4.5: Accuracy of the discriminant analysis approach for classifying motion classes.

4.5 Summary

We have presented an in-depth study of visual motion complexity by proposing a novel set of motion complexity features for both prediction and classification. Based on a user study of visual motion complexity, these features were used toward the creation of a weighted-sum model of complexity scores for a dataset of human actions and a dataset of spider movements. In addition, linear discriminant analysis was used

to train classifiers for the purpose of both predicting complexity scores as well as classifying motion classes. The complexity features were shown to be effective in many cases for correctly classifying motion classes as well as predicting complexity scores, notably when increasing the allowed error range to 1 or 2. It was also shown that using a set of "best" features leads to increased accuracy for predicting complexity scores, but produces mixed results for classification. Even greater accuracy gains were made when using the mean class score instead of the individual video scores.

This chapter also revealed interesting results about specific motion complexity features that contribute the most to the overall complexity value. Specifically, it was observed that features involving the number of areas of movement, the kurtosis of motion strength, and the dominant frequency were identified to be the most useful features both by human participant belief and the feature selection algorithm. These features, we believe, hold the most useful information about the motion signatures of visual complexity.

In Chapter 5, we abandon optical flow for spatial-temporal measures as the basis for motion complexity features. Spatial-temporal features integrate both space and time to determine where interesting and significant motion is happening within the video volume. It is our hope that interesting motion complexity information is hidden in the space-time domain, and that the utilization of space-time interest points will identify those hidden signatures. The efficacy of using the new features for measuring complexity is demonstrated again using trained linear-discriminant classifiers for distinguishing motion classes and predicting complexity scores for new motion samples. A sequential feature selection algorithm is utilized to identify the complexity features that contribute the most toward correctly predicting complexity scores and accurately classifying motions.

Chapter 5

A Motion Complexity Metric Using Spatial-Temporal Features

In this chapter, we investigate the creation of a measure for quantifying the observed articulated motion complexity of a single subject in video. Uses for having such a measure include video indexing, motion classification, motion comparison, and advanced biological study of visual communication. In addition, to the best of our knowledge, no current standardized measure for visual articulated motion exists.

While the majority of previous attempts have utilized optical flow for capturing the unique signatures of motion, our approach utilizes a novel set of motion complexity features generated from a set of space-time interest points. By incorporating information from both the spatial and temporal domains, we demonstrate the efficacy of this set of features towards capturing the various signatures of articulated motion complexity. The accuracy is shown on a set of human and spider videos through the creation of a set of classifiers aimed at predicting both the complexity score of an observed motion as well as classifying the motion class. As ground truth data is nonexistent, a user study on motion complexity is conducted for obtaining ground

truth information for the complexity values of the dataset videos. The work in this chapter is planned for publication in [13].

5.1 Introduction

The analysis of motion is a critical component of many computer vision systems. Motion estimation has made it possible to estimate three-dimensional structure, recognize visual patterns for classification, and even recognize security threats or other emergency situations that may be in progress or imminent. However, a component of this analysis that has received little attention is the visual analysis of motion complexity. That is, the vast majority of previous work has focused on how to estimate motion and use the motion estimation for real-world problems instead of analyzing the complexity of the motion itself. An important motion class that is exemplified by the movement of many living beings is articulated motion. In such cases, the object is composed of a set of segments connected by joints. The existence of a measure that could quantify the visual complexity of articulated motion has many potential uses in a variety of real-world domains. For example, an articulated motion measure could allow for comparing one dance routine to another, indexing videos in a search database based on the motion complexity, or studying the subtle differences of the visual communication patterns of one species from another based only their movements.

In this chapter, we investigate the creation of a complexity measure for articulated motion complexity. The aim is to be able to accurately quantify the complexity of the observed motion of any general articulated movement of a single subject recorded with a non-moving camera. Throughout the process, we also aim to identify which aspects of motion contribute toward the overall measure, as well as to what degree. There is no

agreed upon definition of what it means to be visually complex with regards to motion. Thus, we aim to not only identify a complexity value for a given set of motions, but also the complexity domains that contribute to the idea of being "complex". For example, we investigate if larger/smaller, shorter/faster, or periodic/non-periodic motions indicate more or less complexity. In order to do so, we generate a set of motion complexity features that together cover all areas of the complexity domains.

Our approach relies on a new set of features that are generated from a set of *space-time interest points* (STIPs). STIPs have long been used in the activity recognition domain for learning a set of movements that accurately describe a motion. While a large percentage of previous work has focused on using optical flow for estimating motion, the approach presented in this chapter utilizes STIPs to locate the points in the space-time volume of video data that are significantly "interesting". By incorporating both space and time in the feature set, the goal is to create a measure that can capture both the spatial and temporal signatures of the displayed motion complexity. We specifically investigate two uses of such a measure for demonstrating its usefulness: 1) predicting the complexity scores for a set of videos, and 2) classifying videos into their respective motion classes. The accuracy is obtained by utilizing a user study on motion complexity for obtaining the ground truth information. We also investigate each motion complexity feature separately to observe its usefulness as a stand-alone feature in terms of accuracy.

The vast majority of work in motion analysis has revolved around humans subjects. Human subjects already have a large and readily available collection of videos demonstrating a wide variety of movements. In addition, a prioritized desire exists to study humans due to real-world applications in the security, entertainment, and health domains. Other interesting domains exist, however, such as the analysis of spider movements. A desire exists in the biological domain for more advanced ways to

study the differences between a variety of species, and spiders provide a challenging and exciting domain of exploration. One way to provide this more advanced analysis of species is through the creation of a complexity measure, and this chapter explores the creation and application of a complexity measure to both human subjects and wolf spider subjects. By exploring two vastly different subjects, we can identify complexity features that vary in importance depending on the domain. That is, a feature that contributes greatly toward the complexity of a spider may have little contribution toward the complexity of a human.

5.1.1 Problem Definitions

This work presents an analysis of visual motion complexity by utilizing space-time interest points toward the prediction of motion complexity scores, as well as the classification of motion classes (classification). Here we provide formal definitions for each of these problems scenarios.

Complexity Score Prediction The problem of predicting motion complexity scores is defined as follows: Given a set of videos $\mathbf{V}$ where each video $V = [F^1, F^2, \ldots, F^t]$, $F^i_{x,y}$ is the pixel in the x^{th} row and y^{th} column of the i^{th} image frame, and t is the total number of frames in the video, our goal is to create a complexity model C that takes a motion sequence of images (video) as input and generates a value between 1 (lowest possible complexity) and 10 (highest possible complexity) as output. Thus, we aim to find or train a function $C : V \rightarrow [1, 10]$, where $[1, 10]$ is the set of integers between 1 and 10, inclusive.

Motion Class Classification The problem of classifying the motion class of a video is defined as follows: Given a set of videos $\mathbf{V}$ defined as above, our goal is to

train a classification model C using feature set M and assigned motion classes taken from the set L for the purposes of classifying unknown motion instances. Formally, we aim to train a function $C : V \rightarrow \{L_1, L_2, \ldots, L_n\}$ using a set of motion complexity features and videos from a dataset that, given a new video as input, predicts a motion class from L. Here, n is the number of motion classes that could be assigned.

5.1.2 Approaches

The proposed set of motion complexity features is computed based on the detection of selective space-time interest points (S-STIPs) [9]. We utilize these S-STIPs to investigate their efficacy toward describing motion complexity signatures. Our goal is to detect hidden complexity information from the spatial-temporal domain that might be otherwise hidden using only the spatial domain.

As visual motion complexity has no standard definition, our approach relies on a user study (Appendix B) where the complexity values for videos are obtained from a group of participants based on human opinion. This user study on visual motion complexity is introduced in Section 4.3 and detailed further in Appendix C. A set of motion complexity features are then computed for each video, which are defined in Section 4.2. We utilize a sequential feature-selection algorithm to choose the subset of features that give the best accuracy in terms of prediction and classification, and compare a scenario using the best features against a scenario using all of the features. The selected motion complexity features are 1) combined with the user supplied complexity ratings to train a discriminant analysis classifier for prediction, and 2) combined with the motion classes to train a discriminant analysis classifier for classification. A visual overview of our approach is shown in Figure 5.1.

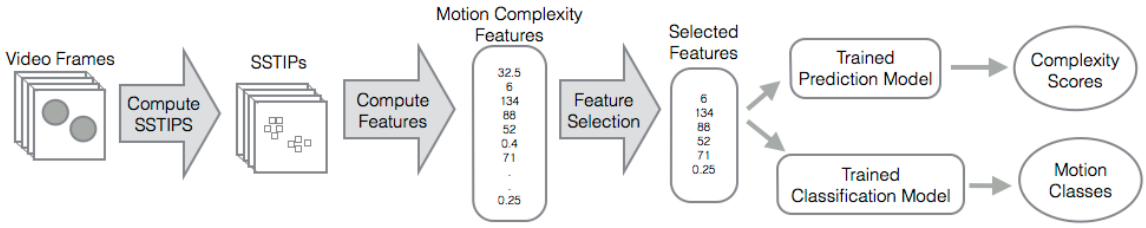


Figure 5.1: Overview of the three approaches.

Complexity Score Prediction (Linear Discriminant Analysis) We propose an alternative approach for motion complexity prediction that utilizes a pattern-recognition-based technique. We train a classification model using the human-assigned complexity scores from the user study. Specifically, we use linear discriminant analysis as the learning algorithm for the classifier. This technique was chosen using empirical testing among several classification algorithms including decision trees, clustering techniques, and support vector machines (SVMs). This trained classifier attempts to correctly predict the complexity score of an unseen motion class.

It is worth noting that the prediction problem is being treated as a classification problem. That is, instead of training a regression-based function, the prediction values are rounded to the nearest integer and used in a trained classification model. While the ultimate goal that this work progresses toward is a specific, real-number-based score, many application domains only require a higher-level categorization of the complexity scores. For example, many applications may only require the knowledge of whether the computed score is low complexity, medium complexity, or high complexity. Other domains may only need the complexity score on a scale of one to ten. This work presents the categorized version of the problem that can ultimately lead to a regression-based analysis in future work.

Motion Class Classification (Linear Discriminant Analysis) Instead of predicting complexity scores, our second approach toward visual motion analysis attempts

to correctly classify the motion class of the video. For example, it will attempt to distinguish between a walking motion video and a running motion video. Similar to our approach for complexity score prediction using a pattern-recognition-based approach, we use the same approach here. That is, we train a linear discriminant classifier using the motion complexity features and the motion classes. The classifier attempts to correctly determine the motion class of an unseen video.

5.1.3 Contributions

The goal of this work is to investigate two pattern-recognition-based approaches towards the creation of a model for both predicting the complexity scores of articulated motion in video as well as classifying the motion class using a novel set of *motion complexity features*. Such a model could be useful as a measure in a number of applications such as providing a numerical value that can be integrated in the understanding of various species (such humans, mice, or wolf spiders), indexing motion/activity videos, or classifying a wide range of movements (such as one dance routine from another). The overall contributions of this chapter are as follows:

1. Presents a novel set of motion complexity features based for use in analyzing complexity in general articulated motion based on features from the space-time domain (selective space-time interest points)

2. Provides a comparison of the accuracy power of each individual feature over several scenarios, revealing the interesting features that contribute the most toward a complexity measure

3. Demonstrates the performance of a pattern-recognition (linear discriminant analysis) model for predicting articulated motion complexity scores

4. Demonstrates the performance of a pattern-recognition (linear discriminant analysis) model for classifying video motion classes

5.1.4 Datasets

We again use the two datasets from the previous chapter: a dataset of human actions and a dataset of spider motions demonstrated during their courtship routine. We briefly review them here for completeness. A detailed description is provided in Appendix C.

