SPATIO-TEMPORAL MAXIMUM AVERAGE CORRELATION HEIGHT TEMPLATES IN ACTION RECOGNITION AND VIDEO SUMMARIZATION

by

MIKEL RODRIGUEZ
B.A. Earlham College, Richmond Indiana
M.S. University of Central Florida

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the School of Electrical Engineering and Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Summer Term
2010

Major Professor: Mubarak Shah

ABSTRACT

Action recognition represents one of the most difficult problems in computer vision given that it embodies the combination of several uncertain attributes, such as the subtle variability associated with individual human behavior and the challenges that come with viewpoint variations, scale changes and different temporal extents. Nevertheless, action recognition solutions are critical in a great number of domains, such video surveillance, assisted living environments, video search, interfaces, and virtual reality.

In this dissertation, we investigate template-based action recognition algorithms that can incorporate the information contained in a set of training examples, and we explore how these algorithms perform in action recognition and video summarization.

First, we introduce a template-based method for recognizing human actions called Action MACH [91]. Our approach is based on a Maximum Average Correlation Height (MACH) filter. MACH is capable of capturing intra-class variability by synthesizing a single Action MACH filter for a given action class. We generalize the traditional MACH filter to video (3D spatiotemporal volume), and vector valued data. By analyzing the response of the filter in the frequency domain, we avoid the high computational cost commonly incurred in template-based approaches. Vector valued data is analyzed using the Clifford Fourier transform, a generalization of the Fourier transform intended for both scalar and vector-valued data.

Next, we address three seldom explored challenges in template-based action recognition. The first is the recognition and localization of human actions in aerial videos obtained from unmanned aerial vehicles (UAVs), a new medium which presents unique challenges due to the small number of pixels per human, pose, and moving camera. The second issue we address is the incorporation of multiple positive and negative examples of a target action class when generating an action template. We address this issue by employing the Fukunaga-Koontz Transform as a means of generating a single quadratic template which, unlike traditional temporal templates (which rely on positive examples alone), effectively captures the variability associated with an action class by including both positive and negative examples in the template training process.

Third, we explore the problem of generating video summaries that include specific actions of interest as opposed to all moving objects. In doing so, we explore the role of action templates in video summarization in an effort to provide a means of generating a compact video representation based on a set of activities of interest. We introduce an approach [90] in which a user specifies the activities that interest him and the video is automatically condensed to a short clip which captures the most relevant events based on the user's preference. We follow the output summary video format of non-chronological video synopsis approaches, in which different events which occur at different times may be displayed concurrently, even though they never occur simultaneously in the original video. However, instead of assuming that all moving objects are interesting, priority is given to specific activities of interest which pertain to a user's query. This provides an efficient means of browsing through large collections of video for events of interest.

To my loving parents and wife Nancy

ACKNOWLEDGMENTS

I am very grateful to Dr. Mubarak Shah for all of his support and advice over the past years. I feel very fortunate to be part of Dr. Shah's research group and for having had the opportunity to learn from and be involved in a wide range of projects (FCS, FDOT, VACE, Harris, VIRAT) during my time with the group. I would like thank Dr. Niels Lobo, Dr. Fernando Gomez and Dr. Raymond Surette, for serving as my PhD committee members and for their valuable comments. I thank my wife, Nancy, and my parents for their love and constant encouragement. I would also like to thank my colleagues, many of which I have collaborated with over the years, including Dr. Saad Ali, Vladimir Reilly, Ramin Mehran, Dr. Jingen Liu and several others.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Recognizing human activities constitutes one of the most challenging problems in computer vision, yet it is widely recognized that effective solutions capable of recognizing actions in uncontrolled environments could lend themselves to a host of important application domains, such as video indexing, surveillance, human-computer interface design, analysis of sports videos, and the development of intelligent environments. In the context of video surveillance, robust action recognition constitutes an essential capability which can improve upon current manual inspection processes. Although video surveillance systems are already prevalent, videos recorded by these surveillance systems are usually only recorded to be used in a post-factum manner, thereby losing an important benefit as an active real-time warning. Action recognition systems which are both robust and efficient will likely have a great impact on the transition of video surveillance from a forensic tools that are used after the fact to active crime prevention systems.

Within human-computer interfaces, action recognition can prove to be a useful extension to existing speech-based control systems. Action recognition would allow for more detailed visual cues that can be received through action and gesture recognition as well as facial action unit classification. Robust methods for recognizing human motion patterns can also be used as a means of providing automatic sign-language translation between agents and signaling specific instructions in high-noise environments.

Another area that could benefit from robust human action recognition is virtual reality and gaming. In particular, the ability to automatically recognize actions in real-world environments would allow for more streamlined form of interaction with other agents by automatically capturing human actions and facial action units as cues. Similarly, along these same lines, action recognition can serve as an interface in computer and console games, teleconferencing and simulation and training.

1.1 Challenges

The problem of action recognition is one of great difficulty given that it represents the combination of several uncertain attributes, namely the subtle variability associated with individual human behavior and the challenges that come with viewpoint variations, scale changes, execution rate, as well as anthropometric attributes that vary with age and gender. All of these factors coupled with the range and complexity of human actions renders developing human action recognition algorithms a difficult challenge.

Several standard action datasets capture the aforementioned challenges. Traditional action datasets are captured under controlled environments, where some of variability (such as viewpoint) can be controlled. Examples of this class of datasets are the KTH dataset and the Weizmann action dataset (Figure 1.1). In these datasets we observe a wide range of actors, each of which interprets a particular action class with differing execution rates. Another collection of action examples captured in controlled environments is the IXMAS dataset (Figure 1.2).

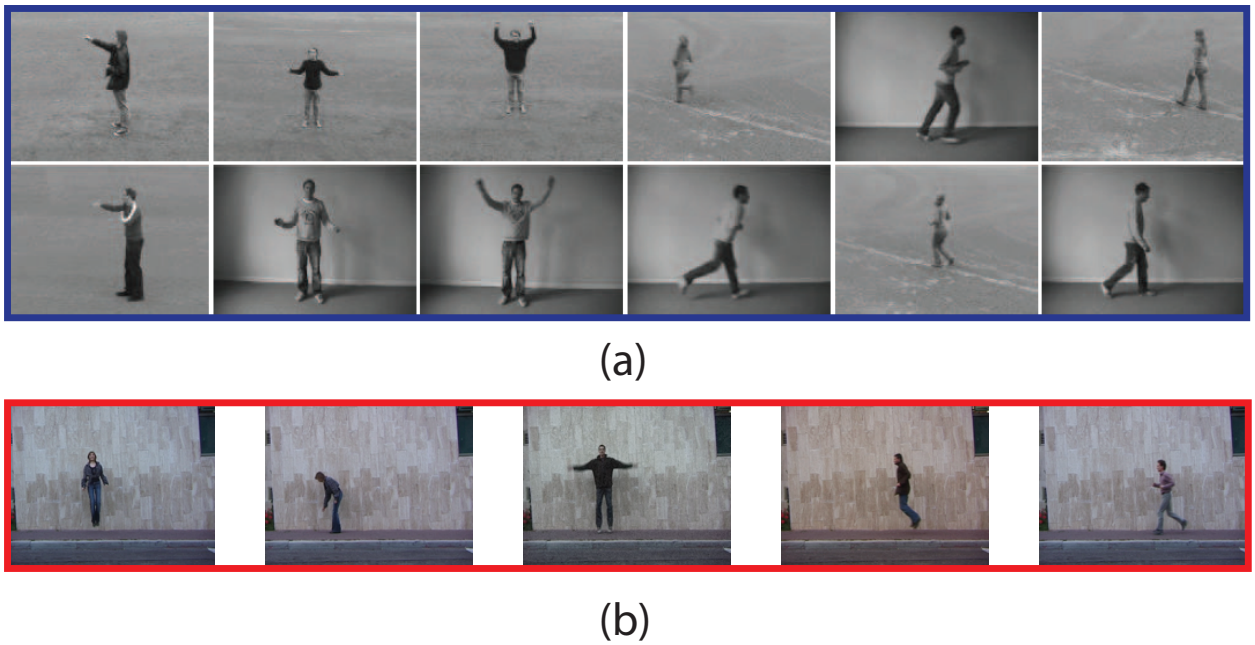


Figure 1.1: Instances of human actions in different traditional action recognition datasets captured in controlled environments. The KTH action dataset (top row), Weizmann Action Dataset (bottom row).

Despite the fact that recognizing actions in controlled environments remains a challenging problem, recently research on action recognition has moved toward more unconstrained and realistic action datasets. In these datasets we encounter most of the challenges and variability associated with human actions.

Figure 1.3 demonstrates some action examples from UCF's Sports Action Dataset, the UCF aerial actions dataset and UCF's feature film action dataset which consists of unconstrained videos collected from archive footage web sites, moving platforms and feature films. Generally, these unconstrained action datasets contain significant camera motion, complicated and cluttered backgrounds, as well as variation in sales, and viewpoints.

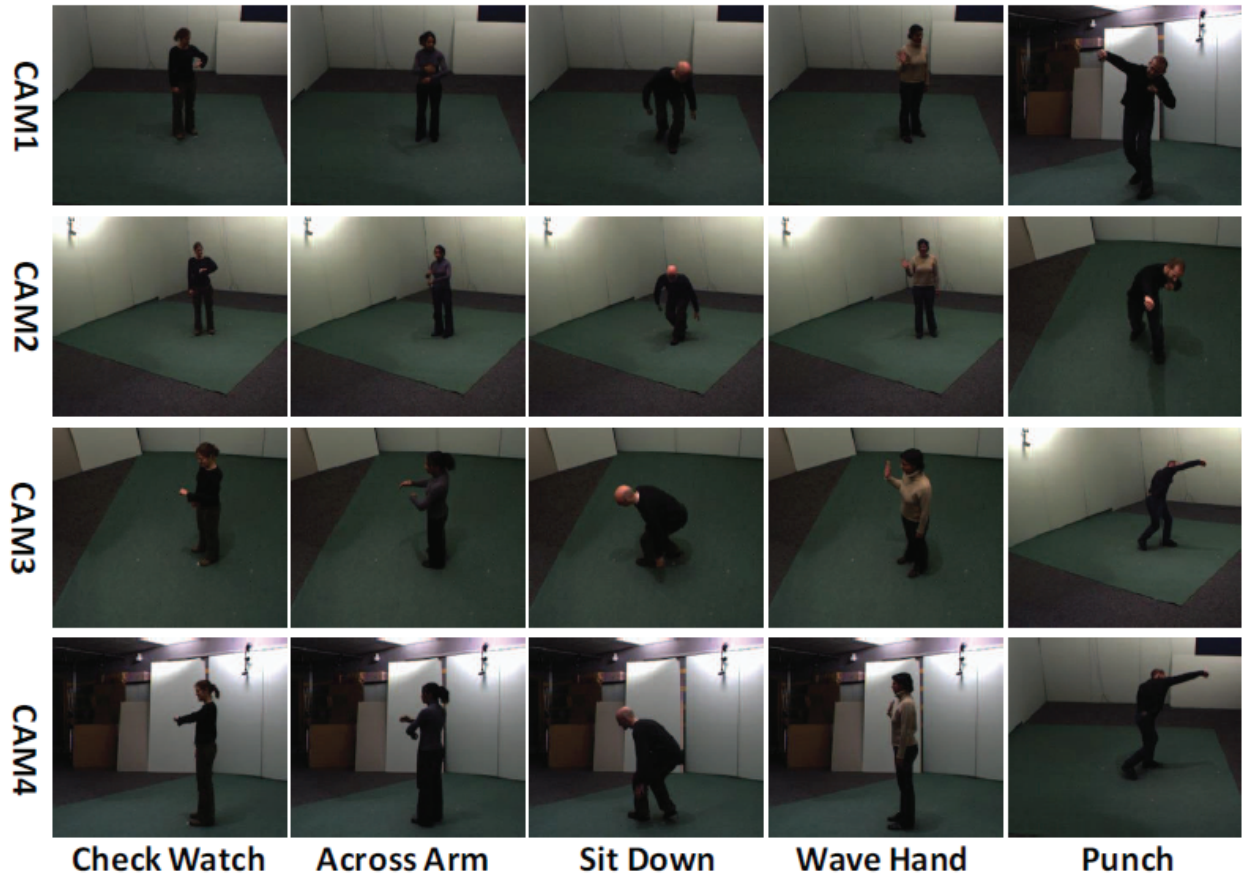


Figure 1.2: Examples of four views of some actions present in the IXMAS dataset.

Unlike the standardized performance of actors in traditional constrained action datasets, actions found in broadcast television sports (Figure 1.3-a) and feature films (Figure 1.3-c) tend to occur over cluttered backgrounds and usually display wide variation in camera pose as well as different interpretations of the same action by different actors.

Although actions featured in broadcast television and movies tend to contain complex background and a wide range of intra-class capability, actors are usually the center of focus and take up a significant number of pixels of each frame. Furthermore, cameras are usually placed at ground-

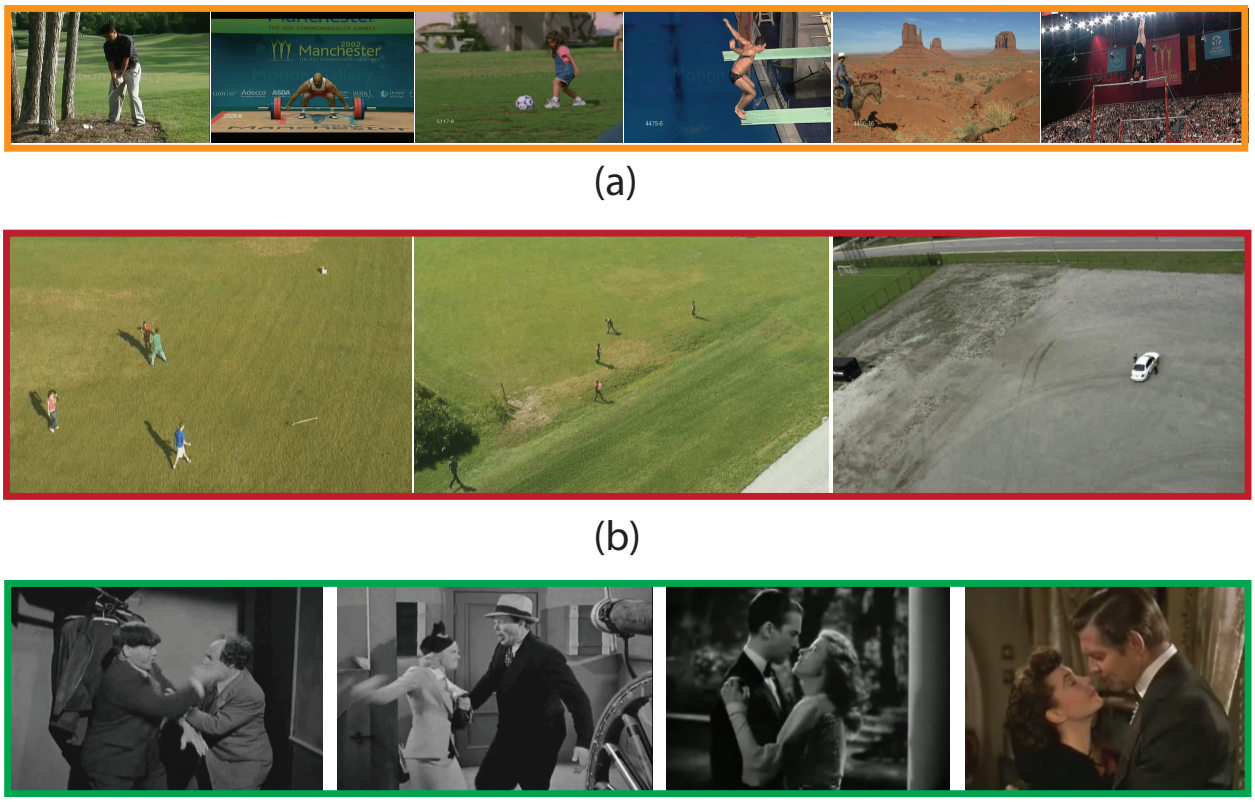


Figure 1.3: Unconstrained action datasets. The UCF sports dataset (top row), the UCF aerial actions dataset (middle row, and the UCF feature films dataset (bottom row).

level and sensor motion tends to be minimal and smooth. These are factors which are usually not present in aerial videos. As can be seen in Figure 1.3-b, the main challenges associated with recognizing actions in aerial videos reside in the fact that articulated human motions are very difficult to observe at high altitudes. Depending on the flight pattern of the mobile platform viewpoint can drastically change the appearance of an action (Figure 1.4). Other issues, such as moving camera and low number of pixels on target (which typically averages less than 8 pixels wide and 15 pixels tall) also factor in making this a complex problem.

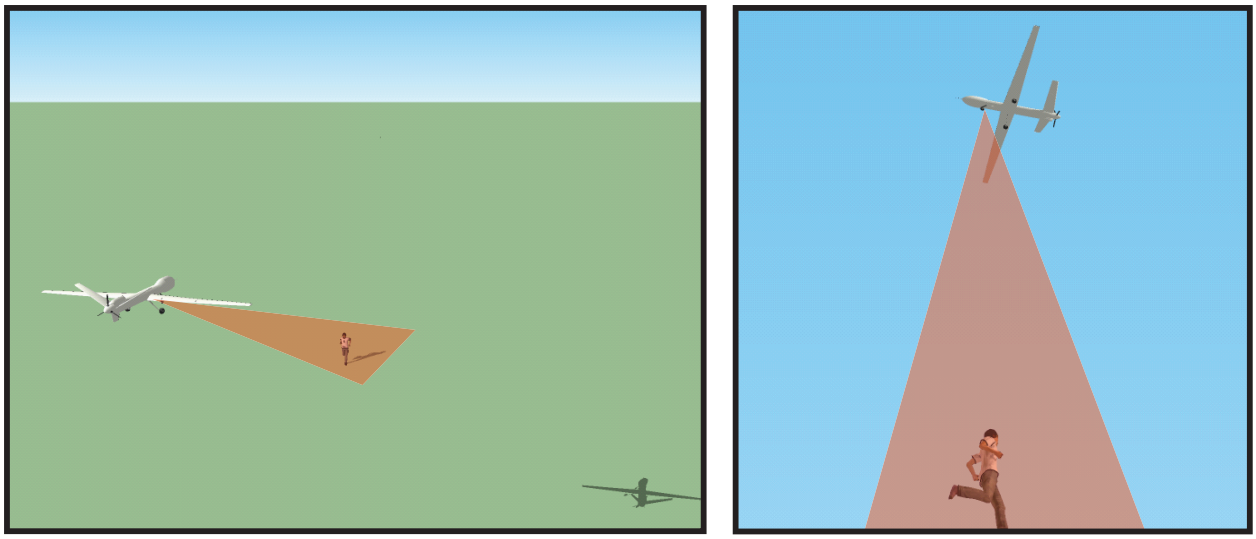


Figure 1.4: One of the challenges of aerial action recognition includes the wide range of viewpoints. (a) Low oblique mobile plarform angle, (b) Nadir perspective

1.2 Objective

One of the earliest solutions to action recognition that emerged from the computer vision community was the use of temporal template matching. In this class of approaches a template of an action is matched with a sub-region of a video. A gamut of approaches which fall under this general denomination has been proposed over the years.

In this dissertation we address three unexplored challenges in template-based action recognition. The first challenge we address is the ability to generalize from positive exemplars of an action class in order to effectively generate a single action template which captures the intra-class variability of an action using a collection of examples. Recent popular action recognition methods which employ machine learning techniques such as SVMs and AdaBoost, provide one possibility

for incorporating the information contained in a set of training examples. However, learning from multiple labeled examples has not been addressed in template-based action recognition methods.

As we explore algorithms that enable us to learn templates from multiple examples, we will focus our training and testing on unconstrained real-world datasets. Unlike traditional action recognition datasets, which consist of short testing and training video clips that are only a few seconds long, in this dissertation we attempt to approximate a more realistic scenario. Therefore, we explore the use of raw, unsegmented videos sequences, which are typically several minutes in length and contain multiple instances of different action classes occurring concurrently at any moment.

The second issue we intend to address in this dissertation is the study of alternative correlation filter paradigms that are capable of explicitly incorporating not only positive samples of an action class but also negative exemplars. We propose a new quadratic spatio-temporal action template framework which, unlike traditional template-based approaches, generalizes from multiple examples of positive instances of an action class as well as a set of negative examples (clutter) present in a dataset. The quadratic correlation filter will be composed of a set of dominant bases in the feature domain which contain maximal information about a given action class, while at the same time containing minimal information about the negative class. In order to test this new approach we propose an extensive set of experiments on unconstrained aerial videos.

Finally, a third issue we address pertains to exploring the role of template-based action recognition in compact video representations. Currently, most of the work on activity recognition has focused mainly on detection in short pre-segmented video clips commonly found in publicly available action datasets. In this dissertation, we attempt to move beyond only performing action de-

tection in an effort to understand how template-based action recognition methods can provide as basis for generating compact video representation based on a set of activities of interest, while preserving the scene dynamics of the original video.

1.3 Contributions

In this dissertation we explore the role of spatio-temporal templates in action recognition of both aerial and ground videos, and we study how action templates can employed to generate action-specific video summaries. Unlike the traditional methods of generating action templates, which typically don't generalize from a number of labeled examples, we address the ability to effectively generate a single action template which captures the general intra-class variability of an action by using a collection of examples. In this work we generalize the traditional MACH filter to operate on spatiotemporal volumes as well as vector valued data by embedding the spectral domain into a domain of Clifford algebras. Before the detailed discussion, we would like to summarize our major contributions in this dissertation:

- We propose a new template-based action recognition approach that is capable of incorporating multiple labeled example to generate a single template that is capable of capturing the variability associated with an action class.

- We extend traditional MACH filters to include vector-valued data by employing the Clifford Fourier Transform, which extents the traditional Fourier transform by including vector-valued data.

- Unlike most action recognition approaches, we focus our study of action templates and testing on a collection of new datasets recorded in uncontrolled environments. All of these datasets contain multiple actors performing actions concurrently and are not limited to temporally short segments.

- We introduce activity-specific video summaries, which provide an effective means of browsing and indexing video based on a set of events of interest. Our method automatically generates a compact video representation of a long sequence, which features only activities of interest while preserving the general dynamics of the original video.

1.3.1 A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition

The first algorithm developed in this dissertation generalizes the traditional MACH filter to operate on spatiotemporal volumes as well as vector valued data. Traditionally, MACH filters have been employed mainly in object classification using scalar data. In this approach a two dimensional template is generated such that it expresses the general shape or appearance of an object. In our work we generalize the MACH filter by synthesizing a template estimated from the spatio-temporal volumes of action sequences. We show that such filters can be trained using raw pixel values, edges, temporal derivative, or optical-flow in the spatiotemporal volume. When each pixel in this volume contains multiple values it is not possible to synthesize a MACH filter using traditional Fourier transform. In order to deal with this problem, we propose to employ the Clifford transform, which is a generalization of the standard Fourier transform for vector valued functions.

The process of incorporating multiple examples to generate a single Action MACH filter for a given action class begins with the creation of a series of spatio-temporal volumes from the testing action sequences by concatenating the frames of a complete cycle of an action. Subsequently, we compute features which can be either scalar or vector-based such as optical flow, resulting in a volume for each training sequence. Following the construction of the spatio-temporal volumes for each action in the training set, we proceed to represent each volume in the frequency domain by performing a 3-D FFT operation if the features are scalars or a 3-D CFT if the features are vectors.

Once the feature volumes are created and represented in the frequency-domain we proceed to convert the resulting 3-D FFT matrix into a column vector by concatenating all the columns of the 3-D matrix. Once the column vectors are obtained for all the examples of the action, the Action MACH filter (which minimizes average correlation energy, average similarity measure, output noise variance; and maximizes average correlation height) can be synthesized in the frequency domain. Once an Action MACH filter has been synthesized, we can proceed to detect similar actions in a testing video sequence by correlating the template with an input sequence, normalizing the resulting correlation space and applying a threshold.

1.3.2 Aerial Action Recognition Using Spatio-temporal Quadratic Correlation Filters

The second algorithm developed in this dissertation addresses three seldom explored challenges in template-based action recognition. The first is the recognition and localization of human actions in aerial videos obtained from unmanned aerial vehicles (UAVs), a new medium which presents unique challenges due to the small number of pixels per human, pose, and moving camera. The

second issue we address is the incorporation of multiple positive and negative examples of a target action class when generating an action template. We address this issue by employing the Fukunaga-Koontz Transform as a means of generating a single quadratic template which, unlike traditional temporal templates (which rely on positive examples alone), effectively captures the variability associated with an action class by including both positive and negative examples in the template training process. Finally, we explore a range of low-level and mid-level motion features and assess their effectiveness within the context of aerial action recognition by introducing a first-of-its-kind dataset which features actions captured from a UAV at different flying altitudes.

1.3.3 Compact Representation of Actions in Movies

The third problem we address is the role of action templates in video summarization. In our approach [90], a user specifies which activities interest him and the video is automatically condensed to a short clip which captures the most relevant events based on the user's preference. We follow the output summary video format of non-chronological video synopsis approaches, in which different events which occur at different times may be displayed concurrently, even though they never occur simultaneously in the original video. However, instead of assuming that all moving objects are interesting, priority is given to specific activities of interest which pertain to a user's query. This provides an efficient means of browsing through large collections of video for events of interest.

Our approach to generating compact action-specific video representations is composed of three main phases. First, we begin by determining a set of regions in space-time which contain dynamic

objects. Subsequently, we detect specific activities and actions of interest within the long video sequence. Finally, we select dynamic regions which contain events of interest and optimize the temporal shifts of the video summary via an energy minimization. In the subsequent chapters we describe each of these in more detail.

1.4 Organization of Dissertation

The rest of the dissertation is organized as follows. In Chapter 2 we cover relevant research in the related areas. We also detail how the proposed methods contribute to the literature in relation of the previous work. Chapter 3 presents the Action MACH filter, a template-based method for action recognition which is capable of capturing intra-class variability by synthesizing a single Action MACH filter for a given action class. In this chapter we generalize the traditional MACH filter to video (3D spatiotemporal volume), and vector value data. In this chapter we also describe an extensive set of experiments on both publicly available datasets such as the KTH action dataset, the Weizmann action dataset, the Cohn-Kanade facial expression database, and on two new homegrown datasets. These include a collection of sports-related actions as featured on broadcast television channels, and a pool of actions found in feature films.

In Chapter 4 we discuss recognizing actions in aerial videos by including both positive *and* negative examples of an action class in order to obtain a quadratic correlation filter that generalizes the variability associated with an action. We address this issue by employing the Fukunaga-Koontz Transform as a means of generating a single quadratic template which, unlike traditional temporal templates that rely only on positive examples, effectively captures the variability associated with

an action class by including both positive and negative examples in the template training process. We present an extensive set of experiments that explore a range of low-level and mid-level motion features and assess their effectiveness within the context of aerial action recognition and quadratic correlation filters.

Chapter 5 presents our study of the role of template-based action recognition in video synopsis. We introduce activity-specific video summaries, which provide an effective means of browsing and indexing video based on a set of events of interest. In this chapter we describe a new approach to summarizing video which automatically generates a compact video representation of a long sequence, that features only activities of interest while preserving the general dynamics of the original video.

CHAPTER 2: RELATED WORK

In this chapter we review a number of relevant methods in the literature related to correlation filters, activity recognition, and video synopsis. We describe the advantages and drawbacks of these approaches. We also describe where our work fits within the context of the current literature.

2.1 Human Activity Recognition

Action recognition and event classification in video have been studied extensively in Computer Vision; a comprehensive review can be found in recent surveys on the topic [65, 1]. Many of the existing methods can be categorized based on the type of representation employed by the approach. The most common representations in action recognition are: bag of words representations, feature-based representations, shape-based, abstract body models, and temporal-templates. In this section we discuss each of these leading representations of actions in videos.

Appearance-based models were a common class of representation in the first generation of action recognition approaches [17, 26, 100, 110, 33]. A common approach within this class of representations is based on learning the visual model of specific postures of the human body which are then matched with testing video sequences [68, 67, 43]. A major drawback of this set of static methods is the loss of temporal information which is essential to recognizing complex activities.

A common approach used to capture the temporal dynamics and interdependencies of events involves training some variant of a Hidden Markov Model (HMM, CRF, BRF). Nevertheless, despite the fact that these methods do encode temporal transitions across the extent of a video, appearance-based representation have always been found to be limited in their robustness in the presence of unconstrained videos as they are sensitive to background noise, and changes due to clothing.

Another important paradigm is that of shape-based representations of actions, which are also related to the appearance-based models. In this class of approaches actions are modeled as a series of postures which are detected via edges or silhouettes in a given frame [20, 24].

The concept of using silhouettes in a frame to encode posture has also been generalized to represent the spatio-temporal evolution of the outline of an actor's body [113, 9]. This is done layering consecutive silhouettes from each frame leading to a 3D volume. In general, this class of approaches is focussed on modeling the overall form of an action and do not consider motion information in their representation.

Another category of volume-based approaches focusses on computing spatio-temporal features. Laptev *et al.* [61] and Dollar *et al.* [27] focus on spatio-temporal interest points, and Ke *et al.* [47] define a set of spatio-temporal difference operators and learn a cascade of boosted classifiers. The advantage of this class of approaches lies in its ability to handle partial occlusions given that the entire silhouette of a human need not be tracked.

2.1.1 Correlation Filters

Temporal template matching emerged as an early solution to the problem of action recognition, and a gamut of approaches which fall under this general denomination has been proposed over the years. Early advocates for approaches based on temporal matching, such as Polana and Nelson [83], developed methods for recognizing human motions by obtaining spatio-temporal templates of motion and periodicity features from a set of optical flow frames. These templates were then used to match the test samples with the reference motion templates of known activities. Essa and Pentland [32] generated spatio-temporal templates based on optical flow energy functions to recognize facial action units. Bobick et al [14, 13] computed Hu moments of motion energy images and motion-history images to create action templates based on a set of training examples which were represented by the mean and covariance matrix of the moments. Recognition was performed using the Mahalanobis distance between the moment description of the input and each of the known actions. Recently, Weinland *et al.* [106] extended this representation to handle different viewpoints.

The drawback of this early generation of temporal templates based purely on appearance is that movements performed by different subjects can lead to very different intensity patterns over time, regardless of whether these movements belong to the same action class. Factors such as clothing color and style render the use of appearance as a means for classifying actions ineffective. Therefore, numerous approaches which advocate for optical-flow-based temporal templates [52, 31, 96] have been proposed which leverage the underlying flow fields induced in the presence of different actions.

Efros et al. [31] proposed an approach to recognizing human actions at low resolutions which consisted of a motion descriptor based on smoothed and aggregated optical flow measurements over a spatio-temporal volume centered on a moving figure. This spatial arrangement of blurred channels of optical flow vectors is treated as a template to be matched via a spatio-temporal cross correlation against a database of labeled example actions.

In order to avoid explicit computation of optical flow, a number of template-based methods attempt to capture the underlying motion similarity amongst instances of a given action class in a non-explicit manner. Shechtman and Irani [96] avoid explicit flow computations by employing a rank-based constraint directly on the intensity information of spatio-temporal cuboids to enforce consistency between a template and a target. Given one example of an action, spatio-temporal patches are correlated against a testing video sequence. Detections are considered to be those locations in space-time which produce the most motion-consistent alignments.

Given a collection of labeled action sequences, a disadvantage of these methods is their inability to generalize from a collection of examples and create a *single* template which captures the intra-class variability of an action. Effective solutions need to be able to capture the variability associated with different execution rates and the anthropometric characteristics associated with individual actors. Recent popular methods which employ machine learning techniques such as SVMs and AdaBoost, provide one possibility for incorporating the information contained in a set of training examples.

2.2 Video Summary

A common theme in all of the approaches mentioned above is their focus on detection. That is, given a learned model of an action class, emphasis is placed on detecting instances of the learned action within small testing clips typically found in standard action datasets. After performing detection, most methods do not go beyond placing a bounding box delimiting the spatio-temporal extent of the detected action. Our present work aims at moving beyond detection by examining the role of action recognition in efficient video representations. More specifically, we are interested in the generation of compact video summaries which contain specific actions of interest.

A very common approach employed for video summarization is based on some form of fast forwarding, where multiple frames are skipped during periods in which no moving objects are present. A special case of this class of approaches are the time-lapse methods which tend to depict an event that occurs across a long temporal span in a short period of time. Given that fast forwarding frames at this type of global level can lead to the loss of important events that occur quickly (like running or opening a door), several researchers have suggested multi-level fast forward approaches [71, 81]. This class of approaches attempt to skip frames which correspond to periods of no interest based on criteria such as sound, color and camera motion.

Another leading framework for video summarization uses still images as the basis of representation. A common method is the use of collection of select key frames which represent typical instances of a long video [53, 116]. Another group of approaches which employ still images as the basis of their representation focus on mosaic images as a compact depiction of a video [84, 75].

One of the main drawbacks of this set of static approaches resides in the fact that one losses the temporal flow of a video.

Recently, a new class of video summarization approaches has emerged which is focused on generating non-chronological video synopses [87, 88, 86]. In most of these methods moving objects are segmented and then shifted along the time axis of the video in order to obtain a compact representation of a long video which may contain concurrent events that don't occur simultaneously in the original video. Most of these approaches are geared towards providing a compact representation of a video as a whole and do not distinguish between different classes of events. Therefore most of the existing approaches are best suited for uncrowded scenes that contain periods of inactivity and where events are sparse.