Basic human actions The human action dataset used in this work is the Weizmann dataset [23], a widely used collection of basic human motions for comparing action classification systems. It contains 81 low-resolution (180 $\times$ 144) video sequences, recorded at 25 FPS, displaying nine different people performing nine basic actions. The displayed action classes are "running (run)", "walking (walk)", "jumping jack (jack)", "jumping forward on one leg (skip)", "jumping in place on two legs (jump)", "galloping sideways (side)", "waving one hand (1wave)", "waving two hands (2wave)", and "bending (bend)".

Spider courtship movements The second dataset contains high frame rate videos of samples of two species of *Schizocosa* wolf spider: *S. bilineata* and *S. crassipalpata*. There are 52 total grayscale videos in the dataset, where each video is roughly six seconds in length. The dataset is divided into two halves (one half for each species), while each of those halves is further divided into two (high diet and low diet). The separation of high diet from low diet comes from the expectation that nutrient intake could influence the degree to which spiders can engage in complex courtship displays. Thus, by varying the diet of individuals, we can assess whether there is a link between

nutrient intake and courtship complexity. Each video has a temporal resolution of 250 FPS for capturing the quick movements of the spiders, and a varying spatial resolution due to cropping out the areas of interest in each clip. When computing optical flow vectors for each frame, any vector with a very small magnitude (< 0.02) is discarded before computing the motion complexity features to eliminate noise. Adjusting this threshold could cause significant changes in the final results, and should be carefully selected depending on the application domain.

5.2 Motion Complexity Features

In this section, we present and detail a novel set of motion complexity features based on a computed set of selective-STIP points (S-STIPs). We specifically aim to create a set of features that cover a set of motion complexity domains we believe contribute significantly to visual motion complexity. This set of motion complexity domains is shown in Table 5.1. Every feature we propose can be mapped into one of these complexity domains.

Motion Domain	**Description**
Movement amount	Degree of spatial-temporal motion
Movement stability	Stability of the motion intensity
Movement periodicity	Repetition of motion
Movement synchrony	Multiple motion units moving simultaneously
Movement parts	Number of moving areas

Table 5.1: Spatial-temporal motion complexity domains.

The majority of these features are based on a set of *motion units*, which are themselves computed from the S-STIPS. We define a motion unit as a connected component of the space-time volume of S-STIPs. Specifically, we assume we have a set of S-STIPS S (each element a set of (x, y, t) coordinates) computed on a video

volume V of dimensions $x \times y \times t$. We then compute a corresponding binarized copy B from V where a '1' is assigned if an S-STIP exists at the corresponding (x, y, t) location, or a '0' otherwise. We then input B into a three-dimensional connected-component-labeling algorithm to detect spatial-temporal clusters of points. These three-dimensional clusters of points are the units of motion (motion units) M over both space and time used to compute the motion complexity features. We now define the set of motion complexity features as follows (where t is used to represent any given frame, T is the number of video frames, PC^t is the number of S-STIPS is video frame t):

Point Count (Mean) – The average number of S-STIPS over all frames, providing a measure of the overall interesting motion in a video:

$$\frac{1}{T}\sum_{t=1}^{T} PC^t \tag{5.1}$$

Point Count (STD) – The standard deviation of the number of S-STIPS over all frames, providing a global measure of motion stability over time:

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}(PC^t - PCM)^2} \tag{5.2}$$

where PCM is the Point Count (Mean) feature for the video.

Large Scale Percentage – The percentage of points that are large scale (> 3), where a point detected at a large scale has greater spatial and temporal response:

$$\frac{length(S_{large})}{length(S)} \tag{5.3}$$

where S_{large} is the subset of S where the spatial scale of the points is greater than three.

Motion Unit Count – The number of motion units (three-dimensional space-time clusters) over the length of the video, normalized by the frame count:

$$\frac{length(M)}{T} \tag{5.4}$$

Motion Unit Synchrony – The number of frames containing more than one motion unit ($B^t_{\geq 2}$) divided by the number of frames with at least one motion unit ($B^t_{\geq 1}$):

$$\frac{B^t_{\geq 2}}{B^t_{\geq 1}} \tag{5.5}$$

Burst Count – The number of movement bursts, normalized by frame count, where a burst is defined as the number of S-STIP frame sums that are larger than one standard deviation from the mean:

$$\frac{\sum_{t=1}^{T}(PC^t > std(PC))}{T} \tag{5.6}$$

Primary Frequency – The frequency (in hertz) of the largest frequency response from the mean-subtracted and unit normalized PC vector, as computed by a short-time discrete Fourier analysis algorithm [28]. As the algorithm reveals frequency strengths at different windowing scales, we average all of the frequency responses together, then record the largest frequency.

Primary Frequency Strength – The frequency response value of the primary frequency.

Frequency Peak Count – We compute the number of frequency responses (of any strength) by taking the vector of averaged frequencies from a short-time discrete Fourier analysis algorithm, and count the number of peaks (values that are larger than both values sequentially to the left and right. This provides a measure of how many frequencies had a response of any size.

Point Count Peak Width (Mean) – The average distance between every successive pair of peaks in PC in the temporal direction (peak width), providing a measure of temporal motion stability.

Point Count Peak Width (STD) – The standard deviating distance between every successive pair of peaks in PC in the temporal direction (peak width), providing a measure of temporal motion stability.

Point Count Peak Height (Mean) – The average distance between every successive pair of peaks in PC in the spatial direction (peak height), providing a measure of motion intensity stability.

Point Count Peak Height (STD) – The standard deviating distance between every successive pair of peaks in PC in the spatial direction (peak height), providing a measure of motion intensity stability.

Motion Unit Size (Mean) – The average spatial-temporal size (in pixels) of the motion units for a video:

$$\frac{1}{length(M)} \sum_{m=1}^{length(M)} pixelCount(M^t) \tag{5.7}$$

Motion Unit Size (Max) – The maximum spatial-temporal size (in pixels) of the

motion units for a video:

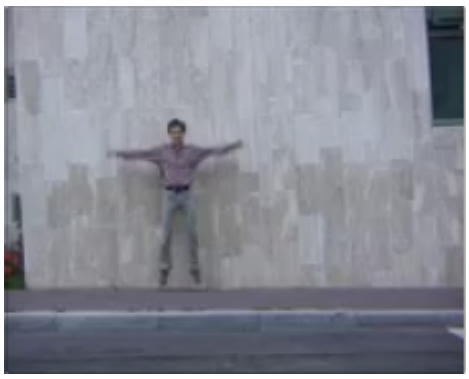
$$max(pixelCount(M)) \tag{5.8}$$

Motion Unit Lifespan (Mean) – The average temporal length of all motion units for a video.

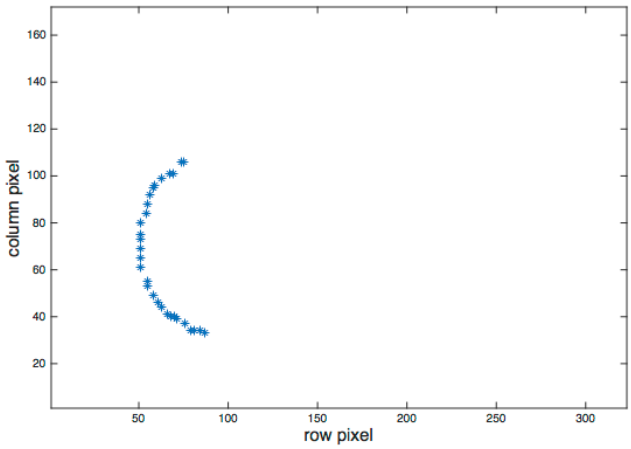
Motion Unit Lifespan (STD) – The standard deviation of the temporal lengths of all motion units for a video.

Motion Unit Trajectory (Mean) – The average amount of distance travelled (in pixel) of all the motion units, determined by computing the centroid (of the binarized connected component) movement distance of each time slice of a motion unit. That is, we follow the centroid of a motion unit over time, and compute how far it travelled in pixels.

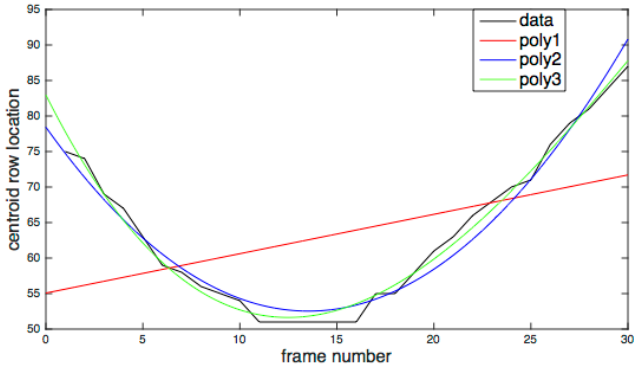
Motion Unit Polynomial Fit (Mean) – For a motion unit, we compute the centroid of each time slice of the unit, and store the centroid 'x' locations (can alternatively be done for the 'y' locations instead). By plotting the centroid 'x' locations as they change over time, we attempt to fit a set of polynomials (from a first-order polynomial up to an ninth-order polynomial) to the points, keeping the lowest possible polynomial that fits the points sufficiently (when the r-square value is ≥ 0.99). For example, a motion cluster that moves in the same direction over time will have a first-order polynomial fit sufficiently. The average polynomial order (from the integer set $1, 2, \ldots, 9$ of polynomial orders) is computed from all motion units for a video. An example visualization of this technique is shown in Figure 5.2.



(a) Sample frame from a jumping jack video.



(b) A selected space-time cluster (motion unit) plotted spatially by ignoring the temporal domain.



(c) S-STIPS visualized over space and time.

Figure 5.2: The first three polynomial orders fitted to the centroid 'x' location over time.

Each feature from this proposed set is mapped into one of the domains listed in Table 5.1. The goal is to have several features included in each motion complexity domain, as we believe these domains to be the most important measures toward quantifying complexity. The mappings of the features to the domains is presented in Table 5.2.

Motion Domain	**Description**
Movement amount	Point Count (Mean) Large Scale Percentage Motion Unit Size (Mean) Motion Unit Size (Max) Motion Unit Lifespan (Mean) Motion Unit Trajectory (Mean)
Movement stability	Point Count (STD) Burst Count Point Count Peak Width (Mean) Point Count Peak Width (STD) Point Count Peak Height (Mean) Point Count Peak Height (STD) Motion Unit Lifespan (STD) Motion Unit Polynomial Fir (Mean)
Movement periodicity	Primary Frequency Primary Frequency Strength Frequency Peak Count
Movement parts	Motion Unit Count Motion Unit Synchrony

Table 5.2: Spatial-temporal motion complexity features mapped into their respective motion complexity domains.

5.3 User Study On Motion Complexity

In this section, we briefly review the user complexity study presented in Chapter 4 (detailed further in Appendix B) for completeness, as the approaches listed in this chapter utilize the ratings provided by the users for training and testing (ground truth).

A user study was conducted on 24 participants from varying backgrounds to investigate what (based on user opinion) makes a motion complex. For each participant, a series of videos was shown from the two datasets described in Appendix C. Each clip was played repeatedly while waiting for the user to rate the displayed motion on a scale of one (low complexity) to ten (high complexity). 25% of the displayed videos were randomly chosen to be shown twice to measure a rater's accuracy. A summary of the ratings given to humans for the nine motion classes is shown in Figure 5.3, while the summary of the ratings given to spiders for the four cases is shown in Figure 5.4.