We are interested in a compact representation of long videos which is based on specific actions of interest. Therefore, it may not be appropriate to rely on static, frame-based or mosaic-based representations of video, given that important events and actions which can only be distinguished upon inspecting a sequence of frames are lost in these static representations. Furthermore, generating a compact video representation based on all moving objects in the scene may lead to the inclusion of irrelevant moving objects. This is especially true in crowded scenes were moving objects abound. Therefore, in this work, we explore the role of action recognition as a means of generating condensed representations of long videos which can efficiently convey only important events and actions of interest which occur over a long period of time.

2.3 Summary

We described some of the most relevant research within the context of action recognition and video summarization. We discussed the drawbacks and advantages of various action recognition and video summarization approaches. We described a set of methods which are geared towards addressing some of the limitations of existing methods. Our approach for template-based action detection is based on learning from a number of labeled examples, it extends traditional maximum average correlation height filters to actions using both scalar and vector training data. We have also presented a novel approach to represent long video sequences by incorporating human activity detection as a means of generating a compact representation of the original sequence that depicts important events in the form of a short non-chronological video synopsis. In subsequent chapters we describe the details of our methods for template-based action detection and action-based video synopsis.

CHAPTER 3: ACTION MACH

3.1 Introduction

In this chapter, we introduce the Action MACH filter [91], a template-based method for action recognition which is capable of capturing intra-class variability by synthesizing a single Action MACH filter for a given action class. We generalize the traditional MACH filter to video (3D spatiotemporal volume), and vector value data. By analyzing the response of the filter in the frequency domain, we avoid the high computational cost commonly incurred in template-based approaches, thereby reducing detection to a matter of seconds. Vector-valued data is analyzed using the Clifford Fourier transform, which is a generalization of the traditional scalar-valued Fourier transform. In order to assess the effectiveness of the proposed approach we perform an extensive set of experiments on both publicly available datasets such as the KTH action dataset, the Weizmann action dataset, the Cohn-Kanade facial expression database, and on two new homegrown datasets. These include a collection of sports-related actions as featured on broadcast television channels, and a pool of actions found in feature films.

Traditionally, MACH filters have been employed in object classification, palm print identification [35], and aided target recognition problems [115, 99]. Given a series of instances of a class, a MACH filter combines the training images into a single composite template by optimizing four performance metrics: the Average Correlation Height (ACH), the Average Correlation Energy

Figure 3.1: Our framework is capable of recognizing a wide range of human actions under different conditions. Depicted on the left are a set of publicly available datasets which include dancing, sport activities, and typical human actions such as walking, jumping, and running. Depicted on the right column are examples of two action classes (kissing and slapping) from a series of feature films.

(ACE), the Average Similarity Measure (ASM), and the Output Noise Variance (ONV). This procedure results in a two dimensional template that may express the general shape or appearance of an object. Templates are then correlated with testing sequences in the frequency domain via a FFT transform, resulting in a surface in which the highest peak corresponds to the most likely location of the object in the frame.

The notion of a traditional MACH filter could be generalized to encompass human actions in a number of ways. A fairly straightforward approach would be to recognize an action class by a succession of two dimensional MACH filters at each frame. However, in order to fully leverage the information contained in a video sequence, the approach we propose in this work consists of generalizing the MACH filter by synthesizing a template estimated from the spatio-temporal volumes of action sequences. Such filters could be synthesized using raw pixel values, edges, temporal derivative, or optical-flow in the spatiotemporal volume. When each pixel in this volume contains multiple values it is not possible to synthesize a MACH filter using traditional Fourier transform. Solutions to this problem could include employing motion magnitude or direction alone (scalar values), instead of complete vector data. In order to deal with this problem, we propose to employ the Clifford transform, which is a generalization of the standard Fourier transform for vector valued functions.

3.1.1 Action MACH Filter for Scalar Data

In this subsection we describe the process of synthesizing an Action MACH filter to to recognize various actions based on scalar data. A typical example of a set of actions which we attempt to recognize is depicted in Figure 4.1. These consist of a set publicly available datasets, as well as a collection of actions featured on sports networks and in feature films.

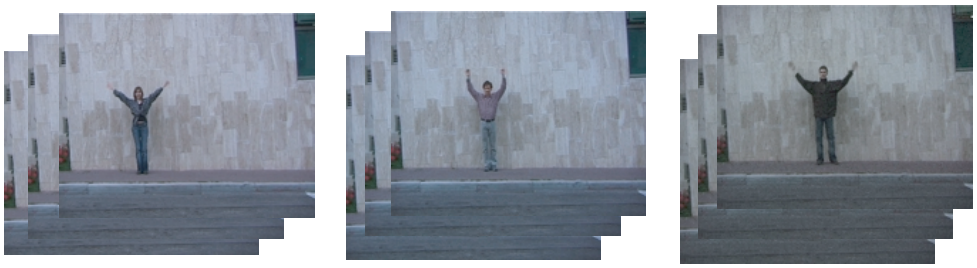
We begin the process of training the Action MACH filter with the creation of a series of spatio-temporal volumes from the testing action sequences by concatenating the frames of a *single* complete cycle of an action. Subsequently, we compute the temporal derivative of each pixel resulting

in a volume for each training sequence. Following the construction of the spatio-temporal volumes for each action in the training set, we proceed to represent each volume in the frequency domain by performing a 3-D FFT operation, which is given by:

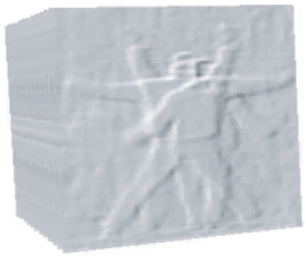
$$F(u,v,w) = \sum_{t=0}^{N-1}\sum_{y=0}^{M-1}\sum_{x=0}^{L-1} f(x,y,t) \exp(-j2\pi(\frac{uv}{L} + \frac{vy}{M} + \frac{wt}{N})), \tag{3.1}$$

where $f(x,y,t)$ is the volume corresponding to the temporal derivative of the input sequence, and $F(u,v,w)$ is the resulting volume in the frequency-domain. L is the number of columns, M the number of rows, and N the number of frames in the example of the action. In an effort to increase the efficiency of this step, we exploit the separability property of the Fourier transform, and compute the multi-dimensional transform by performing one-dimensional transforms in x (horizontal axis), y (vertical axis), and finally t (time axis). Having obtained the resulting volumes in the frequency-domain we proceed to convert the resulting 3-D FFT matrix into a column vector by concatenating all the columns of the 3-D matrix. Let the resulting single column-vector be denoted by x_i (of dimension $d = L * M * N$), where $i = 0, 1, 2..., N_e$, where N_e is the total number of examples of the action in the training dataset. Once the column vectors are obtained for all the examples of the action, the Action MACH filter (which minimizes average correlation energy, average similarity measure, output noise variance; and maximizes average correlation height) can be synthesized in the frequency domain, as follows [99]:

$$h = (\alpha C + \beta D_x + \gamma S_x)^{-1} m_x, \tag{3.2}$$



(a)



(b)

Figure 3.2: The 3D Action MACH filter (b) synthesized for the "Jumping Jacks" action (a) in the Weizmann action dataset using temporal derivatives.

where m_x is the mean of all the x_i vectors, and h is the filter in vector form in the frequency domain. C is the diagonal noise covariance matrix of size $d \times d$, where d is the total number of elements in x_i vector. If the noise model is not available, we can set $C = \sigma^2 I$, where σ is the standard deviation parameter and I is a $d \times d$ identity matrix. D_x is also a $d \times d$ diagonal matrix representing the average power spectral density of the training videos and is defined as:

$$D_x = \frac{1}{N_e} \sum_{i=1}^{N_e} X_i^* X_i, \tag{3.3}$$

where X_i is a $d \times d$ diagonal matrix in which the diagonal elements are the same as the elements of the x_i vector, and $*$ represents the conjugate operation. S_x is the diagonal average similarity

matrix defined as:

$$S_x = \frac{1}{N_e}\sum_{i=1}^{N_e}(X_i - M_x)^*(X_i - M_x), \tag{3.4}$$

where M_x is a diagonal matrix whose elements are the same as those in m_x. Finally, α, β, and γ are the parameters that can be set to obtain the trade-off among the performance measures.

After designing the 1-D filter h, we assemble a complete filter by applying the reverse of the operation that was used to convert the 3D volume of the action example into a column vector. Subsequently, we perform the 3D inverse Fourier transform. The resulting matrix constitutes the Action MACH filter, H, for the particular action (Figure 3.2).

3.1.1.1 Action Classification

Once an Action MACH filter has been synthesized, we can proceed to detect similar actions in a testing video sequence by applying the action MACH filter H to the video:

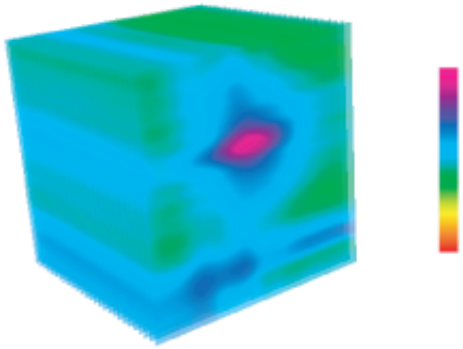
$$c(l,m,n) = \sum_{t=0}^{N-1}\sum_{y=0}^{M-1}\sum_{x=0}^{L-1} s(l+x, m+y, n+t)H(x,y,t), \tag{3.5}$$

where s is the spatio-temporal volume of the test video, H is the spatio-temporal MACH Filter (h is its Fourier transform). P, Q, and R are the dimensions of the of the spatio-temporal volumes.

As a result of this operation, we obtain a response, c, of size $(P-L+1)\times(Q-M+1)\times(R-N+1)$ (Figure 5.4). We denote this location by (l^*, m^*, n^*). Due to varying illumination conditions and noise in the scene, we optimize the response of the filter by normalizing our correlation space:

$$c'(l,m,n) = \frac{c(l,m,n)}{\sqrt{E_H E_S(l,m,n)}}, \tag{3.6}$$

(a)



(b)

Figure 3.3: (a) Frames from a testing sequence containing the "wave2" action from the Weizmann action dataset. (b) The normalized correlation response for the testing sequence depicted in (a) correlated against the "Wave2" Action MACH filter depicted in Figure 3.2.

where $c(l, m, n)$ is given by equation 5. E_H is a scalar value which represents the energy of the filter, and $E_s(l, m, n)$ corresponds to the energy of the test volume at location (l, m, n), given by:

$$E_H = \sum_{t=0}^{N-1} \sum_{y=0}^{M-1} \sum_{x=0}^{L-1} H^2(x, y, t), \tag{3.7}$$

$$E_S(l, m, n) = \sum_{t=0}^{N-1} \sum_{y=0}^{M-1} \sum_{x=0}^{L-1} s^2(l + x, m + y, n + t). \tag{3.8}$$

Each element in the response of the normalized correlation lies within 0 and 1, a fact that can be used as a level of confidence in a pseudo-probabilistic manner. The peak value in the response of the filter is compared with a threshold (τ). If it is greater than the threshold, we infer that the

action is occurring at the corresponding location in the test video. Thresholds for action classes are computed during training as $\tau = \xi * min(p_1, p_2, p_3, ..., p_{N_e})$, where p_i is the peak value obtained from the correlation response when ith training volume was correlated with the 3D MACH filter, ξ is a constant parameter, and N_e is the number of all the training volumes.

3.1.2 Action MACH Filter for Vector Fields

In the previous section we described the process of synthesizing an Action MACH filter based on scalar data. In this section we extend our approach to include vector data.

3.1.2.1 Spatiotemporal Regularity Flow

The estimation of motion in video sequences constitutes an integral task in numerous applications, including human action recognition. The ability to estimate motion accurately and consistently has numerous challenges associated with it, such as motion discontinuities, aperture problems, and large illumination variations. Several of these challenges lead to direct violations of the assumptions embedded in the formulation of the classical flow estimation methods, such as Horn and Schunck's optical flow [37]. Therefore, numerous flow estimation approaches have been proposed which deal with large motion discontinuities [55], complex illumination variation, iconic changes [7], and three-dimensional scene flow [104].

In this work, we capture the temporal regularity flow information of an action class using a recently proposed "Spatio-temporal Regularity Flow" (SPREF)[4]. SPREF computes the directions,

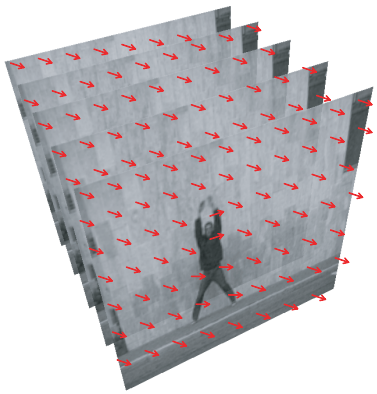


Figure 3.4: The directions of regularity obtained by xy-parallel SPREF for an instance of the "jumping jacks" action in the Weizmann dataset.

along which the sum of the gradients of a video sequence is minimized:

$$E = \int\int\int \left| \frac{\partial (F * \mathcal{G})(y, x, t)}{\partial \zeta(x, y, t)} \right|^2 dxdydt, \tag{3.9}$$

where F is the spatiotemporal volume, and $\mathcal{G}$ is a regularizing filter (Gaussian). This formulation results in a 3-D vector field in which each location is composed of three values that represent the directions along which intensity in a spatiotemporal region is regular, i.e., the pixel intensities in the region change the least.

SPREF is designed to have three cross-sectional parallel components in order to handle the regularities that depend on the motion and the scene structure. These components are: *xy-parallel*($\mathcal{F}_t$), *xt-parallel*($\mathcal{F}_y$), and *yt-parallel*($\mathcal{F}_x$). For our experiments we employ the *xy-parallel* component of SPREF. A slice from from 3D flow field generated from the *xy-parallel* component of SPREF is depicted in Figure 3.4.

This approach to regularity flow estimation does not rely on edge detection, hence its success does not depend on the presence of strong edges in the scene. Instead it analyzes the entire spatio-temporal volume, and tries to find the best directions that model the overall regularity of the volume. Even when the local gradient of a pixel is not significant, the global analysis of the region assigns a well-defined direction to it. The strength of SPREF lies in treating the data not as a sequence of 2D images, but as a 3D volume, and processing all of its information simultaneously.

In order to incorporate SPREF flow vectors directly into the synthesis of Action MACH filters, the MACH framework must be generalized to incorporate vector valued data. In the next subsection we employ an extension to Euclidean n-space based on a Clifford algebra which allows for a generalization of traditional Fourier transform functions to vector fields.

3.1.2.2 Clifford Embedding

Unlike Action MACH filters derived from scalar values as defined in Section 3.1.1, the process of synthesizing a filter based on a vector field cannot employ the traditional Fourier transform which is defined on scalar values. Both synthesis and correlation operations of MACH filters are performed in the frequency domain, therefore we require an analog to the classical Fourier transform for vector fields. For this purpose, we follow the framework proposed in [29], which consists of applying an extension of the traditional Fourier transform to include vector-valued data. This class of Fourier transform is commonly referred to as the "Clifford Fourier transform." Using this embedding we preserve the full information of both magnitudes as well as directions of our vector dataset while learning the action MACH filters.

Clifford algebra extends the Euclidean n-space to a real algebra. For a three-dimensional Euclidean vector space E^3 we obtain an eight-dimensional $\mathbb{R}$-algebra G^3 that has the following bases of a real vector space:

$$\{1, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1\mathbf{e}_2, \mathbf{e}_2\mathbf{e}_3, \mathbf{e}_3\mathbf{e}_1, \mathbf{e}_1\mathbf{e}_2\mathbf{e}_3\} \tag{3.10}$$

Elements belonging to this algebra are referred to as multivectors, and the structure of the algebra is given by: $1\mathbf{e}_j = \mathbf{e}_j; \mathbf{e}_j\mathbf{e}_j = 1; \mathbf{e}_j\mathbf{e}_k = -\mathbf{e}_k\mathbf{e}_j$, where $j = 1, 2, 3$. Based on this algebra, a set of of basic operators can be defined to generalize Euclidian space to encompass vector fields. These include not only the basic operations such as Clifford multiplication and integrals, but also composite operations such as Clifford Convolution and the Clifford Fourier Transform.