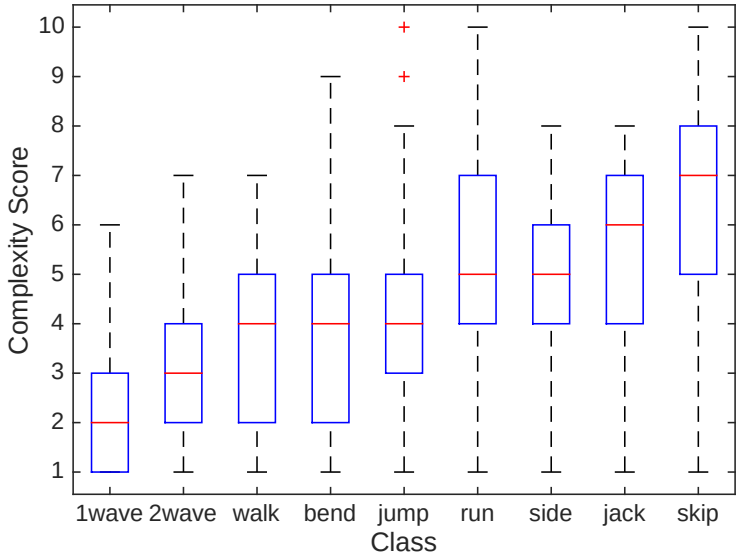


Figure 5.3: Overview of user ratings for the nine human motions.

5.4 Implementation & Results

Here, we demonstrate the potential of using S-STIPs for quantifying motion complexity. We first look at the S-STIP detection process to gain a better understanding of what the interest points are representing. We also present a visualization of the S-STIPs over time, where important motion signatures can be observed. We then present the

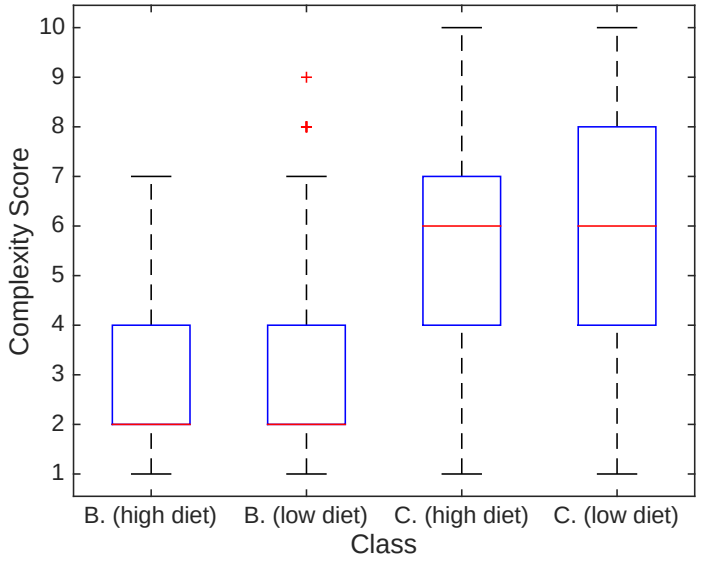


Figure 5.4: Overview of user ratings for the spider movements of *S. Bilineata* (B) and *S. Crassipalpata* (C).

results for the two problem scenarios: 1) predicting motion complexity scores and 2) predicting motion classes. Accuracies are reported for several sub-scenarios, and the prediction and classification power of the individual features is shown.

5.4.1 S-STIP Detection

We first visualize the detected S-STIPs on samples from the two datasets of humans and spiders. The S-STIP detection process is detailed in Chapter 2.2. We provide example visualizations of two human samples (walking from left to right and jumping jack), as well as two spider samples. For each sample, we show (a) a sample frame with the S-STIPs superimposed on the image frame, (b) the S-STIPS for the entire video collapsed into the spatial domain by ignoring the temporal domain to show any location with interesting movement, and (c) the entire set of S-STIPS plotted over both space and time. These visualizations are shown in Figure 5.5 (a walking pattern can be seen as a linear plane of S-STIPs), Figure 5.6 (the repetition pattern of the

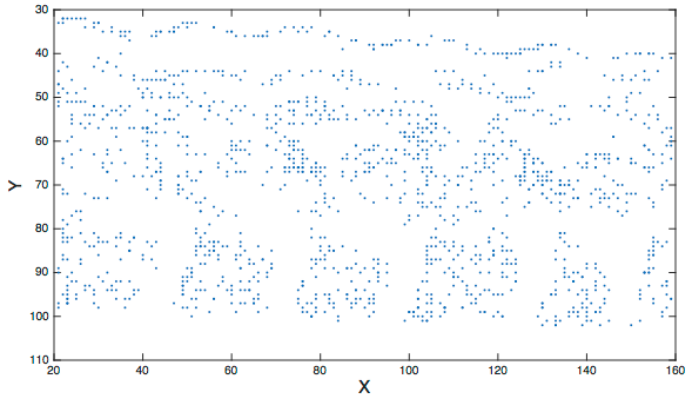
jumping jack movement is visible), Figure 5.7 (the path of the spider leg and pedipalp vibration is visible), and Figure 5.8 (showing two significant periods of motion over time).

5.4.2 Predicting Complexity Classes

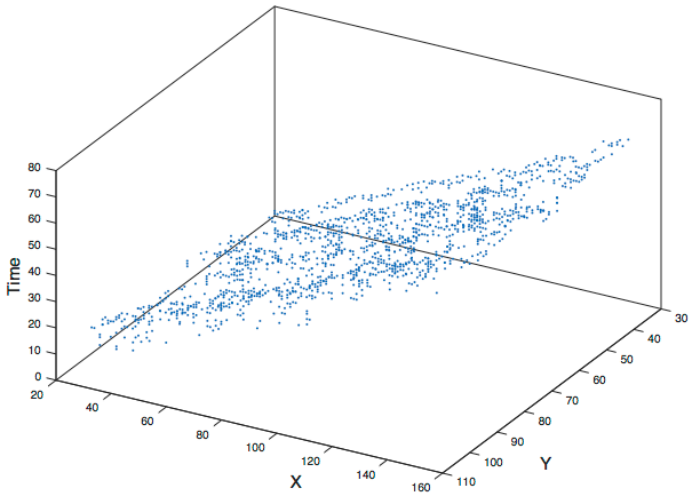
A pattern recognition approach was used to learn the important features for each dataset. Specifically, a linear discriminant classifier was trained on the motion complexity features using the average human complexity scores (rounded to the nearest integer) as the training labels. The training/testing split used was $\frac{2}{3}/\frac{1}{3}$. To compensate for the random selection of the training and testing sets, each classifier was trained and tested 1000 times, with the average classification score used for the classifier accuracy.

Several tests were performed to observe how the accuracy changes. For both the human and the spider datasets, the training score was first set to be the average human complexity score for each individual video. The second case set the training score to be average human complexity score for the given video class. For example, if the average "jumping jack" action for the human videos was a score of 0.4, then every "bend" video was assigned a 0.4. The number of features used was also varied from using all of the features to only using the "best" (top) features determined by a sequential feature selection algorithm. The top features for the human dataset were Point Count (STD), Motion Unit Synchrony, Primary Frequency, and Motion Unit Lifespan (Mean), while the best features for the spider dataset were Point Count (Mean), Primary Frequency Strength, Point Count Peak Width (STD), Motion Unit Size (Mean), Motion Unit Size (STD), and Motion Unit Lifespan (Mean). Between the two, standard deviation proves useful as a measure of stability for both datasets. In addition, primary frequency is revealed to be a strong signature for predicting

(a) Detected S-STIPs (in red) from a single frame.



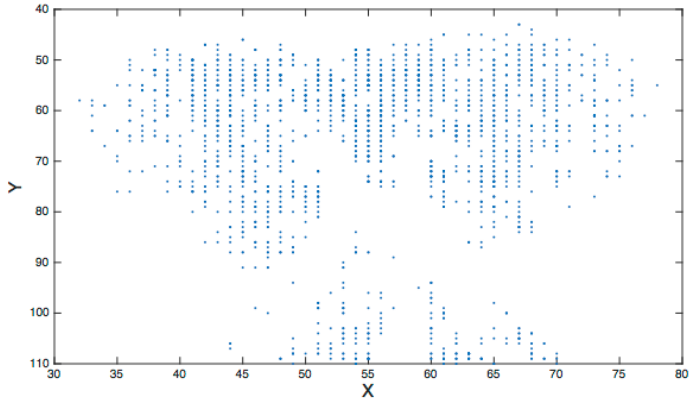
(b) S-STIPs visualized spatially for all frames by ignoring the temporal domain.



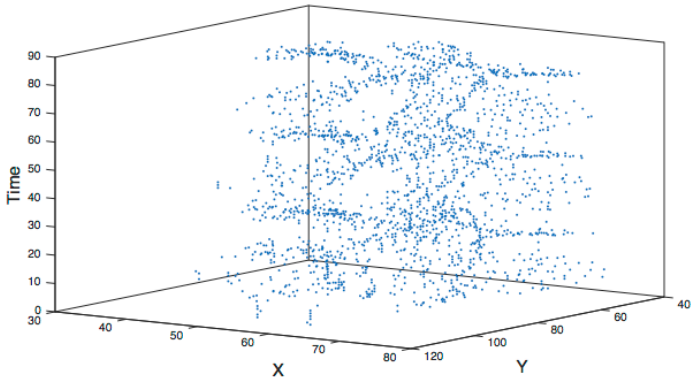
(c) S-STIPS visualized over space and time.

Figure 5.5: S-STIP detection of a human walking.

(a) Detected S-STIPs (in red) from a single frame.

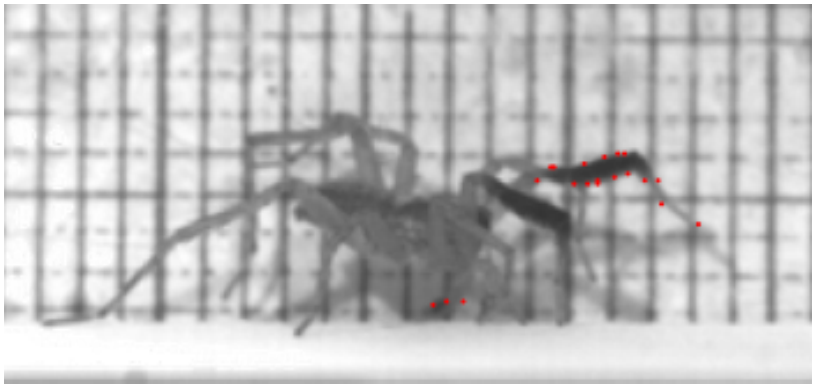


(b) S-STIPs visualized spatially for all frames by ignoring the temporal domain.

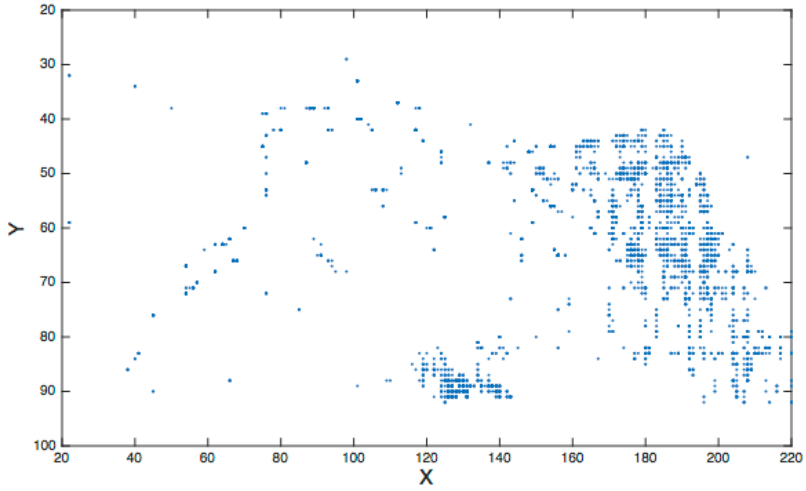


(c) S-STIPS visualized over space and time.