The Clifford Fourier transform (CFT) for multivectors-valued functions in 3D is defined as:

$$\mathcal{F}\{\mathbf{F}\}(\mathbf{u}) = \int \mathbf{F}(\mathbf{x}) \exp(-2\pi \mathbf{i}_3 \langle x, u \rangle) |d\mathbf{x}|, \tag{3.11}$$

where $\mathbf{i}_3$ represents the the analog of a complex number in clifford algebra, such that $\mathbf{i}_3 = \mathbf{e}_1\mathbf{e}_3$ and ${\mathbf{i}_3}^2 = -1$. The inverse transform is given by

$$\mathcal{F}^{-1}\{\mathbf{F}\}(\mathbf{x}) = \int \mathbf{F}(\mathbf{x}) \exp(-2\pi \mathbf{i}_3 \langle x, u \rangle) |d\mathbf{x}|. \tag{3.12}$$

A multivector field $\mathbf{F}$ in Clifford space corresponding to a thee-dimensional Euclidian vector field can be regarded as four complex signals which are independently transformed by a standard complex Fourier transformation. Therefore, the Clifford Fourier transform can be defined as a linear combination of several classic Fourier transforms:

$$
\begin{aligned}
\mathbf{F}(\mathbf{x}) &= [\mathbf{F}_0(\mathbf{x}) + \mathbf{F}_{123}(\mathbf{x})\mathbf{i}_3]1 + \\
&\quad [\mathbf{F}_1(\mathbf{x}) + \mathbf{F}_{23}(\mathbf{x})\mathbf{i}_3]\mathbf{e}_1 + \\
&\quad [\mathbf{F}_2(\mathbf{x}) + \mathbf{F}_{31}(\mathbf{x})\mathbf{i}_3]\mathbf{e}_2 + \\
&\quad [\mathbf{F}_3(\mathbf{x}) + \mathbf{F}_{12}(\mathbf{x})\mathbf{i}_3]\mathbf{e}_3, \qquad (3.13)
\end{aligned}
$$

which can be interpreted as belonging to $\mathbb{C}^4$. Given the linearity property of the Clifford Fourier Transform, the Fourier transform for multivector is given by:

$$
\begin{aligned}
\mathcal{F}\{\mathbf{F}\}(\mathbf{u}) &= [\mathcal{F}\{\mathbf{F}_0(\mathbf{x}) + \mathbf{F}_{123}(\mathbf{x})\mathbf{i}_3\}(\mathbf{u})]1 + \\
&\quad [\mathcal{F}\{\mathbf{F}_1(\mathbf{x}) + \mathbf{F}_{23}(\mathbf{x})\mathbf{i}_3\}(\mathbf{u})]\mathbf{e}_1 + \\
&\quad [\mathcal{F}\{\mathbf{F}_2(\mathbf{x}) + \mathbf{F}_{31}(\mathbf{x})\mathbf{i}_3\}(\mathbf{u})]\mathbf{e}_2 + \\
&\quad [\mathcal{F}\{\mathbf{F}_3(\mathbf{x}) + \mathbf{F}_{12}(\mathbf{x})\mathbf{i}_3\}(\mathbf{u})]\mathbf{e}_3. \qquad (3.14)
\end{aligned}
$$

Therefore, the Clifford Fourier transform of a SPREF vector field can be computed as a linear combination of four classical Fourier transforms. As a result, all of the well-known theorems that apply to the traditional Fourier transform hold for the CFT. In our experiments we map vector fields in 3D Euclidian space to $\mathbf{F}_{123}$ in the Clifford domain, and the remaining components are set to zero. For this purpose we use the publicly available GluCat library.[1] We exploit space decomposition properties described in this section to apply traditional Fast Fourier algorithms

[1] glucat.sourceforge.net

in order to accelerate the computation of the CFT, thereby reducing computation to a matter of seconds for a $320 \times 240 \times 300$ flow field.

Since scalars and vectors are part of multivectors, both scalar and vector-valued fields can be regarded as multivector fields. Therefore, the described Clifford embedding becomes a unifying framework for scalar, vector, and multivector-values filters.

3.1.2.3 Filter Synthesis

Given the SPREF flow field volumes in the frequency-domain we can proceed to convert the corresponding Clifford Fourier matrix into a column vector by concatenating all the columns of the matrix. This results in a single column-vector denoted by x_i, where $i = 0, 1, 2..., N_e$ and N_e represents the number of training examples of an action class. We then proceed to synthesize the Action MACH filter in the frequency domain using the same methodology described in section 3.1.1. Similarly, detection of new action instances for a given class is performed as described in section 3.1.1.1, replacing traditional scalar convolution with Clifford convolution. We evaluate the performance of Action MACH filters synthesized on vector fields and compare it with scalar-based filters in our experimental section.

3.1.3 Action MACH Using Spatio-Temporal Regularity Flow

In the next section we evaluate the performance of each of these approaches to synthesizing action MACH filters and compare our results with existing action recognition methods.

3.2 Experiments and Results

We performed an extensive set of experiments to evaluate the performance of the proposed method on a series of publicly available datasets and on a collection of actions found in feature films and broadcast television.[2] Details about the datasets and the experiments performed are given below.

3.2.1 KTH Dataset

The KTH human action dataset [95] contains 25 people performing six action classes, namely: walking, running, jogging, hand waving, boxing, and hand clapping. Each video sequence contains one actor repeatedly performing an action. The dataset contains a varied set of challenges including scale changes, variation in the speed of execution of an action, and indoor and outdoor illumination variations. Each sequence averages about 4 seconds in length.

Action classification is performed by cross correlation in the Clifford Fourier domain. We used the 5-fold cross-validation framework [56] to partition the dataset into K subsamples. We report the mean of the results obtained from MACH filters synthesized on spatio-temporal regularity flow vectors in Table 1. We achieve a mean accuracy of 88.66%, outperforming all other methods that rely on flow-based features alone.

A second set of experiments on the KTH dataset was geared towards evaluating the effect of using different features to train the action MACH filter. A 5-fold cross validation framework was

[2] http://www.cs.ucf.edu/~mikel/datasets.html

Action	Walk	Jog	Run	Box	Clap	Wave
Walk	0.91	0.04	0.04	0.01	0.00	0.00
Jog	0.05	0.84	0.11	0.00	0.00	0.00
Run	0.01	0.12	0.87	0.00	0.00	0.00
Box	0.01	0.00	0.04	0.95	0.00	0.00
Clap	0.00	0.00	0.01	0.09	0.85	0.05
Wave	0.00	0.00	0.00	0.04	0.06	0.9

Table 3.1: Confusion matrix using our method synthesized with SPREF vectors for the KTH actions database. Mean accuracy=86.66%

employed to obtain mean accuracy averages for MACH filters trained using block-based optical flow vectors, temporal derivatives (scalar data), and spatio-temporal regularity flow vectors.

Action MACH filters were synthesized for each action based on optical flow vectors by employing the 2-dimensional formulation of the Clifford transform [4]. The mean accuracy of the optical-flow based filter was 87.2%. Finally, an Action MACH filter trained on scalar data obtained from temporal derivatives yielded a mean accuracy of 80.9%.

3.2.2 Feature Films

We have compiled a dataset of actions performed in a range of film genres consisting of classic old movies such as “A Philadelphia Story,” “The Three Stooges,” and “Gone With the Wind,” comedies such as “Meet the Parents,” a sci-fi movie titled “Star Wars,” a fantasy movie “The Lord of the Rings: The Return of the King,” and romantic films such as “Connie and Carla.” This dataset provided a representative pool of natural samples of action classes such as “Kissing” and “Hit-

	Mean accuracy
Temporal derivatives	80.9%
Optical flow	87.2%
SPREF	88.66%

Table 3.2: Comparison of various feature sets used for the MACH filter on the KTH dataset.

Figure 3.5: Detections of the kissing actions (a) and the slapping actions (b) in classic feature films.

ting/slapping." We extracted 92 samples of the "Kissing" and 112 samples of "Hitting/Slapping." The extracted samples appeared in a wide range of scenes and view points, and were performed by different actors. Instances of action classes were annotated by manually selecting the set of frames corresponding to the start and end of the action along with the spatial extent of the action instance.

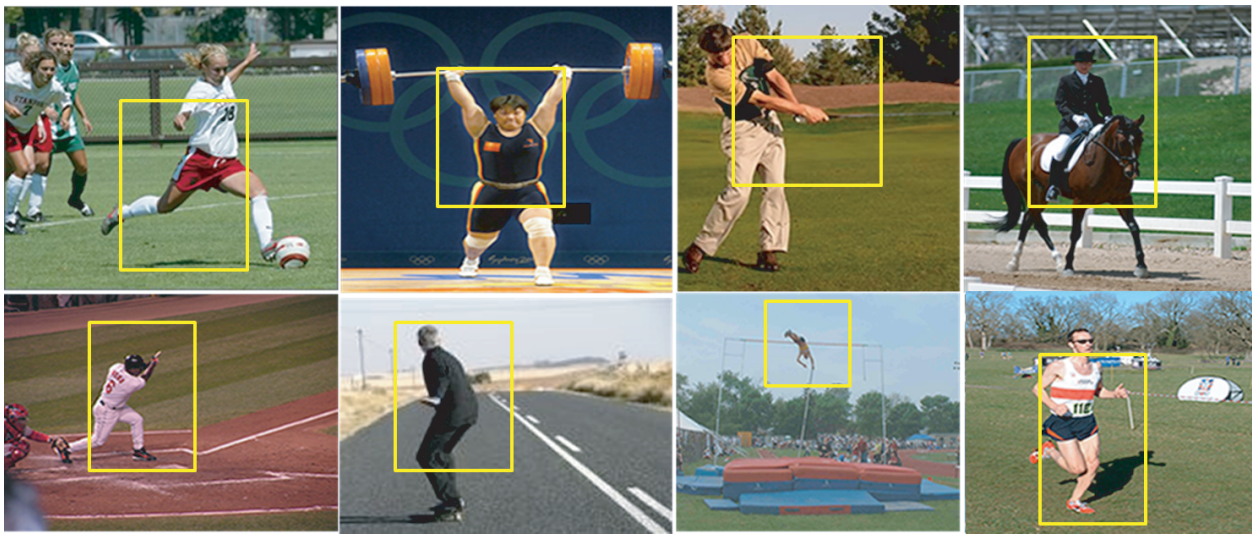


Figure 3.6: The collection of broadcast sports actions videos represents a set of typical network news videos featured on the BBC and ESPN.

Testing for this dataset proceeded in a leave-one-out framework. Given the significant intra-class variability present in the movie scenes, the recognition task is challenging. In our experiments using SPREF, we achieved a mean accuracy of 66.4% for the "Kissing" action, and a mean accuracy of 67.2% for the "Hitting/Slapping"action.

3.2.3 Broadcast Television Action Dataset

We have collected a set of actions from various sports featured on broadcast television channels such as the BBC and ESPN. Actions in this dataset include diving, golf swinging, kicking, lifting, horseback riding, running, skating, swinging a baseball bat, and pole vaulting (Figure 3.6). The dataset contains over 200 video sequences at a resolution of 720×480. The collection represents a natural pool of actions featured in a wide range of scenes and view points. To our knowledge this

dataset is the first of its kind; by releasing it we hope to encourage further research into this class of action recognition in unconstrained environments.

Testing for this dataset was performed using the leave-one-out cross-validation framework. The confusion matrix for this set of experiments is depicted in Figure 3.9. The overall mean accuracy for this dataset was 69.2%. Given the difficulty of the dataset, these results are rather encouraging.

3.2.4 Weizmann Action Dataset

We tested the proposed method on the Weizmann action dataset [10]. Data from this collection was partitioned into testing and training using 5-fold cross validation, the results are depicted in Figure 3.7.

The average run-time for a $144 \times 180 \times 200$ testing video from this dataset was 18.65 seconds on a Pentium 4, 3.0 GHz. Whereas [10] reports a runtime of 30 minutes on the same architecture for this dataset, our results represent a considerable increase in performance over existing template-based methods.

3.2.5 Cohn-Kanade Facial Expression Database

Although our main goal is to detect and locate human actions, our framework is well suited for other application domains which involve spatio-temporal matching. We adapted our algorithm to perform classification of different facial action units (AU). To test the method we used data from the commonly used subset of the Cohn-Kanade facial expression database [45]. This database consists of gray scale recordings of subjects displaying basic expressions of emotion on command.

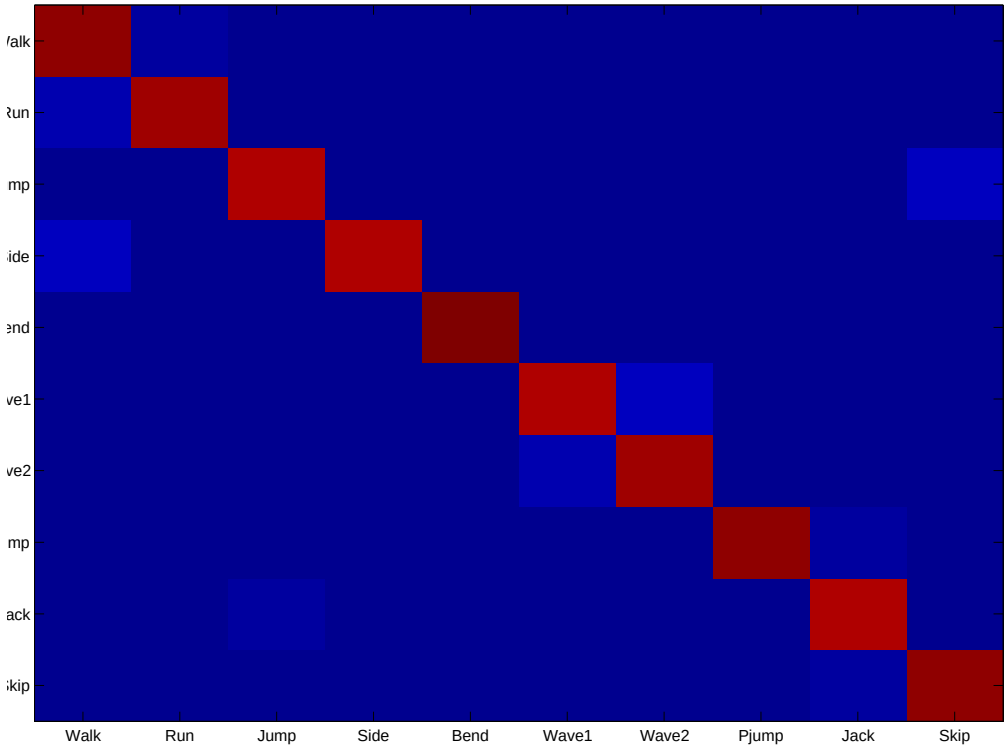


Figure 3.7: The confusion matrix depicting the results of action recognition for the Weizmann dataset.

The data included a set of upper face action units: AU1 (inner portion of the brows is raised), AU2 (outer portion of the brow is raised), AU4 (brows lowered and drawn together), AU5 (upper eyelids are raised), AU6 (cheeks are raised), and AU7 (lower eyelids are raised). The action units were partitioned into training and testing using 4-fold cross validation.

Unlike a significant number of existing works [46, 25, 5], no prior facial model or feature tracking was used in training. Additionally, we do not require manual marking of feature points

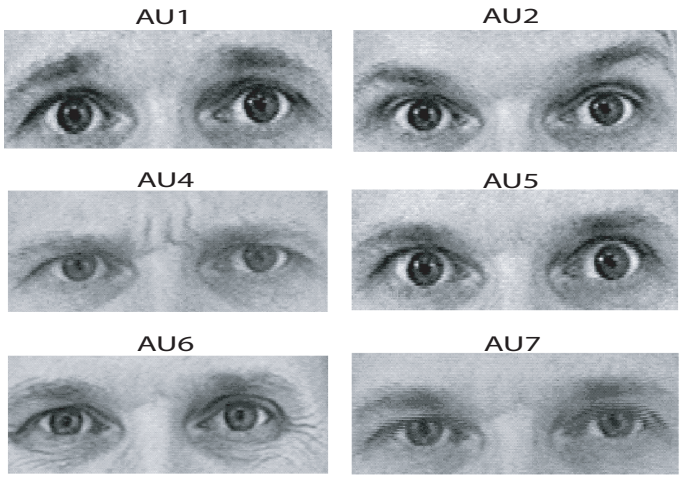


Figure 3.8: Example frames from the seven facial action units .

Action Unit	AU1	AU2	AU4	AU5	AU6	AU7
AU1	0.78	0.2	0.00	0.00	0.00	0.02
AU2	0.16	0.73	0.11	0.00	0.00	0.00
Au4	0.00	0.00	0.88	0.10	0.02	0.00
AU5	0.00	0.00	0.00	0.92	0.05	0.03
AU6	0.00	0.00	0.00	0.00	0.79	0.21
AU7	0.00	0.00	0.00	0.02	0.022	0.76

Table 3.3: Confusion matrix for 7 upper face AU. Acuracy=81.0%

around face landmarks or alignment with a standard face image, yet the performance on the standard dataset (as observed in Table 3) was comparable to current state-of-the-art systems. Despite the fact that our main focus lies in recognizing human body motion patterns, these results indicate that our approach provides enough discriminating power, even when subtle motions of the face are involved.

3.3 Conclusion

In this chapter we have introduced the Action MACH filter, a method for recognizing human actions which addresses a number of drawbacks of existing template-based action recognition approaches. Specifically, we address the ability to effectively generate a single action template which captures the general intra-class variability of an action using a collection of examples. Additionally, we have generalized the traditional MACH filter to operate on spatiotemporal volumes as well as vector valued data.

The results from our extensive set of experiments indicate that the proposed method is effective in discriminating a wide range of actions. These include both whole-body motions (such as jumping jacks or waiving) and subtle localized motions (such as smiling or raising eyebrows). Additionally, by analyzing the response of the Action MACH filter in the frequency domain, we avoid the high computational cost which is commonly incurred in template-based approaches.

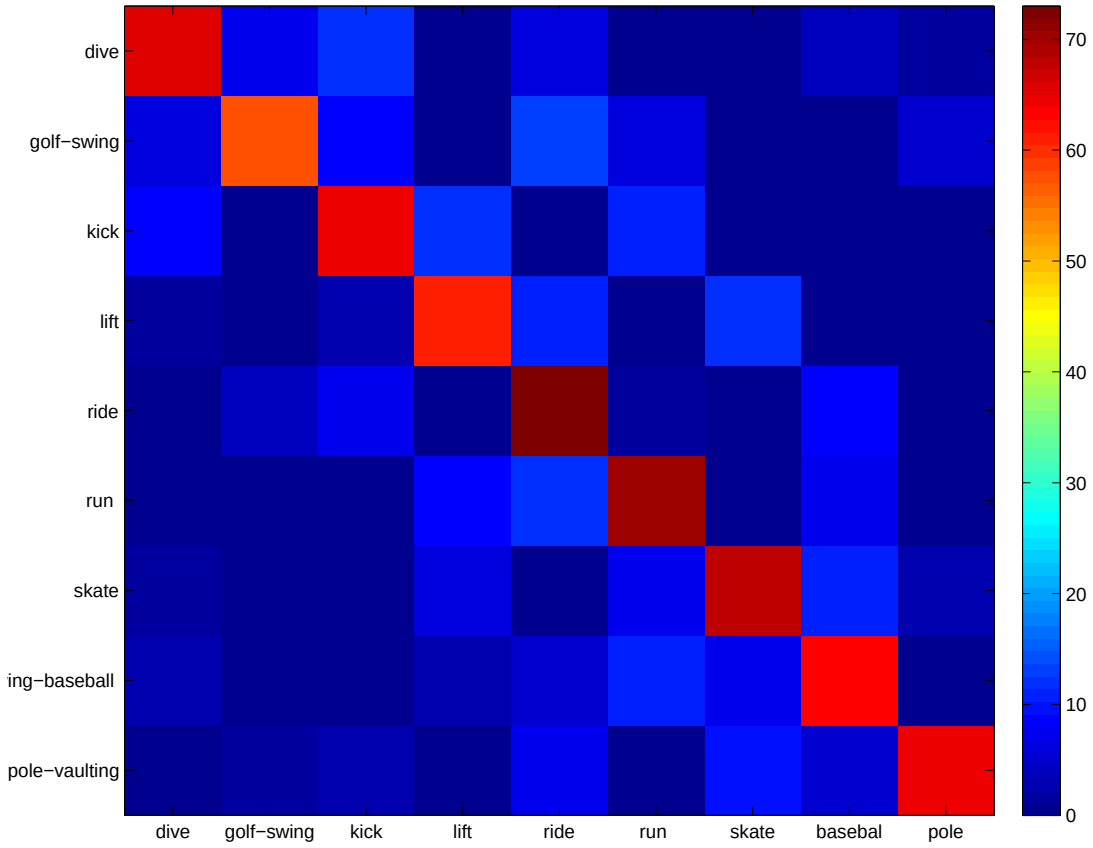


Figure 3.9: The confusion matrix depicting the results of action recognition for diving, golf swinging, kicking, lifting, horseback riding, running, skating, swinging a baseball bat, and pole vaulting.

CHAPTER 4: AERIAL ACTION RECOGNITION USING SPATIO-TEMPORAL QUADRATIC CORRELATION FILTERS

In this chapter, we discuss the second problem that we address in this dissertation which focusses on recognizing actions in aerial videos by including both positive *and* negative examples of an action class in order to obtain a quadratic correlation filter that generalizes the variability associated with an action. Specifically, we address three seldom explored challenges in template-based action recognition. The first is the recognition and localization of human actions in aerial videos obtained from unmanned aerial vehicles (UAVs), a new medium which presents unique challenges due to the small number of pixels per human, pose, and moving camera. The second issue we address is the incorporation of multiple positive and negative examples of a target action class when generating an action template. We address this issue by employing the Fukunaga-Koontz Transform as a means of generating a single quadratic template which, unlike traditional temporal templates (which rely on positive examples alone), effectively captures the variability associated with an action class by including both positive and negative examples in the template training process. Finally, we explore a range of low-level and mid-level motion features and assess their effectiveness within the context of aerial action recognition by introducing a first-of-its-kind dataset which features actions captured from a UAV at different flying altitudes.

Aerial video collected by electro-optical and infrared cameras deployed on small UAV platforms is rapidly becoming a low-cost and up-to-date source of imagery. This increase in aerial

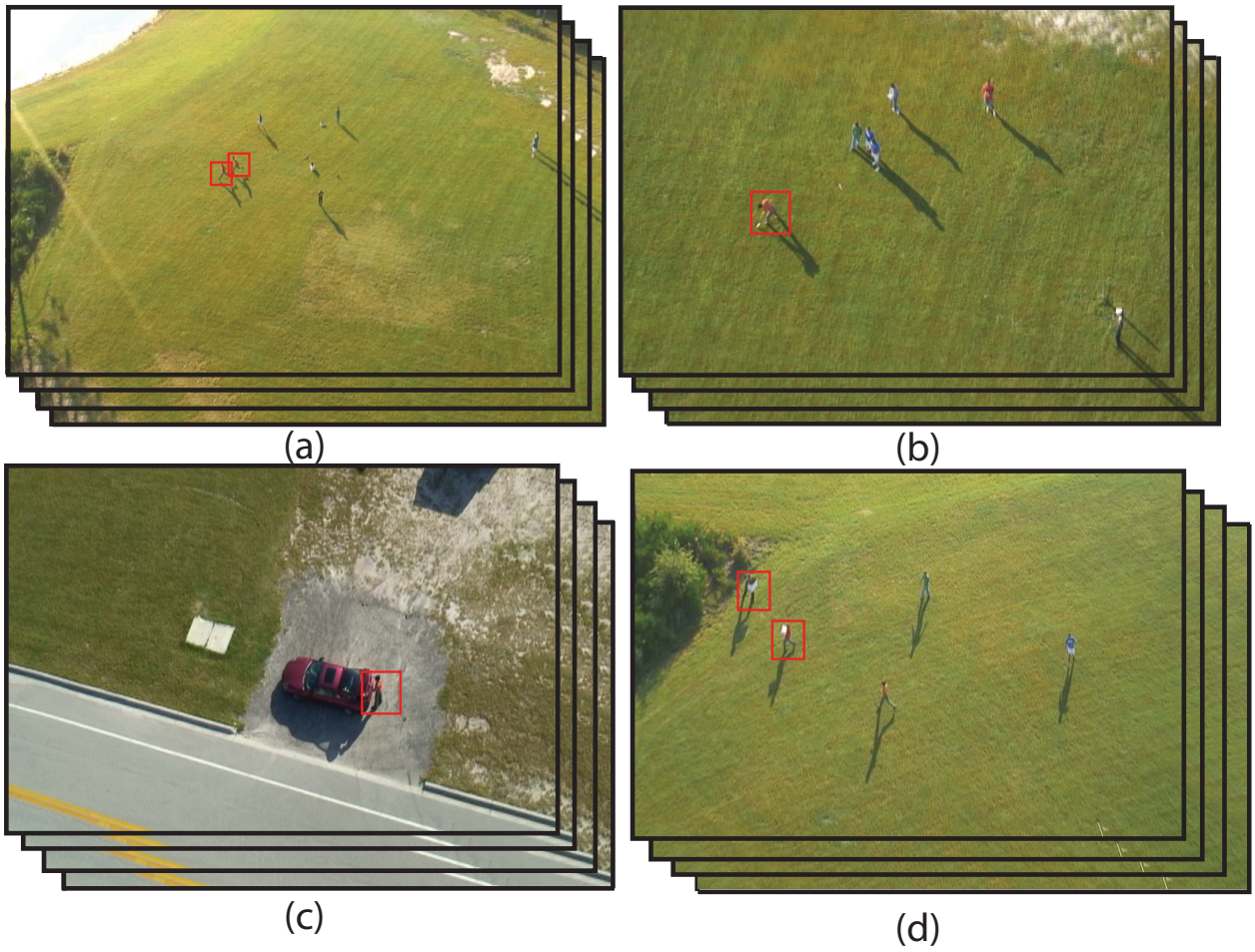


Figure 4.1: Detecting actions from video obtained from a UAV. (a) Two instances of the "running" action, (b) an instance of the "digging" action, (c) an "open trunk" action, (d) two instances of the "carrying" action.

videos has led to the development of systems for natural disaster remediation and surveillance monitoring. Given the large volume of video which is currently being generated by such platforms, it has become infeasible for users to sort through thousands of hours of video for events or activities of interest. Therefore, vision-based action recognition methods capable of detecting specific actions of interest in aerial videos are needed.

Chapter 2 described the large body of work on action detection which mainly focuses on ground-level cameras. However, there has been almost no attention has been placed on understanding how the challenges associated with traditional action recognition relate to actions in videos obtained from aerial platforms. Similarly, no work as been done on including positive and negative samples of an action class to generate action templates. Given a collection of positive examples of an action sequence and clutter samples, a disadvantage of the method described in Chapter 3 is its inability to generalize from a collection of positive and negative examples and create a template which captures the variability of an action.

As part of the second component of this dissertation we focus on addressing three seldom examined challenges in template-based action recognition. The first issue we intend to address is the capability of explicitly incorporating multiple positive *and* negative examples of a target action class in order to generate a template which effectively captures the variability associated with the action. For this purpose we intend to study the Fukunaga-Koontz Transform which provides us with a systematic framework for including data from both a positive and a negative training examples to establish a set of optimal basis vectors that maximize the separation of the classes.

The second issue we intend to explore will focus on recognizing human actions from video captured at high altitudes. To obtain a full grasp of the problem at hand, it is useful to contrast "actions from above" with traditional ground-camera mediums. In traditional near-field, ground-camera action recognition, humans in the scene are hundreds of pixels both tall and wide. In this class of mediums articulated motions can be clearly distinguished. Most of the existing work on human action classification works best with data of this resolution and viewpoint. Medium-field,

Figure 4.2: A collection of actions captured from an aerial mobile platform at different angles and altitudes. On the top row: "walking," two videos of "open trunk," and "run". On the bottom row: "pick up," "run," "dig," and "open car door".

ground-camera setups typically only have humans which are close to a hundred pixels tall. However, it has been shown that pose in this class of ground camera mediums is such that articulated motions can still be effectively classified.

In this chapter we develop a general approach for recognizing actions from above (Figure 4.1), a medium in which articulated actions are difficult to discern for two main reasons: the low number of pixels per human in the scene, and the overhead viewpoint. At altitudes over 400 feet, humans captured on video are, on average, less than 8 pixels wide and 15 pixels tall, and the overhead camera perspective leads to frequent self-occlusion of the limbs.

Finally, the third issue we explore in this chapter pertains to the unique challenges that need to be addressed when selecting which motion features are to be used to generate an action template. We study the effectiveness of low-level features such temporal derivatives and optical flow, as well

as a number of mid-level motion features which diminish the effect of generating action templates on noisy low-level features.

4.1 Incorporating Negative Training Examples

In our construction proposed in Chapter 3, an action template was generated from a collection of positive samples of an action class. In this chapter we propose a new quadratic spatio-temporal action template framework which, unlike traditional template-based approaches, generalizes from multiple examples of positive instances of an action class as well as a set of negative examples (clutter) present in a dataset. For each action class, we begin by computing a set of features for both positive and negative examples of the class. Subsequently, we determine a set of dominant bases in the feature domain which contain maximal information about a given action class, while at the same time containing minimal information about the negative class (described in section 4.1.1). Finally, a spatio-temporal action template is generated based on the most discriminating basis. In the following subsections, we describe each of the aspects of the spatio-temporal quadratic action template generation and correlation process in more detail.

4.1.1 Fukunaga-Koontz Transform

The structure of the quadratic spatio-temporal action template can be viewed as projecting an input feature vector field onto a set of templates and accumulating the correlation score. In this work, we use the Fukunaga-Koontz Transform (FKT) as a systematic methodology of finding a basis set which maximizes the separation between a target action class and the rest of the dataset. In this

subsection, we briefly review the basic idea and concepts which pertain to our application domain of the FKT.

Similar to the popular Principle Component Analysis (PCA), the FKT is also intended to find an orthogonal basis space which is most efficient for representing the target data. However, unlike traditional PCA, the FKT uses data from both a positive *and* a negative training set to establish a set of optimal basis vectors that represent positive class and simultaneously have no representation of negative class.

Assume we have a collection of positive training examples of a target action class and a set of negative training examples. Let $\mathbf{X}_{d\times m}$ be the matrix of which each column represents a training example of the target action class. Let $\mathbf{Y}_{d\times n}$ be the matrix of which each column represents an example from the negative example class. Then the class-mean-removed covariance matrix $\mathbf{\Sigma}$ of all data is given by the summation of the covarience matrix $\mathbf{\Sigma_X}$ and $\mathbf{\Sigma_Y}$. We can perform eigen analysis on $\mathbf{\Sigma}$ and decompose it into the following form:

$$\mathbf{\Sigma} = \mathbf{\Sigma_X} + \mathbf{\Sigma_Y} = \mathbf{\Phi D \Phi}^T, \tag{4.1}$$

where $\mathbf{\Phi}$ is the matrix containing eigenvectors of $\mathbf{\Sigma}$; $\mathbf{D}$ is matrix in which each diagonal element correspond to the corresponding eigenvalues of $\mathbf{\Phi}$. If we transform all of our data (both the positive and negative action classes) by $P = \mathbf{\Phi D}^{\frac{-1}{2}}$ then we will get new data set $\tilde{\mathbf{X}} = \mathbf{P}^T\mathbf{X}$ and $\tilde{\mathbf{Y}} = \mathbf{P}^T\mathbf{Y}$.

The new class-mean-removed covariance matrix of all action training is now represented by:

$$\begin{aligned} \tilde{\Sigma} &= \Sigma_{\tilde{\mathbf{X}}} + \Sigma_{\tilde{\mathbf{Y}}} = \mathbf{P}^T \Sigma_{\mathbf{X}} \mathbf{P} + \mathbf{P}^T \Sigma_{\mathbf{Y}} \mathbf{P} \\ &= \mathbf{P}^T (\Sigma_{\mathbf{X}} + \Sigma_{\mathbf{Y}}) \mathbf{P} = \mathbf{P}^T \Sigma \mathbf{P} = \mathbf{I}. \end{aligned} \tag{4.2}$$

4.1.2 Quadratic Spatio-temporal Action Template

The process of generating the proposed quadratic spatio-temporal template is as follows: for every action class we collect a set of training examples $(x_i, i = 1, ..., T)$ in which x_i is a spatio-temporal volume (with dimensions $m \times n \times t$) which represents an instance of an action. Similarly, we collect a set of negative examples $(y_i, i = 1, ..., C)$, in which y_i represents a spatio-temporal volume (with dimensions $m \times n \times t$) that holds an instance of a negative example of the action class. The spatio-temporal volume can be converted into an $mnt \times 1$ vector, and we can transform the data with the Fukunaga-Koontz transform (described in the previous section). In our proposed formulation, a small subset of the dominant target action class and negative basis functions will be chosen. Each basis function corresponds to an individual spatio-temporal filter which we combine into a *single* quadratic spatio-temporal action template which is able to generalize from multiple examples of an aerial action class as well as incorporate negative examples of the class (Figure 4.3). Specifically, we choose the N_1 top eigenvectors that best describe the positive action class, and the N_2 eigenvectors that contain the most information about he negative action examples in the training data. Θ is defined as: $\Theta = [\vec{\theta}_1, ...\vec{\theta}_{N_1}, \vec{\theta}_{mnt-N_2+1}, ..., \vec{\theta}_{mnt}]$. Detection of an action is done by projecting a testing video onto this set of filters in order to obtain a correlation space v.