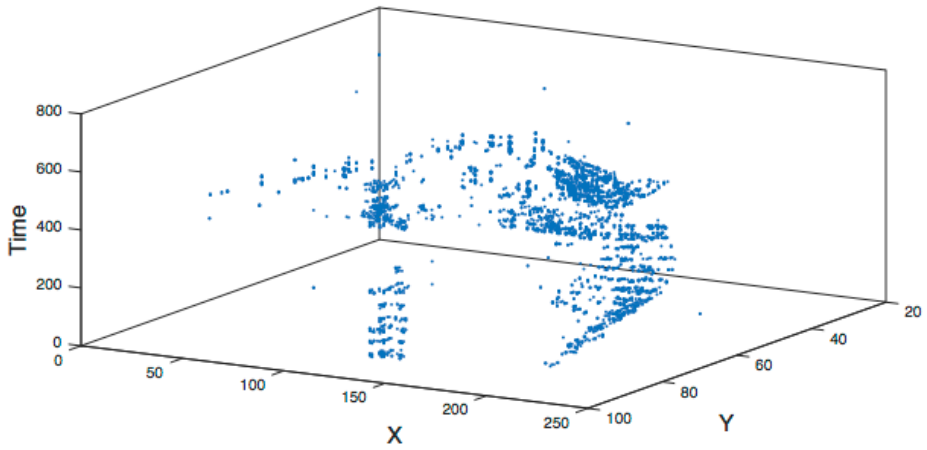
Figure 5.6: S-STIP detection of a human performing jumping jacks.



(a) Detected S-STIPs (in red) from a single frame.

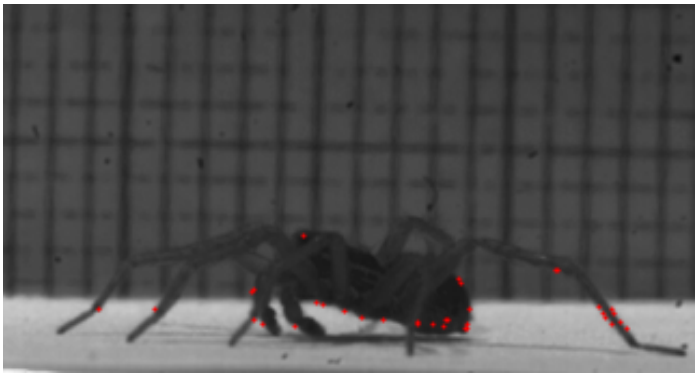


(b) S-STIPs visualized spatially for all frames by ignoring the temporal domain.

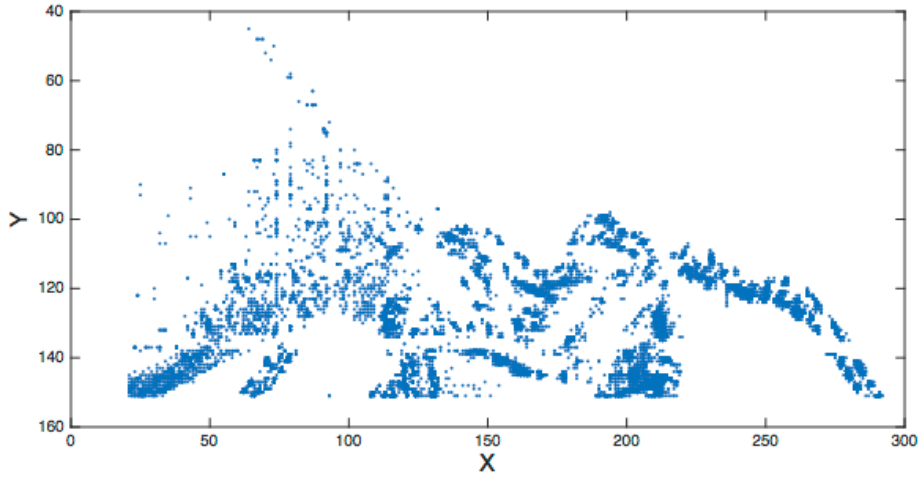


(c) S-STIPS visualized over space and time.

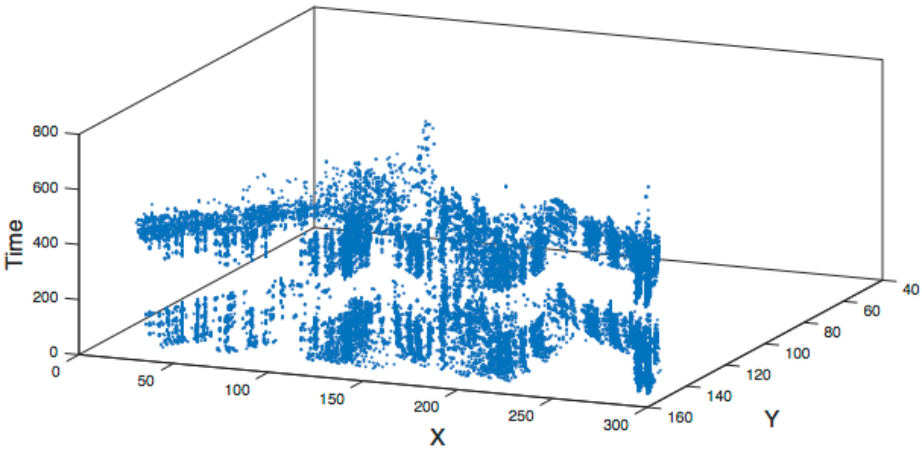
Figure 5.7: S-STIP detection of a spider moving both a leg and its pedipalps.



(a) Detected S-STIPs (in red) from a single frame.



(b) S-STIPs visualized spatially for all frames by ignoring the temporal domain.



(c) S-STIPS visualized over space and time.

Figure 5.8: S-STIP detection of a spider showing several areas of significant movement over time.

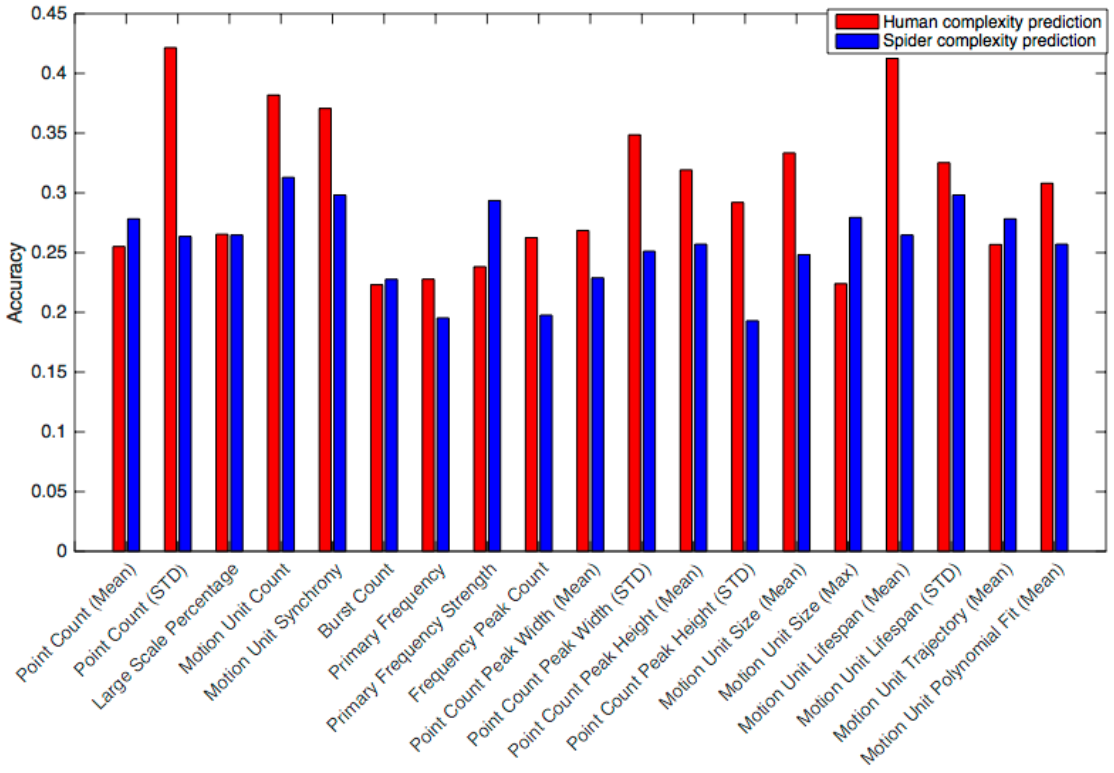


Figure 5.9: Individual feature prediction accuracy for both spider and human complexity scores.

motion complexity scores. The strength (accuracy) of the features used individually for predicting complexity scores is shown in Figure 5.9 for both humans and spiders separately. The results of motion complexity prediction are presented in Table 5.3. As the results show, using only the top features increases the accuracy only slightly in every case. Using the mean average rating for the videos instead of individual video scores showed significantly greater prediction accuracy. In addition, increasing the allowed range to 1 or 2, which can be acceptable in some domains, reveals significantly greater prediction accuracy as well.

Dataset	Training Label	Features	Allowed Range	Accuracy
Human	Video Score	All	0	24%
			1	60%
			2	81%
		Top	0	38%
			1	82%
			2	99%
	Mean Class Score	All	0	63%
			1	83%
			2	97%
		Top	0	63%
			1	87%
			2	98%
Spider	Video Score	All	0	38%
			1	84%
			2	92%
		Top	0	40%
			1	90%
			2	96%
	Mean Class Score	All	0	78%
			1	96%
			2	96%
		Top	0	86%
			1	97%
			2	97%

Table 5.3: Accuracy of the discriminant analysis approach for predicting motion complexity scores.

5.4.3 Classifying Motion Classes

The previous classifier-based approach was used to predict complexity scores. Here, another classifier using linear discriminant analysis is trained, but instead used to learn and classify motion classes. That is, instead of learning and predicting scores for a 'walk' video using motion complexity features, it attempts to learn and classify a video motion as 'walk'. Here, the classifier is trained for five different scenarios: 1) classifying human actions, 2) classifying spiders into the four classes (two species, with each species having low diet and high diet samples), 3) classifying as either

spider species one or species two, 4) classifying between high diet and low diet for species one, and 5) classifying between high diet and low diet for species two. Using sequential feature selection, the top features for each of the five scenarios, respectively, are 1) Point Count (Mean), Point Count (STD), Motion Unit Count, and Motion Unit Synchrony, 2) Point Count (Mean), Point Count (STD), Motion Unit Count, Motion Unit Synchrony, and Point Count Peak Height (Mean), 3) Point Count (Mean), Point Count (STD), Large Scale Percentage, Motion Unit Count, Motion Unit Synchrony, Point Count Peak Width (Mean), Point Count Peak Height (Mean), Motion Unit Size (Max), and Motion Unit Lifespan (STD), 4) Motion Unit Synchrony, Frequency Peak Count, Point Count (Mean), Point Count (STD), and Point Count Peak Height (Mean), and 5) Primary Frequency Strength, Point Count (Mean), Point Count (STD), Burst Count, and Point Count Peak Height (Mean). These results are visualized in detail in Figure 5.10 and Figure 5.11. Overall, this identifies Motion Synchrony as very important feature for general articulated motion, as well as Point Count (Mean) and Point Count (STD).

The motion classification results are shown in Table 5.4. For classification, it can be seen that using only the top features produces mixed results, causing only minor accuracy improvements for the spider cases and causing a significant drop in accuracy for human motions. While the classifier does well at classifying one spider species from another, it struggles to correctly classify between high diet and low diet. Visual observation by humans also identifies this to be a difficult problem. However, we again note that classification is only a desired effect of the measure, while complexity score prediction is our main goal.

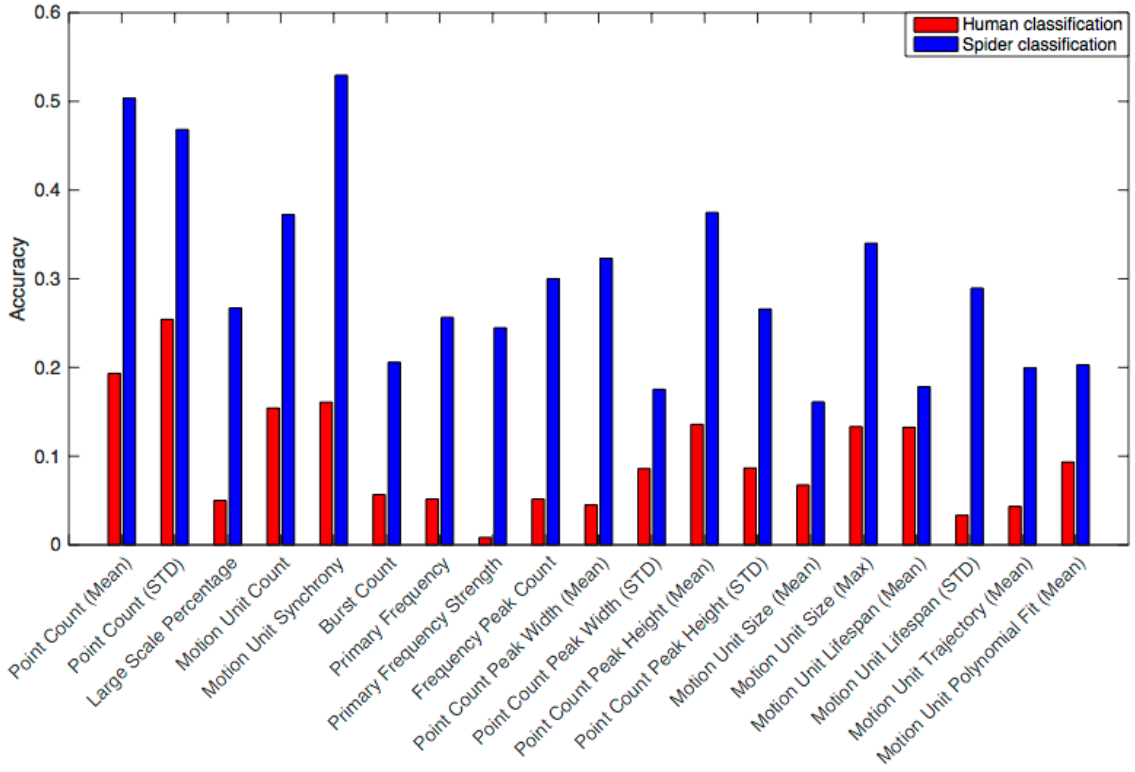


Figure 5.10: Individual feature classification accuracy for both spider and human motion classes.

Dataset	Label Domain	Features	Accuracy
Human	Nine human actions	All Top	61% 58%
Spider	$\{B_H, B_L, C_H, C_L\}$	All Top	50% 50%
Spider	$\{B, C\}$	All Top	95% 95%
Spider	$\{B_H, B_L\}$	All Top	47% 54%
Spider	$\{C_H, C_L\}$	All Top	56% 61%

Table 5.4: Accuracy of the discriminant analysis approach for classifying complexity classes.

5.5 Summary

We have presented an in-depth study of visual motion complexity by proposing a novel set of motion complexity features based on a spatial-temporal feature technique for

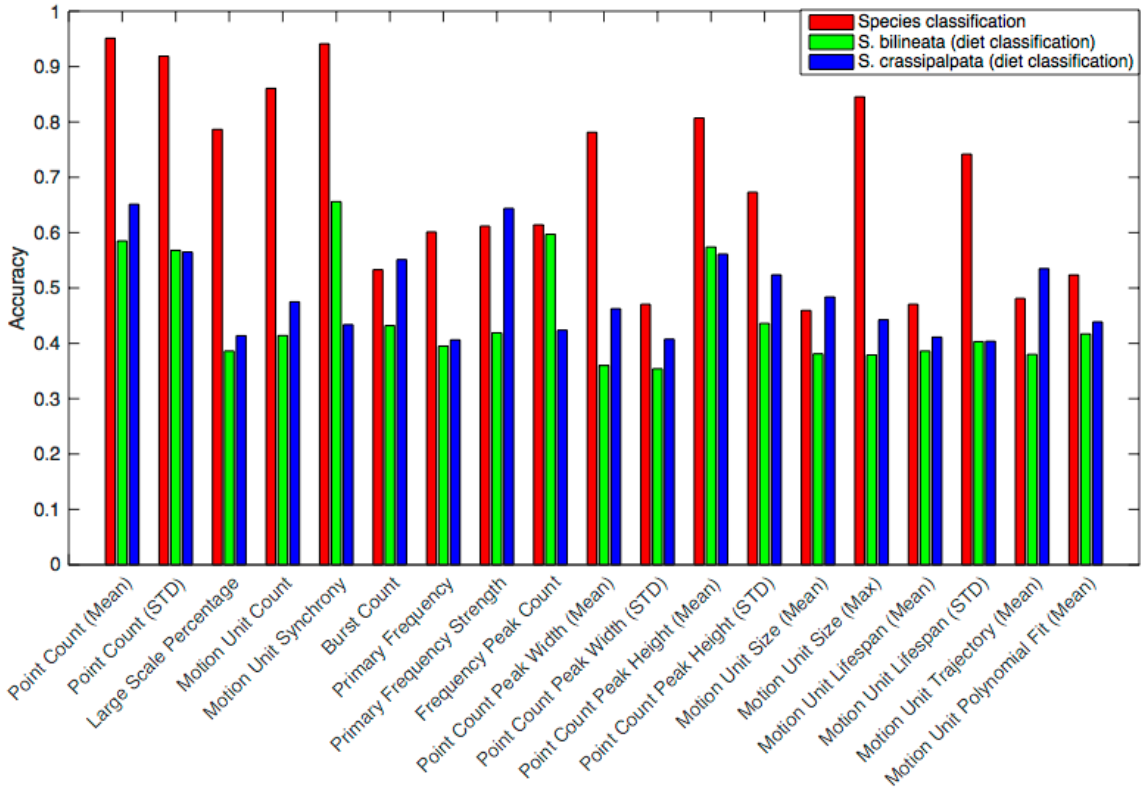


Figure 5.11: Individual feature classification accuracy for three alternative spider scenarios (species vs. species, species 1 high vs. low diet, and species 2 high diet vs. low diet).

both prediction of complexity scores and motion classification. Based on a user study of visual motion complexity, these features were learned using linear discriminant analysis to train classifiers for the purpose of both predicting complexity scores as well as classifying motion classes on a dataset of human actions and a dataset of spider motions. The complexity features were shown to be effective in many cases for correctly classifying motion classes as well as predicting complexity scores, notably when increasing the allowed error range to 1 or 2. It was also shown that using a set of "best" features leads to increased accuracy for predicting complexity scores, but produces slightly mixed results for classification. Even greater accuracy gains were made when using the mean class score instead of the individual video scores.

This chapter also revealed interesting results about specific motion complexity features that contribute the most to the overall complexity value. Specifically, it was observed that features involving motion synchrony, frequency analysis, and point-count statistics were identified to be the most useful features by the feature selection algorithm and individual feature classification. These features, we believe, hold the most useful information about the motion signatures of visual complexity in the spatial-temporal domain.

Chapter 6

Conclusion

In this chapter, a concise review of the work presented in this dissertation is provided. In addition, some closing remarks are given regarding the contributions of this work and the interesting results. Finally, a few possible interesting directions for continuing this research are mentioned.

6.1 Summary & Closing Remarks

This dissertation presented an in-depth study of visual articulated motion complexity by proposing a set of measures for quantifying the observed complexity. The foundation for a general articulated motion measure for complexity is provided that can benefit communities ranging from computer vision researchers to biologists by providing a deeper understanding of complexity. This research was divided into three main bodies of work, together providing the following contributions:

1. Identified a novel set of motion complexity features based on optical flow that encodes the various aspects of articulated motion complexity

2. Defined a measure for quantifying general motion complexity by integrating the motion features as a weighted sum based on feature contribution

3. Demonstrated the performance of a pattern-recognition (linear discriminant analysis) model based on optical flow for predicting motion complexity scores and distinguishing motion classes

4. Summarized the results of two user studies on visual motion complexity: 1) an expert poll on statistical feature importance for complexity, and 2) a user study where participants rate a dataset of videos for further analysis of what a typical person believes contributes to complexity

5. Presented a novel set of motion complexity features that utilize spatial-temporal features for integrating hidden complexity information not visible using a strictly optical flow-based strategy

6. Demonstrated the accuracy of a spatial-temporal feature approach for predicting motion complexity scores and distinguishing motion classes

7. Demonstrated the efficacy of the defined complexity measures in a real-world problem domain (the biological study of visual signals from spider movement)

In Chapter 3, an optical flow-based measure was proposed that relied on statistical values computed from the motion estimation. This measure demonstrated a weighted-sum approach, where a set of features was computed and weighted based on the belief of a group of domain experts. While the measure showed little potential for use in classifying a set of wolf spider movements, the computed complexity values were believed to be representative of the corresponding observed motion in the dataset videos. A feature-selection process provided insight toward which of these features was most critical toward contributing to the visual complexity.

In Chapter 4, a new set of optical flow-based measures were proposed with the goal of both predicting motion complexity scores as well as classifying motions based on their motion class. Higher-order features were defined to identify hidden aspects of motion complexity, such as repeating patterns of motion by incorporating Fourier analysis, motion cluster analysis, and motion synchrony. The efficacy of these features was demonstrated on both human and spider datasets, revealing the potential for using motion complexity signatures for classification. In addition, a user study on visual motion complexity was summarized for the purposes of providing ground truth information and revealing the beliefs humans have toward complex versus simple motions.

In Chapter 5, an optical flow-based approach was abandoned in favor of an approach based on space-time interest points. While optical flow provides an estimation of the speed and direction of motion for every pixel of a video frame, space-time interest points reveal the locations in the space-time volume where significant and interesting motion is taking place. This alternative approach provided many new and interesting insights into motion complexity by analyzing the clusters of space-time interest points (motion units) and how they change over time. The same user study from Chapter 4 was used to provide human-belief ground-truth information. Classifier-based measures were created for the purpose of both predicting visual complexity scores as well as classifying observed motions into motion classes.

Many useful results have been provided in this work that lead to a deeper understanding of visual motion complexity. We have identified several specific complexity features that contribute more greatly toward the complexity value than others. Specifically, the most useful features were shown to be statistical entropy, statistical kurtosis, primary frequencies and their strength, motion synchrony, point count mean and standard deviation, point count peak width, and motion unit lifespan. We also con-

clude that directional information may not be as critical as we initially predicted to the complexity measure as the non-directional-based features, although alternative directional-based features not proposed here may still hold potential.

We also demonstrated a deeper understanding of human belief regarding complexity. The expert poll in Appendix A revealed the belief in directional changes, statistical entropy, number of movement runs, number of motion clusters, directional distribution test (Rayleigh test), and statistical kurtosis as being the important contributors to complexity. The user study on motion complexity presented in Appendix B revealed the belief in amount of movement and number of moving parts contributing the most to complexity, with movement periodicity and motion synchrony contributing the least.