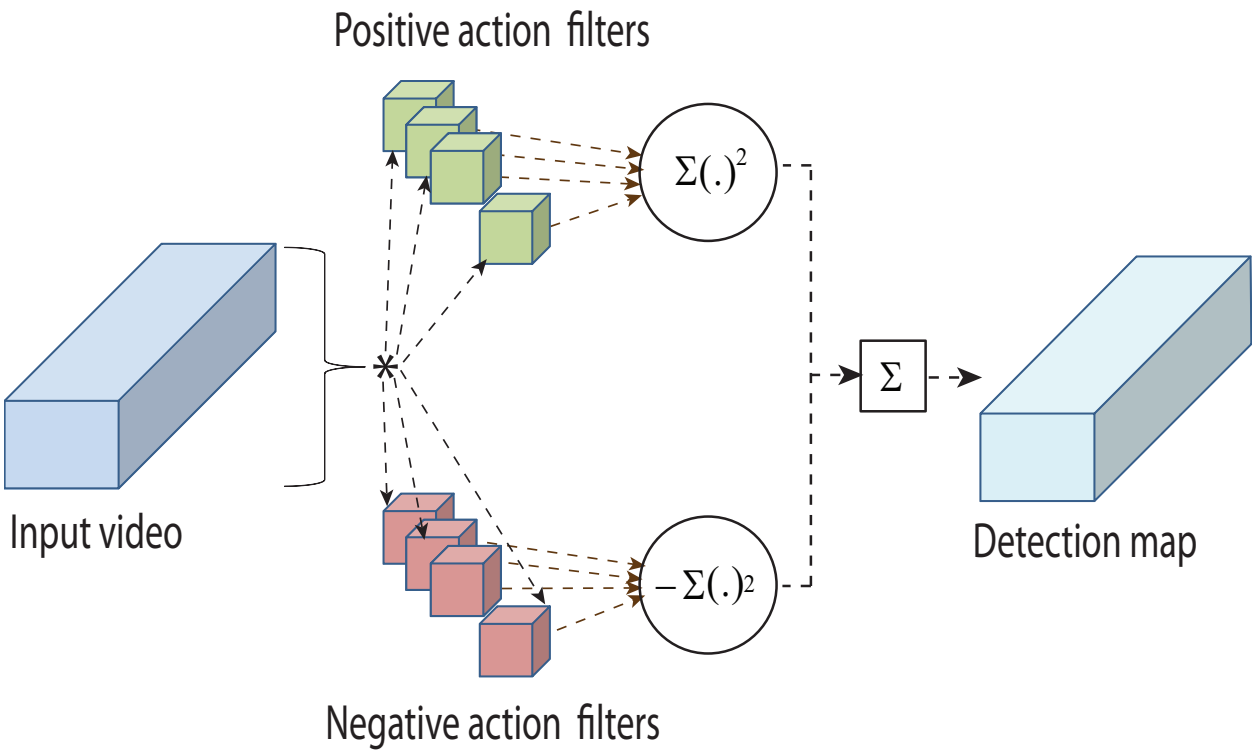


Figure 4.3: Structure of the proposed quadratic action template.

The correlation space for a given action class is therefore given by:

$$\phi = \sum_{i=N_1+1}^{N_1+N_2} v_i^2 - \sum_{i=1}^{N_1} v_i^2, \tag{4.3}$$

see Figure 4.3. The summation on the rhs is the correlation score of the test video on the positive-action-class basis function. The second sum represents the correlation score of the negative-action-class basis functions. The correlation space for the quadratic spatio-temporal template is given by the substraction of the two correlation spaces. Values in this correlation space are expected to be large in the presence of the action class and small across negative samples of the action.

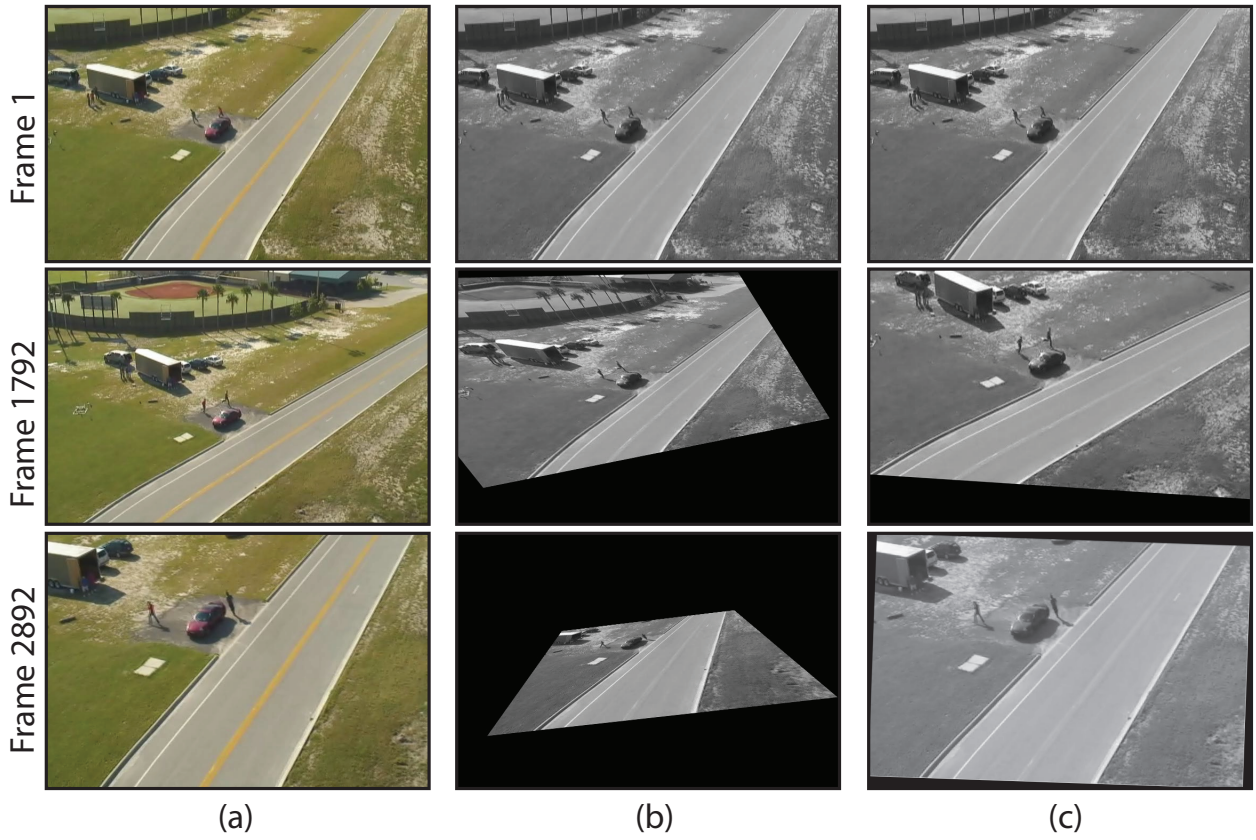


Figure 4.4: Planar homography only holds for the ground plane, out-of plane objects such as people may be distorted. (a) Input frames, (b) Ego-motion compensation without transformation reset, (c) Ego-motion compensantion using temporally local frames.

4.1.3 Motion Features

Recognizing actions from high altitudes presents several unique challenges that need to be addressed when selecting which motion features are to be used to generate an action template. In particular, the use of high-level features such as contours, local shape, and body joint tracks may not be appropriate given the relatively low number of pixels per human in aerial videos. Con-

versely, the use of low-level motion features such as raw optical flow may be of limited use given the noisiness of the data and the considerable amount of shake in videos obtained from UAVs. Therefore, we focus on a number of mid-level motion features which can diminish the effect of generating action templates on noisy low-level features in aerial videos.

4.1.3.1 Motion Magnitude and Direction

In an effort to minimize the effect of noisy aerial videos caused by UAV jitter and motion estimation errors, we treat motion channels (magnitude and direction) as spatio-temporal patterns of noisy measurements which are then used to construct a template which is robust to noise and clutter while capturing the general variability of an action class. In our experiments we have computed optical flow for each input sequence using the the approach described in [76], resulting in a 3D flow field. Subsequently, each optical flow vector (**F**) in the three-dimensional vector field is first divided into a set of scalar fields, u and v, each of which is then blurred with a Gaussian kernel and normalized. Once the two scalar fields are blurred, they are used to compute the two channels, magnitude ($F_M = \sqrt{u^2 + v^2}$) and direction ($F_D = \arctan v/u$), which serve as the basis of the action template.

4.1.3.2 Spatio-temporal Haar-like Features

A second class of mid-level features which builds on the motion magnitude and direction volumes described in the previous section is shown in Figure 4.8. In a similar vein to the popular Viola-Jones rectangle features for object classification [105] and their extension [48], we compute multiple

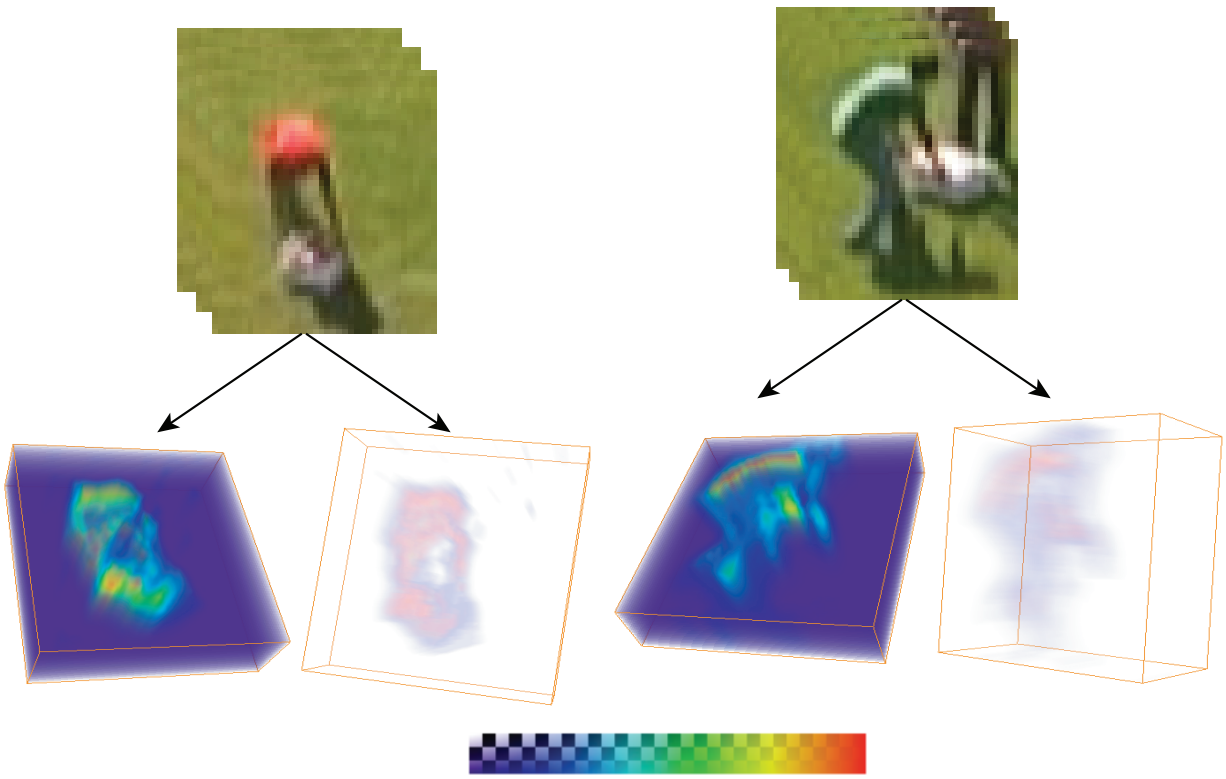


Figure 4.5: Magnitude and direction of optical flow are computed and smoothed independently for each training example. Together, the resulting scalar volumes are treated as a 2D feature vector (in the Clifford domain) when generating a spatio-temporal template.

single-bin and two-bin features within the scalar spatio-temporal volumes of F_M and F_D. The value of the one-bin feature is simply the overall sum of the scalar values within a given motion channel volume (F_M and F_D). Correspondingly, the value of a two-bin feature is the difference of their individual sums. In the experiments we have carried out, given a set of training examples of an action as captured from a UAV, magnitude and direction of optical flow vectors are computed as described in Section 4.1.3.1 and subsequently four types of spatio-temporal haar-like features

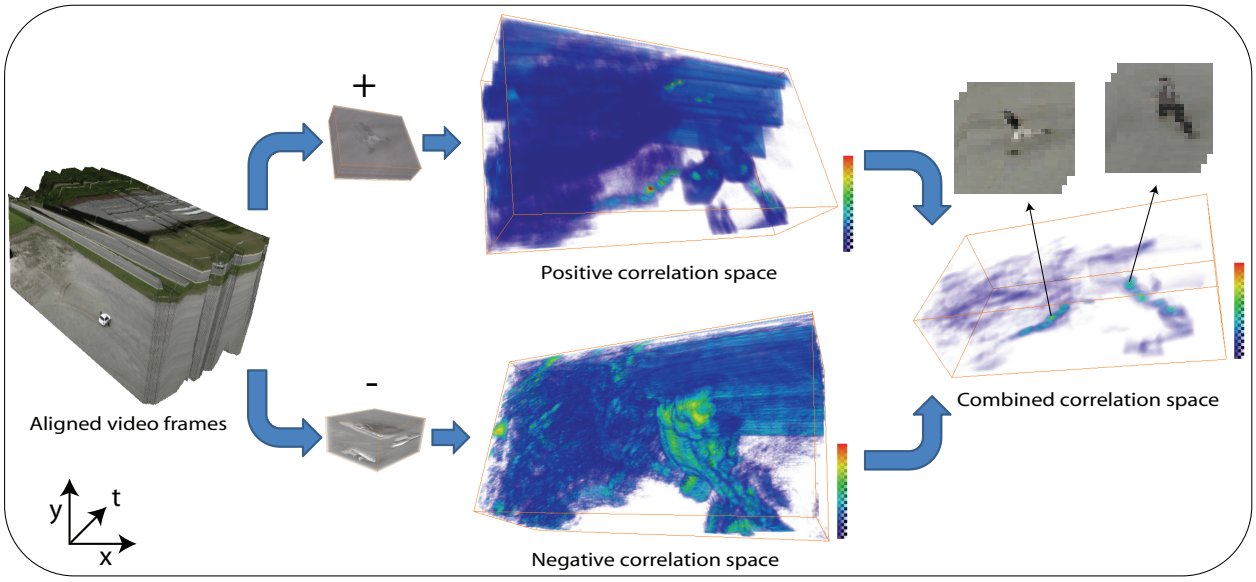


Figure 4.6: Detection of the "running" action. A spatio-temporal quadratic correlation filter represents a combination of several positive (top) and negative (bottom) filters which are combined into a single correlation space.

are computed over the spatio-temporal volume in $6 \times 6 \times 6$ blocks. Therefore, as can be seen in Figure 4.7, for each training sequence we obtain eight independent scalar volumes. The individual feature volumes of all the training examples are combined into a single template via the Clifford algebras embedding described in Section 4.1.3.1.

4.1.4 Ego Motion Compensation

Significant research effort has been expended towards ego-motion compensation (the spatial alignment of successive frames of a video) and it is now largely acknowledged to be a solved problem.

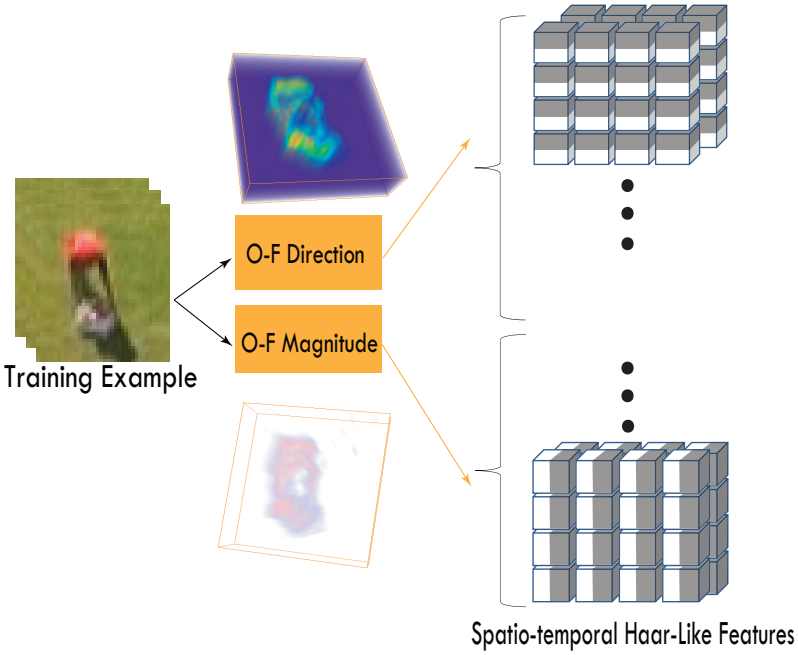


Figure 4.7: Each motion channel volume is subdivided into small spatio-temporal cubes, on which we compute spatio-temporal haar-like features. Individual scalar volumes are combined into a single multi-dimensional feature vector in the Clifford domain.