6.2 Comparison of Results

In general, the results in Chapter 4 showed more accuracy than those in Chapter 3, while the results in Chapter 5 showed more accuracy than those in Chapter 4. While the weighted-sum approaches showed early promise in accurately predicting complexity, the trained models displayed significantly better prediction capabilities. While the optical flow approaches were useful toward predicting complexity, the spatial-temporal approach showed a significant improvement over both of them. It may, however, be possible that a better approach would be a hybrid of spatial-temporal interest points and optical flow. That is, it may be possible to utilize the directional information around the space-time interest points to achieve increases in accuracy. A prediction comparison of the human-provided scores against the computed scores for the spider dataset using the approach from Chapter 4 versus the approach from Chapter 5 is shown in Figure 6.1 with the human dataset results in Figure 6.2.

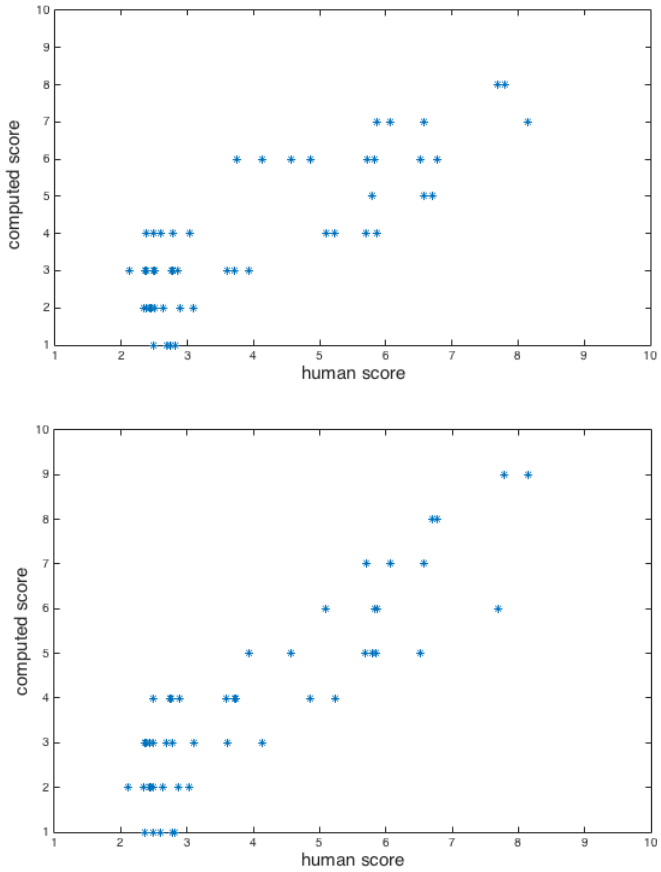


Figure 6.1: Comparison of the human-provided scores against the computed scores for the spider dataset using the approach from Chapter 4 (top) versus the approach from Chapter 5 (bottom).

6.3 Directions for Further Research

Here we note several possible directions in which this research could be extended. While the Horn-Schunck optical flow approach was used in Chapter 3 and Chapter 4, it is one of the first methods used to estimate optical flow. There have been several advances in optical flow in both speed and accuracy [21] that could be used to extend the measures into the real-time domain and improve motion-estimation accuracy. It may also be of interest to use alternative motion-estimation algorithms instead of optical flow. Alternative methods include block matching [5, 48] and phase

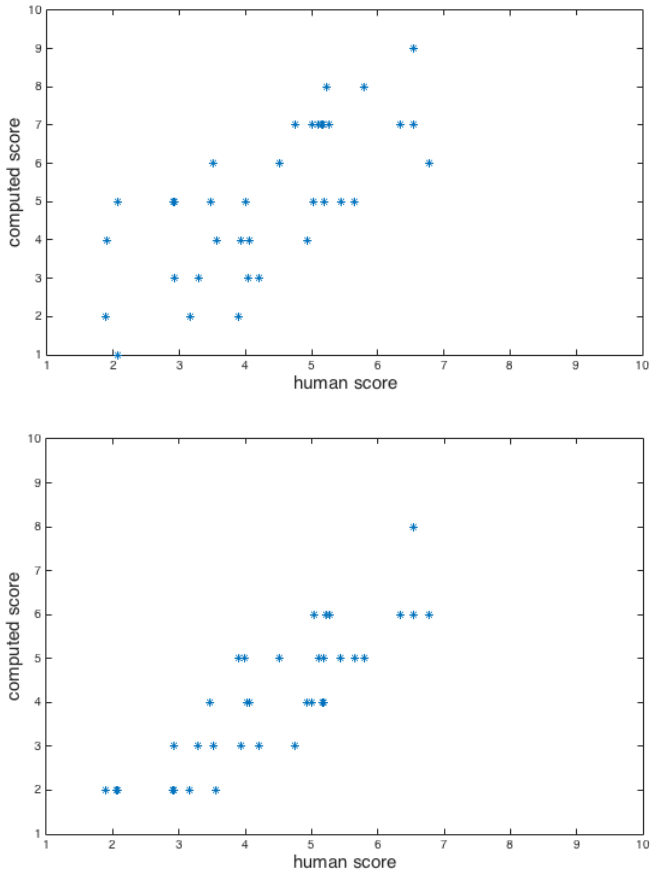


Figure 6.2: Comparison of the human-provided scores against the computed scores for the human dataset using the approach from Chapter 4 (top) versus the approach from Chapter 5 (bottom).

correlation [2, 57].

An interesting direction to pursue that is quite different from the work presented here would be to incorporate action recognition into the complexity measure. By utilizing the ability to learn and recognize specific actions (such as "person raised arm" or "spider quickly tapped leg"), a higher level of complexity understanding could be obtained. In addition, it would interesting to observe not only the actions that are recognized, but also the amount of times that they occur and the order in which they happen. A motion sequence could be then described as a string of characters, where

each character represents the specific action that was happening at that point in time.

While the directional-based features we presented in this dissertation did not show as much promise as the non-directional-based features, we still believe that direction plays an important role in the final complexity measure. An interesting route to take would be to utilize a space-time interest point approach to detect the interest points of movement in time, but extend it to focus on the directional values of motion at those points in time. This could essentially be a hybrid of space-time interest points and optical flow, similar to the idea of cuboids in Dollár et al. [17].

A weakness of the approaches described in this work is the limitation of fixed-size video segments. That is, it is assumed that the samples in a video dataset are approximately contain the same number of frames. While the approaches and motion complexity domains presented here would still be applicable, significant redesign of the motion complexity features would be needed to allow for variable-length video samples. This would be a logical extension of this work, and greatly expand the useful applications for the complexity measures.

The work presented in this dissertation was focused on videos from two datasets. Specifically, we focused on computing all of the features from a video image stack that was loaded ahead of time. Another interesting research path would be the application of these features to real-time complexity prediction and classification. While optical flow can be computed in real time, space-time interest points typically need the entire video volume to be present ahead of time. It may be of interest to pursue near-real-time computation of complexity using either optical flow or a modified version of a spatial-temporal-feature approach for providing a "current" complexity score for live video, or for providing a "current" guess as to which class the displayed motion complexity belongs.

Bibliography

[1] Jake K. Aggarwal and Quin Cai. Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440, 1999.

[2] Dimitrios S. Alexiadis and George D. Sergiadis. Motion estimation, segmentation and separation, using hypercomplex phase correlation, clustering techniques and graph-based optimization. *Computer Vision and Image Understanding*, 113(2):212–234, 2009.

[3] Saad Ali. Measuring flow complexity in videos. In *The IEEE International Conference on Computer Vision (ICCV)*, 12 2013.

[4] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011.

[5] Aroh Barjatya. Block matching algorithms for motion estimation. *IEEE Transactions Evolution Computation*, 8(3):225–239, 2004.

[6] John L. Barron, David J. Fleet, and Steven S. Beauchemin. Performance of optical flow techniques. *International journal of computer vision*, 12(1):43–77, 1994.

[7] Batschelet, editor. *Circular statistics in biology*. Academic Press, London ; New York, 1981.

[8] Philipp Berens. Circstat: a matlab toolbox for circular statistics. *J Stat Softw*, 31(10):1–21, 2009.

[9] Bhaskar Chakraborty, Michael B. Holte, Thomas B. Moeslund, and Jordi Gonzàlez. Selective spatio-temporal interest points. *Computer Vision and Image Understanding*, 116(3):396–410, 3 2012.

[10] I-Cheng . C. Chang and Chung-Lin . L. Huang. The model-based human body motion analysis system. *Image and Vision Computing*, 18(14):1067–1083, 2000.

[11] Chen-Yu Chen, Jia-Ching Wang, Jhing-Fa Wang, and Yu-Hen Hu. Motion entropy feature and its applications to event-based segmentation of sports video. *EURASIP Journal on Advances in Signal Processing*, 2008(1):460913, 2008.

[12] Alberto Chiarle and Marco Isaia. Signal complexity and modular organization of the courtship behaviours of two sibling species of wolf spiders (araneae: Lycosidae). *Behav Processes*, 97:33–40, 2013.

[13] Beau Christ, Ashok Samal, and Eileen Hebets. A motion complexity metric using spatial-temporal features. *International Journal of Computer Vision*, In preparation.

[14] Beau Christ, Ashok Samal, and Eileen Hebets. An optical flow feature-based metric for motion complexity. *Computer Vision and Image Understanding*, In 2nd review.

[15] Beau Christ, Ashok Samal, and Eileen Hebets. Prediction and classification using an optical flow-based complexity metric. *Pattern Recognition*, In preparation.

[16] Mark Claypool. Motion and scene complexity for streaming video games. In *Proceedings of the 4th International Conference on Foundations of Digital Games*, pages 34–41. ACM, 2009.

[17] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE, 2005.

[18] Murat Ekinci and Eyiip Gedikli. Silhouette based human motion detection and analysis for real-time automated video surveillance. *Turk J Elec Engin*, 13(2):199–229, 2005.

[19] Damian O. Elias, Bruce R. Land, Andrew C. Mason, and Ronald R. Hoy. Measuring and quantifying dynamic visual signals in jumping spiders. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol*, 192(8):785–97, 8 2006.

[20] N. I. Fisher. *Statistical analysis of circular data*. Cambridge University Press, Cambridge [England] ; New York, NY, USA, 1993.

[21] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134:1–21, 5 2015.

[22] Eyup Gedikli and Murat Ekinci. Human motion detection, tracking and analysis for automated surveillance.

[23] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE Trans Pattern Anal Mach Intell*, 29(12):2247–53, 12 2007.

[24] D. Gowsikhaa, S. Abirami, and R. Baskaran. Automated human behavior analysis from surveillance videos: a survey. *Artificial Intelligence Review*, 42(4):747–765, 2014.

[25] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Citeseer, 1988.

[26] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[27] Martin J. How, Jochen Zeil, and Jan M. Hemmi. Variability of a dynamic visual signal: the fiddler crab claw-waving display. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol*, 195(1):55–67, 1 2009.

[28] Eric Jacobsen and Richard Lyons. The sliding dft. *IEEE SIGNAL PROCESSING MAGAZINE*, 1053:5888, 2000.

[29] S. Rao Jammalamadaka and Ashis Sengupta. *Topics in circular statistics*. World Scientific, River Edge, N.J., 2001.

[30] Sylvie Jeannin and Ajay Divakaran. Mpeg-7 visual motion descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):720–724, 2001.

[31] Kanav Kahol and Mithra Vankipuram. Hand motion expertise analysis using dynamic hierarchical activity modeling and isomap. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[32] Alexander Klaser and Marcin Marszalek. A spatio-temporal descriptor based on 3d-gradients. 2008.

[33] Jan J. Koenderink and Andrea J. van Doorn. Representation of local geometry in the visual system. *Biological cybernetics*, 55(6):367–375, 1987.

[34] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[35] Ivan Laptev and Tony Lindeberg. Local descriptors for spatio-temporal recognition. In *Spatial Coherence for Visual Motion Analysis*, pages 91–103. Springer, 2006.

[36] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[37] Jonathan Feng-Shun . S. Lin and Dana Kulic. Segmenting human motion for automated rehabilitation exercise analysis. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 2881–2884. IEEE, 2012.

[38] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos in the wild. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1996–2003. IEEE, 2009.

[39] Jingen Liu and Mubarak Shah. Learning human actions via information maximization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[40] Tianming Liu, Hong-Jiang Zhang, and Feihu Qi. A novel video key-frame-extraction algorithm based on perceived motion energy model. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(10):1006–1013, 10 2003.

[41] Yijuan Lu, Ira Cohen, Xiang Sean Zhou, and Qi Tian. Feature selection using principal feature analysis. In *Proceedings of the 15th international conference on Multimedia*, pages 301–304. ACM, 2007.

[42] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981.

[43] Yu-Fei Ma and Hong-Jiang Zhang. A new perceived motion based shot content representation. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 3, pages 426–429 vol.3, 2001.

[44] Ross Messing, Chris Pal, and Henry Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 104–111. IEEE, 2009.

[45] Dimitris Metaxas and Shaoting Zhang. A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing*, 31(6):421–433, 2013.

[46] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2):90–126, 2006.

[47] Peter O'Donovan. Optical flow: Techniques and applications. *The University of Saskatchewan*, 2005.

[48] S. Immanuel Alex Pandian, G. J. Bala, and Becky Alma George. A study on block matching algorithms for motion estimation. *International Journal on Computer Science and Engineering*, 3(1):34–44, 2011.

[49] Kadir A. Peker, Ajay Divakaran, and Thomas V. Papathomas. Automatic measurement of intensity of motion activity of video segments. In *Proc. SPIE, M.M. Yeung, C.-S. Time*, 2001.

[50] R. A. Peters and C. S. Evans. Design of the jacky dragon visual display: signal and noise characteristics in a complex moving environment. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol*, 189(6):447–59, 6 2003.

[51] Ronald Poppe. Vision-based human motion analysis: An overview. *Computer vision and image understanding*, 108(1):4–18, 2007.

[52] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[53] Daniel K. Riskin, David J. Willis, José Iriarte-Díaz, Tyson L. Hedrick, Mykhaylo Kostandov, Jian Chen, David H. Laidlaw, Kenneth S. Breuer, and Sharon M. Swartz. Quantifying the complexity of bat wing kinematics. *J Theor Biol*, 254(3):604–15, 10 2008.

[54] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, pages 357–360. ACM, 2007.

[55] Jonathon Shlens. A tutorial on principal component analysis, december 2005. *URL http://www. snl. salk. edu/shlens/pub/notes/pca. pdf.*

[56] Cristian Sminchisescu. 3d human motion analysis in monocular video. In *IEEE Conference on Advanced Video and Signal Based Surveillance, Sydney, Australia*, 2006.

[57] Harold S. Stone, Michael T. Orchard, Ee-Chien . C. Chang, and Stephen Martucci. A fast direct fourier-based algorithm for subpixel registration of images. *Geoscience and Remote Sensing, IEEE Transactions on*, 39(10):2235–2243, 2001.

[58] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241, 2011.

[59] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Computer Vision–ECCV 2008*, pages 650–663. Springer, 2008.

[60] Shu-Fai . F. Wong and Roberto Cipolla. Extracting spatiotemporal interest points using global information. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[61] Jang-Hee . H. Yoo and Mark S. Nixon. Automated markerless analysis of human gait motion for recognition and classification. *Etri Journal*, 33(2):259–266, 2011.

[62] Junsong Yuan, Zicheng Liu, and Ying Wu. Discriminative subvolume search for efficient action detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2442–2449. IEEE, 2009.

[63] Jerrold H. Zar. *Biostatistical analysis*. Prentice Hall, Upper Saddle River, N.J., 1999.

Appendix A

Expert Poll Questionnaire

This chapter details the expert-polling study presented in Chapter 3 for weighting features based on expert belief. A questionnaire was presented to a group of 11 spider researchers (referred to as "the experts"). This study was completed in order to gain an understanding of which features a group of researchers, who have prior knowledge of working with spiders, believe influences complexity the most and which influence the least. The form displayed in Figure A.1 was presented to each expert. By not discussing their thoughts with other experts, each expert was asked to rate the importance that he/she believes that each motion feature contributes to motion complexity in wolf spiders on a scale of zero (no influence) to three (heavy influence). For each rating provided, each expert was also asked to provide their confidence in the answer they provided with a zero (not confident) or one (confident). The average responses from the 11 experts for both the influence score and response confidence are summarized in Table A.1.

Motion Complexity

Feature	Category	Description	Influence On Complexity (0=Not significant, 1=Low, 2= Medium, 3=High)	Confidence
Total movement percentage (58)	Coverage	The total number of pixels with substantial movement divided by the total number of pixels.		
Right movement percentage (9)	Coverage	The total number of pixels with substantial rightward movement divided by the total number of pixels.		
Left movement percentage (11)	Coverage	The total number of pixels with substantial leftward movement divided by the total number of pixels.		
Right movement kurtosis (45)	Strength	The kurtosis of the rightward movement strength.		
Left movement kurtosis (47)	Strength	The kurtosis of the leftward movement strength.		
Up movement kurtosis (44)	Strength	The kurtosis of the upward movement strength.		
Down movement kurtosis (46)	Strength	The kurtosis of the downward movement strength.		
Maximum movement (52)	Strength	The maximum amount of movement.		
Movement cluster count (59)	Clusters	The number of areas of movement during a given time.		
Movement cluster size (60)	Clusters	The size of the areas of movement during a given time.		
Movement run count (62)	Activity	The number of runs (periods of movement).		
Movement run size (61)	Activity	The average length of the runs (periods of movement).		
Magnitude entropy (55)	Entropy	The entropy of the movement strength.		
Directional change count (63)	Smoothness	The number of significant changes in motion direction.		
Directional average (1)	Direction	The average direction.		
Directional RVL (2)	Direction	The resultant vector length (strength of the average motion direction).		
Directional skewness (4)	Direction	The skewness of the direction.		
Directional Rayleigh test (6)	Direction	The Rayleight test on the directional values (a measure of non uniformity).		

Figure A.1: The form presented to each participant in the expert-polling study.

Feature	**Mean Influence (0-3)**	**Mean Confidence (%)**
Total Movement Percentage	1.73	82%
Right Movement Percentage	1.00	64%
Left Movement Percentage	1.00	64%
Right Movement Kurtosis	1.73	27%
Left Movement Kurtosis	1.73	27%
Up Movement Kurtosis	2.00	45%
Down Movement Kurtosis	2.00	45%
Maximum Movement	1.90	55%
Movement Cluster Count	2.91	82%
Movement Cluster Size	1.36	64%
Movement Run Count	2.18	45%
Movement Run Size	1.36	40%
Magnitude Entropy	2.64	64%
Directional Change Count	2.45	82%
Directional Average	0.64	45%
Directional RVL	1.45	18%
Directional Skewness	1.18	45%
Directional Rayleigh Test	2.27	64%

Table A.1: Summary of the expert poll results.

Appendix B

Complexity Rating Experiment

This chapter details the user study on visual motion complexity presented in Chapter 4 for weighting features based on expert belief and providing labels to videos when training classifiers. A group of 22 people participated in the study. A MATLAB program was written to guide each participant through a set of videos randomly selected from a dataset of spider movements and a dataset of human actions (both described in detail in Appendix C). 25% of the videos were duplicated to assist in measuring participant rating consistency. A detailed set of instructions was initially displayed to the participant, as shown in Figure B.1. The time taken by each participant was about 25 minutes. Inputting the ratings through a one-way ANOVA model revealed which users were not rating duplicate videos accurately. Only one user was indicative of inaccurate rating, and was thus discarded.

Each participant was shown every video from the datasets, then asked to rate the motion shown for each one on a scale of one (low complexity) to ten (high complexity) based on their personal opinion. The video would play through one time initially without allowing the participant to submit a rating, then would play repeatedly until the participant submitted a rating. The participant was also shown the number of

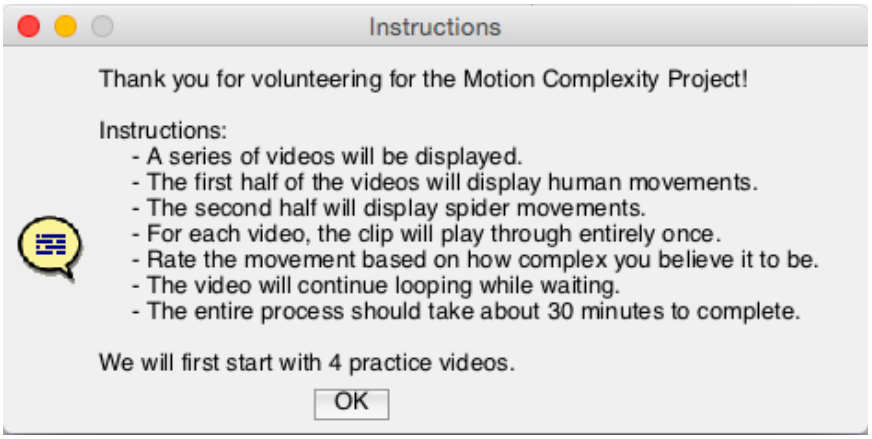


Figure B.1: The initial instruction message presented to each participant to detail the process.

remaining videos. This graphical user interface presented to each participant is shown in Figure B.2. After rating all videos, each participant was shown a final questionnaire asking for beliefs in six identified motion complexity domains. Responses were given on a scale of one (not important) to five (very important). A response was required for all six domains. The graphical user interface displayed for the final questionnaire is shown in Figure B.3.

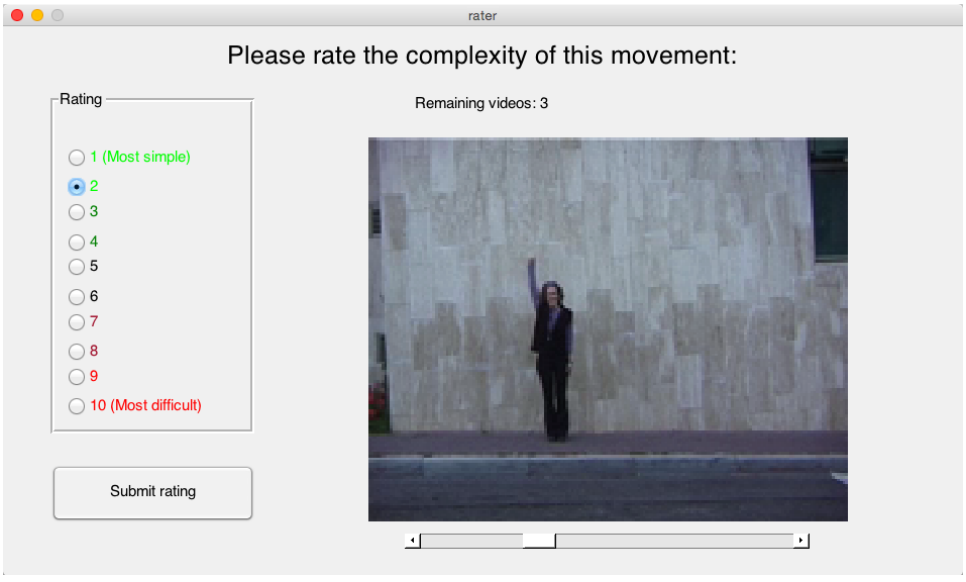


Figure B.2: The complexity rater GUI interface shown to each participant.