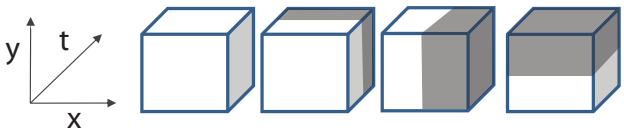


Figure 4.8: Spatio-temporal haar-like operators.

Nevertheless, the relationship between frame-to-frame registration techniques and aerial action recognition remains largely unexplored. We are interested in exploring the effects of different

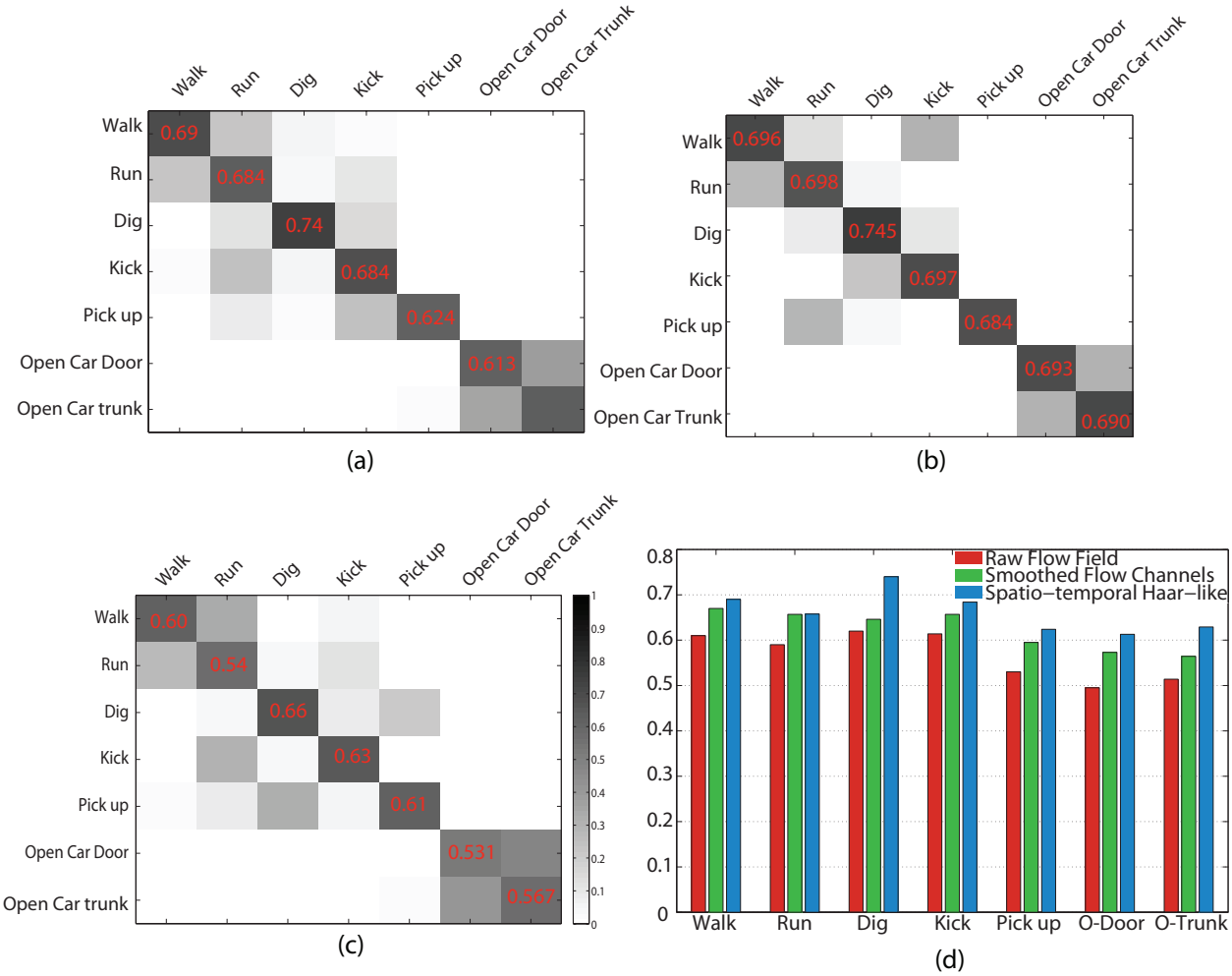


Figure 4.9: Unaligned frames (a), motion compensated video (b), positive training examples alone (c), various motion features (d).

frame-to-frame registration techniques on action recognition, as well as quantifying to what extent ego-motion compensation improves classification rates in specific action classes.

Due to the nature of the flight path of most mobile platforms, throughout the course of a video different action instances will be captured at varying angles, ranging from a Nadir viewpoint in

which the camera axis points directly downwards to a low oblique angle in which the horizon is visible in the frame.

Given this class of flight patterns and the fact that a planar homography only holds for the ground plane, out-of plane objects such as people and buildings can be distorted when performing ego-motion compensation. The greater the difference in camera views between frames the greater the distortion of these out-of plane objects. In our initial experiments we noted how this class of distortions can negatively affect classification accuracy significantly within different action classes. As an alliterative approach to handling frame-to-frame alignment we mitigate the effects of out-of plane distortion by warping only the frames that are temporarily close.

More specifically, given a reference frame f_r of a video, a current frame f_t, and a cumulative homography H that takes f_t to f_r, we define two measures of distortion: curl $C_{r,t}$ and deformation $D_{r,t}$

$$C_{r,t} = |(-H(0,1) + H(1,0))| \tag{4.4}$$

$$D_{r,t} = |(H(0,0) - H(1,1))| \tag{4.5}$$

where $H(0,0), H(0,1), H(1,0), H(1,1)$ constitute the upper-left quadrant of the cumulative homography matrix, which captures rotation and shear.

Given these measures of distortion we define distortion tolerance thresholds such that whenever their values are exceed we update the reference frame such that $f_r = f_t$. As can be seen in Figure 4.4-b, out-of plane distortion can be severe enough to render some action instances irreconcilable. In the example depicted in Figure 4.4, two instances of the running action are seem in frame

1792. When employing frame one as the reference frame this leads to significant out-of-plane distortions (Figure 4.4-b), whereas when using a temporally-local frame (frame 1763) we observe considerably less distortion (Figure 4.4-c). A progressive deterioration can be observed in frame 2892, in which the running action instances are practically irreconcilable when using long-range dependencies for reference frames.

4.2 Actions From Above Dataset

In order to assess the difficulty of the problem, we have begun to construct a new aerial action dataset which includes a representative set of typical actions captured by aerial mobile platforms. The AfA dataset contains fifteen videos that were obtained using a small unmanned aerial vehicle equipped with an HD camera mounted on a gimbal. The dataset contains a diverse pool of actions captured at different heights and aerial viewpoints.

In this dataset, multiple actions occur in the scene at any time, and there is no temporal segmentation of the video into short testing and training clips.

Actions in this dataset include "digging," "running," "walking," "kicking," "picking up an object," "opening a car trunk," and "opening a car door." A number of repetitions of each action were recorded at different flying altitudes which ranged from 100-300 feet and were performed by different actors. Videos in this dataset range from 1-12 minutes in length, unlike traditional single action instance datasets where each clip is only a few seconds long. Finally, each video in the dataset contains multiple instances of different action classes which are occurring concurrently.

Each human in the scene is manually annotated with its corresponding bounding box and action label, providing us with the spatio-temporal extent of each action as it occurs in the video. This information is later used to evaluate our action detection framework.

The main challenges associated with recognizing actions in this class of videos reside in the fact that articulated human motions are very difficult to observe at high altitudes. Other issues, such as moving camera and low number of pixels on target (which typically averages less than 8 pixels wide and 15 pixels tall) also factor in making this a complex problem. In the following sections, we explore several approaches which deal with the challenges mentioned above.

4.3 Experiments

We performed a wide range of experiments to evaluate the various aspects of recognizing actions at high altitudes using spatio-temporal templates. Specifically, experiments were geared towards assessing the relative performance of different motion features, the effect of using motion compensated arial video versus raw video frames, and towards quantifying the improvement obtained by different quantizations of scale space.

Our dataset consists of 2.25 hours of aerial video. Each action is repeated several times by different actors throughout the video. In our experiments we manually select and crop a set of representative positive and negative training examples (20 examples on average) for each action class. Features are computed for each of the training samples as described in Section 4.1.3, and quadratic spatio-temporal action templates are generated in Section 4.1.2. Testing is done on the

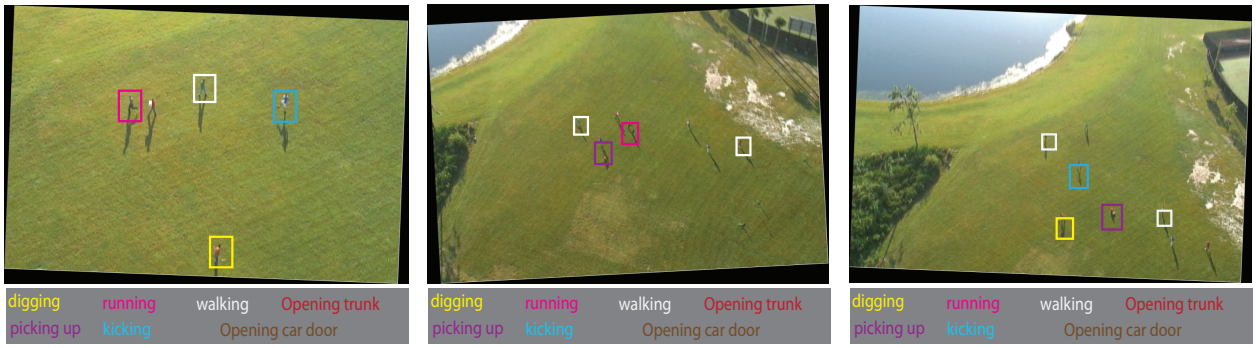


Figure 4.10: Multi-class detection of actions from above at different scales, from left to right: near, medium and far.

unsegmented video clips which are typically five minutes long. Within our experiments detections are considered true positives only if there is an overlap between the detection bounding box and the annotated spatio-temporal extent of an action in the video.

4.3.1 Motion Features

So as to better understand the effect of different spatio-temporal motion features on classification within the context of actions from above, we generate and assess classification rates of action templates generated from different features.

The classification results for the first set of experiments on the AfA dataset are illustrated in Figure 4.9-d. In these experiments we perform ego-motion compensation and generate action templates based on different motion features. We use three templates per class, 20 training examples

per template, and this training set is the same for each feature class. Spatio-temporal non-maximum suppression ensures that only locally maximum correlation values are reported as action detections.

As can be seen in Figure 4.9-d, the best results are obtained using the spatio-temporal haar-like features, followed by magnitude and direction channels. The use of raw optical flow results in relatively poor performance, which can be attributed to the high amount of noise in the estimated motion field of the aerial videos.

4.3.2 Camera Motion Compensation

In a second set of tests we evaluate the effect of ego motion compensation on recognition rates. Two template construction paradigms are tested, the first consisting of temporally short templates, which are typically less than 20 frames long. The intuition behind this approach is to minimize the effect of moving cameras in UAV videos by limiting the temporal extent to a short window. The second template generation approach relies on motion compensation as described in Section 4.1.4. Template construction is then performed on haar-like motion features computed on the registered frames. Similarly, detection is done only after registering the input sequence. In both paradigms we use three templates per class, 20 standardized training examples per template.

As can be seen in Figure 4.9-a and 4.9-b, templates generated on short temporal windows do not underperform by large margins, indicating that it may be possible to avoid camera motion compensation for some actions. Looking closely, we notice that motion compensation benefits actions which inherently have longer temporal dependencies such as “opening trunk,’ and ’ “getting in vehicle (which on average improve by 7.05%),” whereas actions which typically last only a

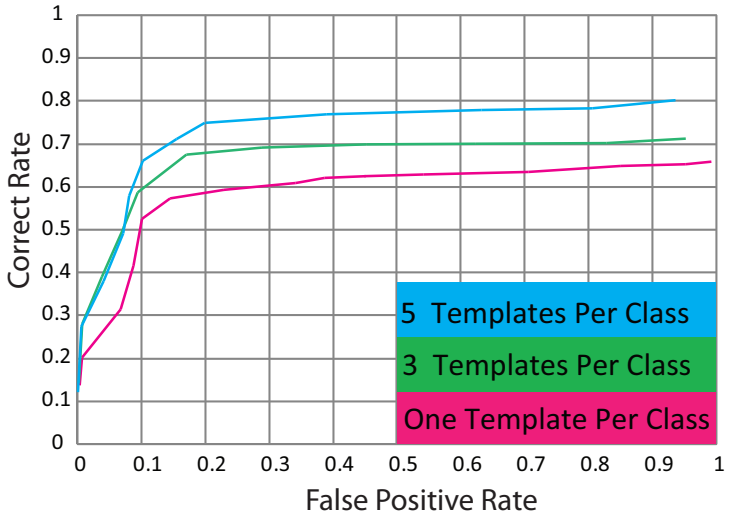


Figure 4.11: Scale space quantizations.

few frames such as “kicking” or “running” are affected the least by the lack of camera motion compensation, improving by only 2.06% on average.

4.3.3 Scale and Viewpoint

Typically, civilian-use unmanned aerial vehicles fly at a wide range of altitudes (anywhere from 200 to 1000 feet). Therefore, actions are usually captured at different scales. Furthermore, depending on the flight pattern of the mobile platform a given action instance can be captured at a variety of viewpoints. In our third round of experiments we evaluate the performance of different approaches to dealing with scale and viewpoint changes associated with different flying altitudes and poses. In particular, we quantify how detection rates improve as we generate filters for a set of discrete scales and viewpoints.

The first configuration of this experiment corresponds to the use of a single template per action class generated based on examples of all scales. The second configuration consists of the division (done manually) of the training set into three discrete scale spaces. Templates are then generated for each scale space separately. Similarly, a third configuration consisting of five discrete scales. As can be seen in Figure 4.11, the use of explicit scales in training does improve performance (as expected), however, there are diminishing returns as the number of discrete scale spaces

4.3.4 Positive Examples Only

A final round of experiments was geared towards quantifying the classification rate improvement obtained by explicitly including negative examples in the template training process. For this purpose we compare the classification rate achieved by a set of templates generated based on twenty positive examples of each action class and the classification rate of a set of templates trained on the same set of positive examples in addition to another twenty negative samples of each class.

The results of this comparison can be observed in Figure 4.9-a and 4.9-c, where we see an increase in the inter-class separation between actions that share similar dynamics such as "Walk" and "Run" as well as "Open Car door" and "Open Car Trunk". On average this increased class separation between similar actions achieved by including negative samples leads to a 7.51% improvement in classification rates.

4.4 Conclusion

We explored the role of template-based action recognition in video obtained from UAVs flying at altitudes of over 400 feet. We also explored the explicit incorporation of negative samples in template-based action recognition and found that significant improvements can obtain over positive sample templates. Given the low resolution of humans in aerial videos, we did not rely on high level features. Conversely, low level motion features were forgone due to the extensive amount of jitter of the UAV. Instead, we focused on mid-level motion features which diminished the effect of noisy low-level features in aerial videos. We introduced a new actions from above dataset, which contains numerous actions captured at different flying altitudes, along with their respective annotations.

CHAPTER 5: ACTIONS IN VIDEO SUMMARY

In this chapter we address the third problem of this thesis: the role of the action templates in video summarization. We introduce an activity-specific video summary approach which provides an effective means of browsing and indexing video based on a set of events of interest [90]. We describe a new approach to summarizing video which automatically generates a compact video representation of a long sequence, that features only activities of interest while preserving the general dynamics of the original video.

5.1 Motivation

Every day millions of hours of video are captured around the world by CCTV cameras, webcams, and traffic-cams. In the United States alone, an estimated 26 million video cameras spit out more than four billion hours of video footage every week. In the time it takes to read this sentence, close to 20,000 hours of video have been captured and saved at different locations in the U.S. However, the vast majority of this wealth of data is never analyzed by humans. Instead, most of the video is used in an archival, post-factum manner once an event of interest has occurred.

The main reason for this lack of exploitation resides in the fact that video browsing is very time consuming due to the fact that finding a specific action of interest requires carefully reviewing

hours of data. In most videos, a specific activity of interest may only occur in a relatively small region along the entire spatio-temporal extent of the video.

There exists a large body of work that addresses the topic of activity recognition which focuses mainly on detection in short pre-segmented video clips commonly found in publicly available, standard action datasets. In this work, we attempt to move beyond only performing action detection in an effort to provide a means of generating a compact video representation based on a set of activities of interest, while preserving the scene dynamics of the original video. In our approach, a user specifies which activities interest him and the video is automatically condensed to a short clip which captures the most relevant events based on the user's preference. We follow the output summary video format of non-chronological video synopsis approaches, in which different events which occur at different times may be displayed concurrently, even though they never occur simultaneously in the original video. However, instead of assuming that all moving objects are interesting, priority is given to specific activities of interest which pertain to a user's query. This provides an efficient means of browsing through large collections of video for events of interest.