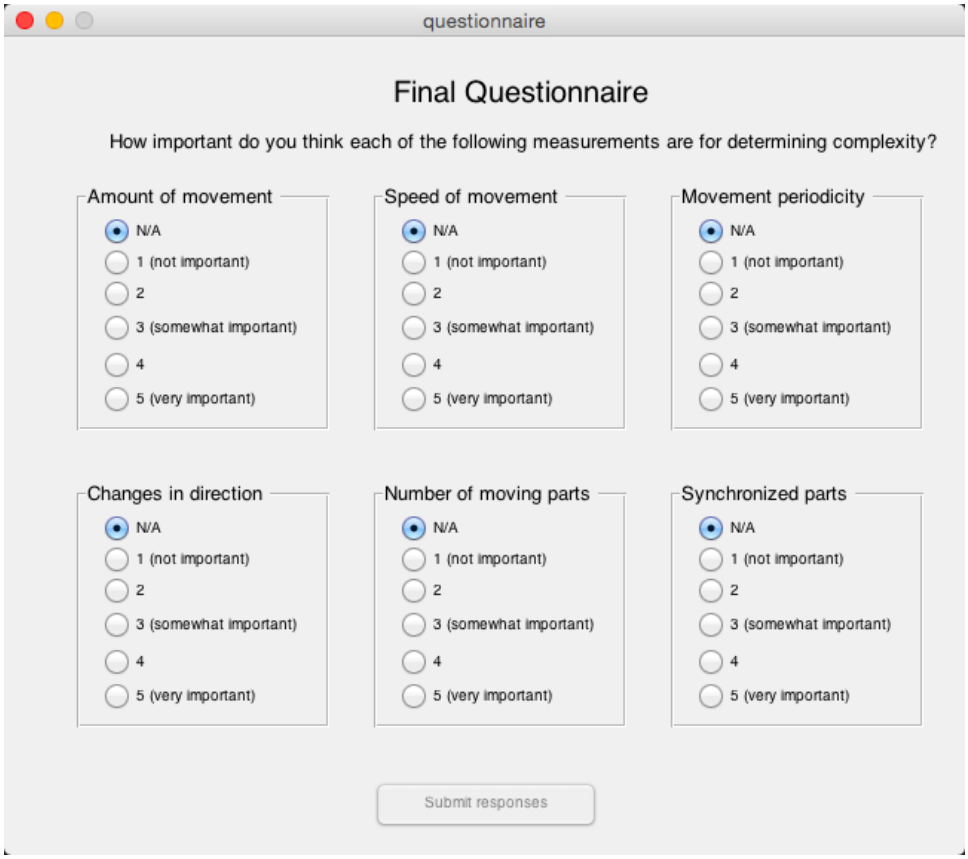


Figure B.3: The questionnaire presented to each participant for obtaining complexity belief of the six motion complexity domains.

Appendix C

Datasets

In this chapter, the two datasets utilized throughout the research are detailed and visualized. The first dataset is a set of videos displaying movements of wolf spiders, while the second dataset displays basic actions of human beings.

C.1 Spider Dataset

This dissertation utilizes a dataset of high frame rate videos containing samples of two species of Schizocosa wolf spider: *S. bilineata* and *S. crassipalpata*. There are 52 total grayscale videos in the dataset, where each video is roughly six seconds in length. The dataset is divided into two halves (one half for each species), while each of those halves is further divided into two (high diet and low diet). The separation of high diet from low diet comes from the expectation that nutrient intake could influence the degree to which spiders can engage in complex courtship displays. Thus, by varying the diet of individuals, we can assess whether there is a link between nutrient intake and courtship complexity. Each video has a temporal resolution of 250 FPS for capturing the quick movements of the spiders, and a varying spatial resolution due to cropping out the

areas of interest in each clip. This dataset is summarized in Table C.1. Samples of videos from the dataset are shown in Table C.2.

Class	**# of Samples**	**Frame Rate**
S. bilineata (high diet)	14	250 FPS
S. bilineata (low diet)	15	
S. crassipalpata (high diet)	10	
S. crassipalpata (low diet	13	
Total	52	

Table C.1: Summary of the spider dataset.

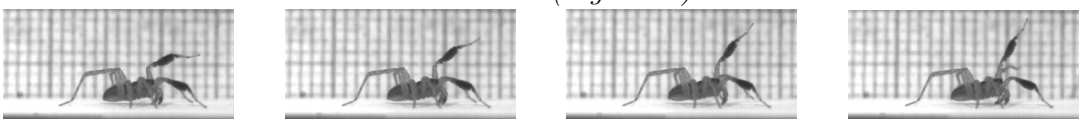
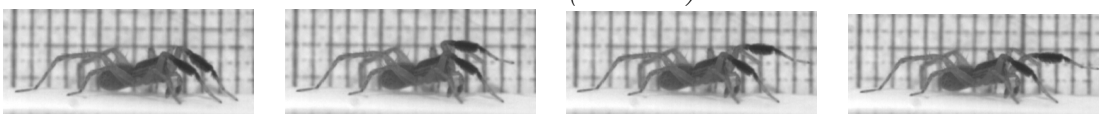
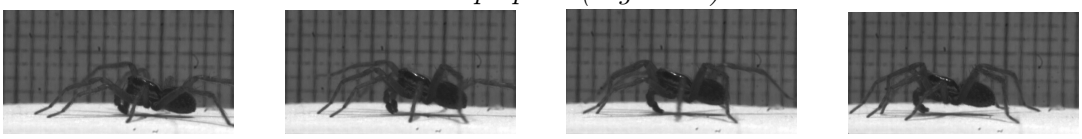
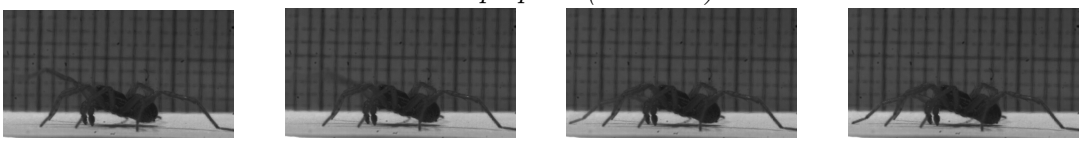
S. bilineata (high diet) 
S. bilineata (low diet) 
S. crassipalpata (high diet) 
S. crassipalpata (low diet) 

Table C.2: Spider dataset samples.

C.2 Human Dataset

This dissertation also utilizes a dataset of standard frame rate videos containing samples of basic human motions. The human action dataset used in this work is the Weizmann dataset [23], a widely used collection of basic human motions for comparing

action classification systems. The videos were recorded with a fixed (non moving) camera. It contains 81 low-resolution (180 × 144) video sequences, recorded at 25 FPS, displaying nine different people performing nine basic actions. The displayed action classes are "running (run)", "walking (walk)", "jumping jack (jack)", "jumping forward on one leg (skip)", "jumping in place on two legs (jump)", "galloping sideways (side)", "waving one hand (1-wave)", "waving two hands (2-wave)", and "bending (bend)". This dataset is summarized in Table C.3. Samples of videos from the dataset are shown in Table C.4.

Motion Class	**# of Samples**	**Frame Rate**
Bend	9	
Jack	9	
Jump	9	
1-wave	9	
Run	9	25 FPS
Side	9	
Skip	9	
2-wave	9	
Walk	9	
Total	81	

Table C.3: Summary of the human dataset.

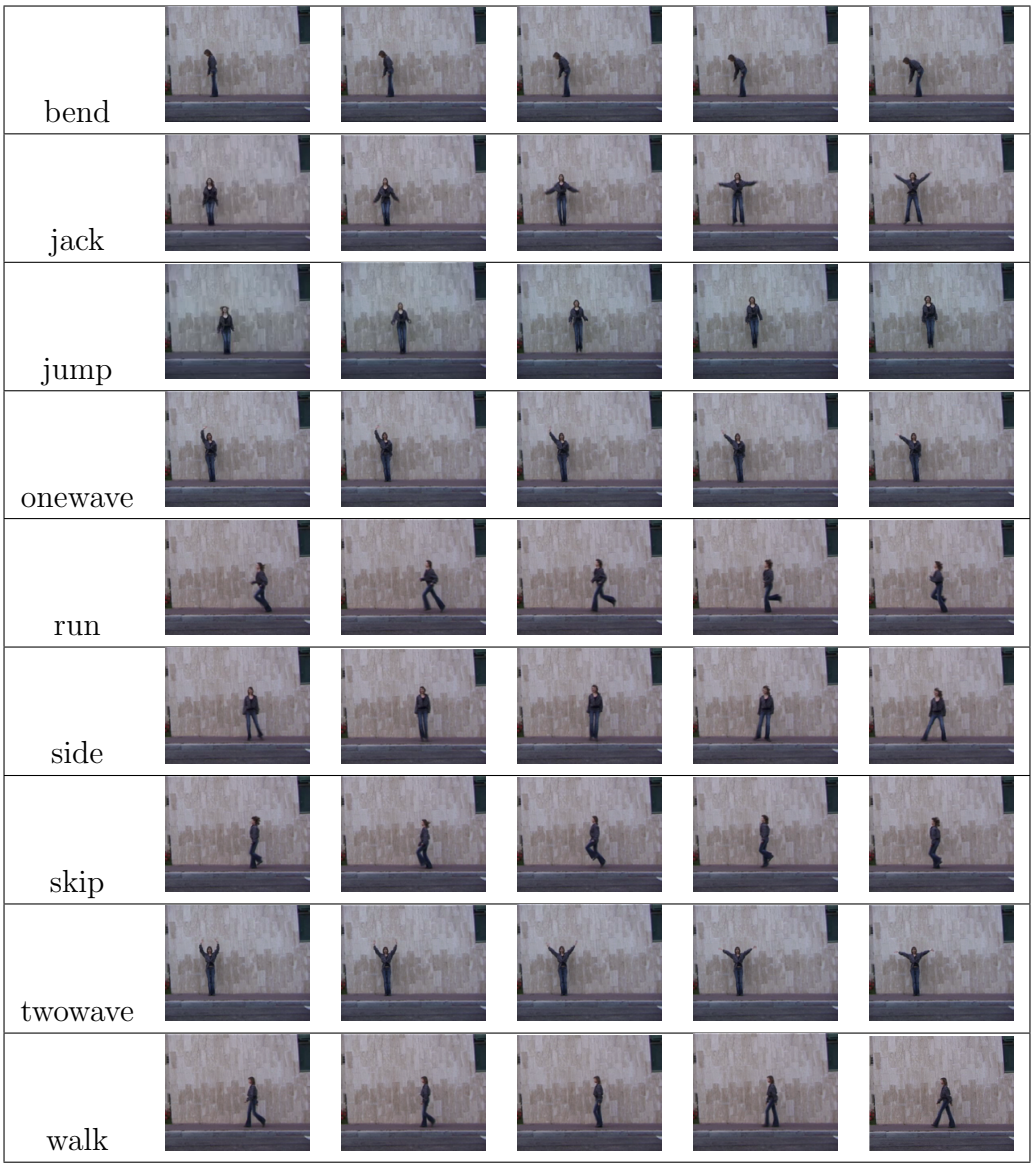


Table C.4: Human dataset samples.