5.2 Compact Action-based Video Representation

Our approach to generating compact action-specific video representations is composed of three main phases. First, we begin by determining a set of regions in space-time which contain dynamic objects of potential interest. Subsequently, we narrow the pool of potential spatio-temporal regions to be included in the final summary video by detecting specific activities and actions of interest.

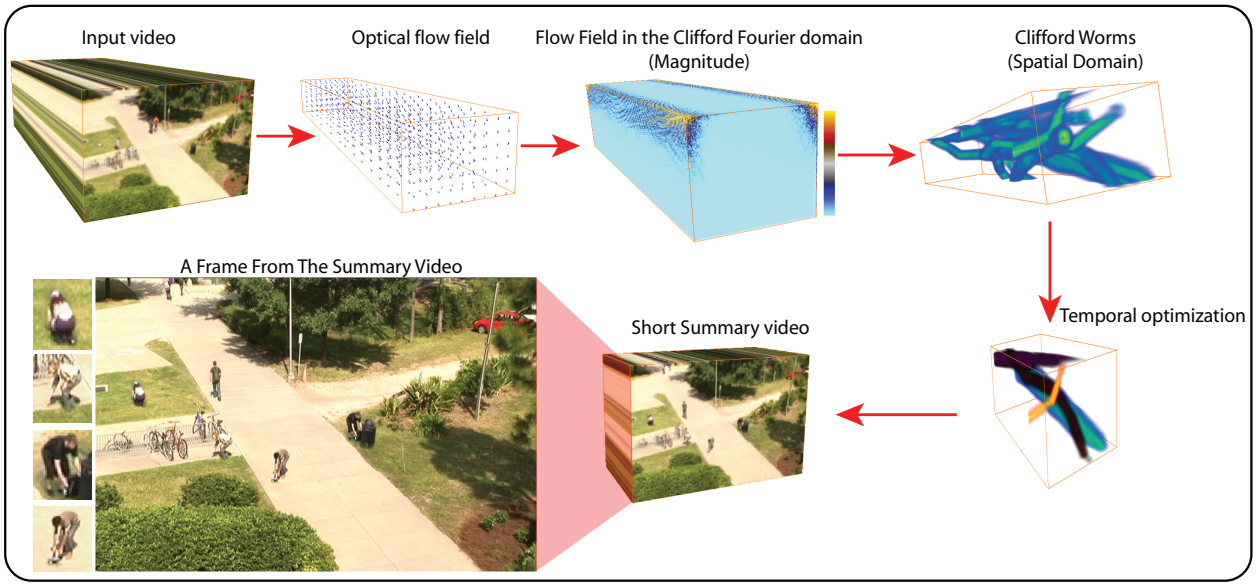


Figure 5.1: A frame from a video summary for the "picking up" action, along with the various steps of the action-specific video synopsis process. Given a long input video sequence (spatio-temporal volume), we compute optical flow and represent the corresponding flow field in the Clifford Fourier domain. Dynamic regions (Clifford worms) are identified within the Clifford domain, and a temporal optimization shifts worms which contain activities of interest in the temporal domain to obtain a compact representation of the original video. Finally, we see the resulting short clip which contains four instances of the "picking up" action of interest.

Finally, we optimize the temporal extent of the video summary via an energy minimization. In the following sections we describe each of these steps in more detail.

5.2.1 Motion Representation

In this section we describe how we identify dynamic regions of a video sequence as potential candidate spatio-temporal locations to be included in the final video summary. For this purpose we begin by computing optical flow for the entire sequence using the flow estimation method

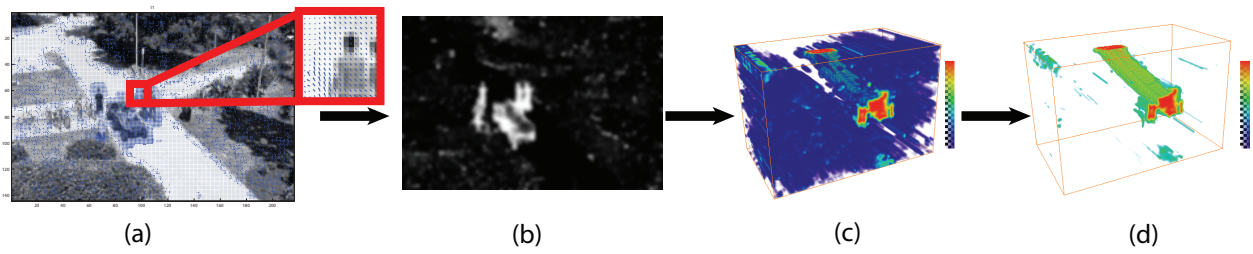


Figure 5.2: (a) The optical flow field for a long video sequence. (b) A 2D slice of the phase spectrum volume (PSV). (c) A 3D segment of the PSV, high values indicate dynamic regions within the flow field. (d) Candidate dynamic regions (worms)

described in [76]. However, instead of identifying dynamic regions within the optical flow in the spatial domain, we efficiently identify such regions in the frequency domain.

Given the fact that we seek to identify salient regions within a 3D optical flow field, where at each point we have two components (d_x, d_y), we cannot employ the traditional Fourier transform which is defined on scalar values as a valid representation without losing any information. In order to efficiently analyze a video sequence in the frequency domain we require an analog to the classical Fourier transform for vector fields. For this purpose, we follow the framework proposed in [30], in that we apply the extension of the original Fourier Transform that is able to deal with vector valued data. This class of Fourier transform is commonly referred to as the "Clifford Fourier transform." Using this embedding, we preserve the full information of relative directions of our vector field while identifying potential regions in space-time which should be included in the summary video.

The Clifford Fourier transform (CFT) for multivectors-valued functions in 3D is defined as:

$$\mathcal{F}\{\mathbf{F}\}(\mathbf{u}) = \int \mathbf{F}(\mathbf{x}) \exp(-2\pi \mathbf{i}_3 \langle x, u \rangle) |d\mathbf{x}|, \tag{5.1}$$

where $\mathbf{i}_3$ represents the analog of a complex number in Clifford algebra, such that $\mathbf{i}_3 = \mathbf{e}_1\mathbf{e}_2\mathbf{e}_3$ and ${\mathbf{i}_3}^2 = -1$. The inverse transform is given by:

$$\mathcal{F}^{-1}\{\mathbf{F}\}(\mathbf{x}) = \int \mathbf{F}(\mathbf{x}) \exp(-2\pi \mathbf{i}_3 \langle x, u \rangle) |d\mathbf{x}|. \tag{5.2}$$

5.2.2 Dynamic Spatio-temporal Regions

Given a long video sequence, we compute optical flow, resulting in a 3D vector field. We employ the Clifford embedding described in section 5.2.1 by performing a 3D Clifford Fourier transform on the optical flow field. In order to identify regions of potential activity of interest for an action-specific video summary, we seek to to carve out a set of spatio-temporal regions, or "worms," from the input flow field which suggest areas of dynamic events. Each worm is, in fact, an object, or group of objects, which carves out a spatio-temporal volume as it moves across the scene over time.

It is well known that the amplitude spectrum of a signal specifies the intensity of the specific sinusoidal components which are present in the signal [38] and the phase provides us with information related to where the components reside inside the original signal, which in our domain corresponds to a flow field. Locations within the flow field which have less periodicity or less homogeneity represent potential dynamic regions of interest in the reconstruction of the flow field, which indicates the location of the worm candidates.

Knowing that the phase spectrum of a flow field in the frequency domain can provide insight as to where dynamic events are occurring in the original video, we identify a set of candidate regions in space-time as follows:

Given a flow field $(\mathbf{u})$ of an input video:

$$f(x,y,t) = \mathcal{F}\{\mathbf{F}\}(\mathbf{u}) \tag{5.3}$$

$$p(x,y,t) = P(f(x,y,t)) \tag{5.4}$$

$$W(x,y,t) = g(x,y,t) * \left|\left|\mathcal{F}^{-1}\{\mathbf{F}\}\left[e^{\mathbf{i}_3 p(x,y,t)}\right]\right|\right| \tag{5.5}$$

where $\mathcal{F}\{\mathbf{F}\}$ and $\mathcal{F}^{-1}\{\mathbf{F}\}$ denote the Clifford Fourier Transform and inverse Clifford Fourier Transform respectively. $P(f(x,y,t))$ represents the phase spectrum of the flow field. $g(x,y,t)$ is a 3D gaussian filter (we use $\sigma = 6$). We obtain potential dynamic regions in the video by converting the phase spectrum into the spatial domain and convolving the resulting scalar field with a gaussian.

The phase spectrum volume (Figure 5.2-c) contains the "innovation" or "pop-out effect" of a specific region in the flow field. Using the Inverse Clifford Fourier Transform, we can construct the output volume which contains primarily the non-trivial, or unexpected spatio-temporal regions of the flow field, where we expect to find events of interest.

Given the phase spectrum volume we segment out a set of "worms" belonging to different objects which trace some movement across the scene over time. We use the normalized cuts toolbox of the algorithm described in [98] to obtain the tightest clusters in this space-time volume. Each spatio-temporal location in the phase spectrum volume (in the Spatial Domain) which is above a threshold forms a node of a completely connected graph. Edge weights are assigned using

Figure 5.3: (a) Frames from the original long video sequence. (b) A non-action-based video summary. (c) An action specific video summary based on the "pickup" action of interest.

the Euclidean distance between connected nodes. Using normalized cuts on this graph we obtain the optimum clustering of dynamic regions in the flow field into a set of worms which can then be shifted individually in time to generate a summary video. In the next section we describe how we narrow the pool of potential worms to be included in the final summary video by detecting specific activities and actions of interest.

5.2.3 Action-Specific Summary

In this work we are interested in compact action-specific video representations. For example, in a parking lot scene, we may be interested in a brief video clip containing all the people entering cars during some time period. Similarly, we may be interested in a quick summary video containing all people who were running through a given scene over the course of a week's worth of video.

In order to generate action-specific summaries we identify the most relevant activities based on pre-defined action templates. Worms containing activities of interest are included in a summary video of a short temporal extent. So as to make the video summary short we need to incorporate many instances of the activity of interest in each frame despite the fact that they may have occurred at different times in the original video.

5.2.3.1 Identifying Activities of Interest

We identify dynamic spatio-temporal regions which contain specific activities of interest by correlating worms with a pool of action templates. Actions such as "run," "walk," "open car door," and "load/unload car trunk" are captured using templates synthesized using a recently proposed [91] generalization of the traditional maximum average correlation height filter to video (3D spatiotemporal volume), and vector value data such as optical flow.

Action templates for activities of interest are generated by computing optical flow of training examples and representing each of the corresponding 3D vector fields in the Clifford Fourier domain using the embedding described in section 5.2.1. Given the resulting volumes in the Clifford Fourier domain, we proceed to convert the resulting 3D CFT matrix into a column vector by concatenating all the columns of the 3D matrix, resulting in a single column-vector (x_i). This process is repeated for each example of an activity of interest. Finally, the template for a given activity of interest can be generated in the Clifford Fourier domain by minimizing:

$$h = (\alpha C + \beta D_x + \gamma S_x)^{-1} m_x, \tag{5.6}$$

where m_x is the mean of all the x_i vectors, and h is the template in vector form in the Clifford Fourier domain. C is the diagonal noise covariance matrix of size $d \times d$, where d is the total number of elements in x_i vector. Given that we do not have a specific noise model for a scene, we set $C = \sigma^2 I$, where σ is the standard deviation parameter and I is a $d \times d$ identity matrix. D_x is also a $d \times d$ diagonal matrix representing the average power spectral density of the training videos.

Having obtained a one-dimensional template (h) for an activity of interest, we proceed to assemble a complete 3D filter by reshaping and then applying the inverse Clifford Fourier transform. The resulting matrix constitutes the template, H, for the particular activity of interest.

Once a set of action templates has been generated, we can proceed to identify worms (dynamic regions) which contain activities of interest and should therefore be included in a summary video.

We determine the likelihood that each worm contains a specific activity of interest by correlating the corresponding template with the spatio-temporal regions in the optical flow field represented by the pool of potential worms described in section 5.2.2. This step is performed in the Clifford Fourier domain; therefore, it amounts to a Clifford multiplication and avoids the high computational cost commonly incurred in template-based approaches. For each potential worm, we normalize the response of the filter to lie within 0 and 1. This normalization is then used as a level of confidence in a pseudo-probabilistic manner to determine which worms contain activities of interest. In this work, we use the response of the template for a given worm in the temporal extent optimization phase in order to give higher priority to worms which are likely to contain events of interest which pertain to a user's query.

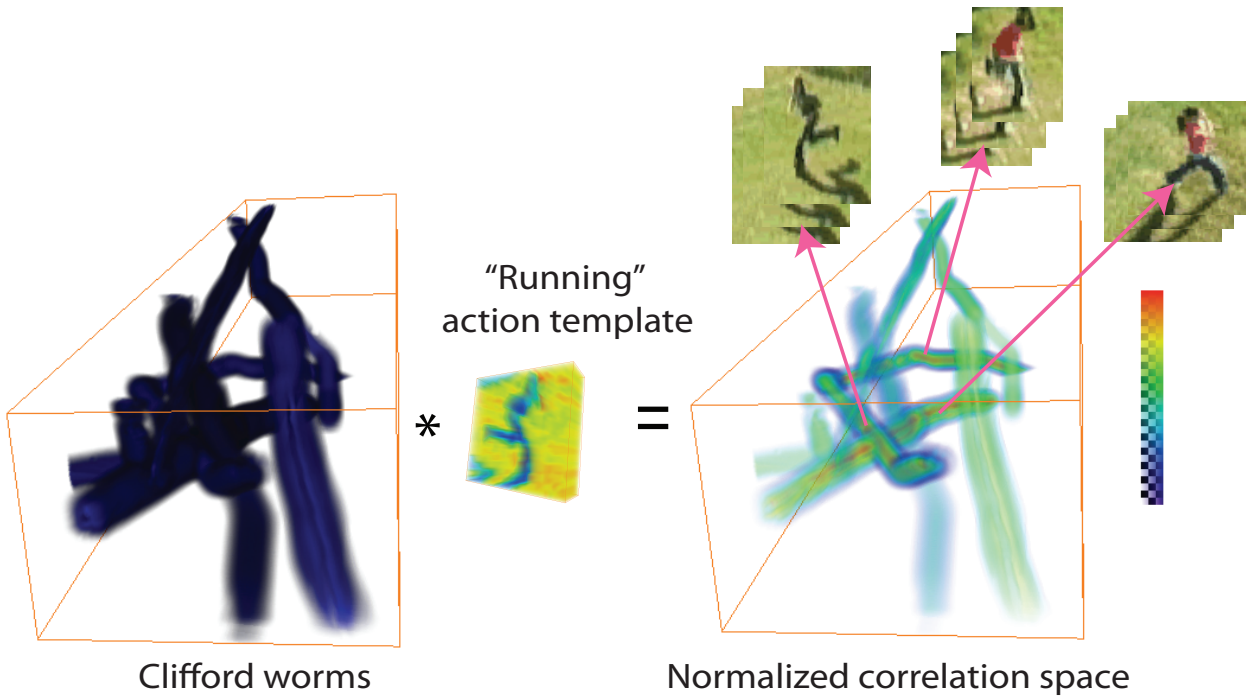


Figure 5.4: We narrow the pool of potential worms to be included in the final summary video by determining the likelihood that worms contain specific activities and actions of interest.

5.2.4 Temporal Extent Optimization

The final summary video is made based on a collection of shifts in time (S) which map spatio-temporal regions to a different time in a summary video such that a more compact representation of the original sequence can be obtained. Given a set of worms (W), we define an action-based video summary using the following energy function:

$$E(S) = \sum_{w \in W} \alpha E_a(w) + \sum_{w,w' \in W} E_o(w, w') \tag{5.7}$$

where E_a is the cost associated with excluding a particular worm which has a high likelihood of containing an action of interest, α is the weight of the activity of interest term, and E_o is the

spatio-temporal overlap cost. E_a is given by

$$E_a(w) = \frac{1}{\left(\sum_{l,m,n} c(l,m,n)\right) D_w^{-1}} \tag{5.8}$$

where D_w is the number of pixels in worm w, and $c(l,m,n)$ corresponds to the response of the template of a particular activity of interest along a dynamic spatio-temporal region, it is given by:

$$c(l,m,n) = \sum_{t=0}^{N-1} \sum_{y=0}^{M-1} \sum_{x=0}^{L-1} s(l+x, m+y, n+t) H(x,y,t), \tag{5.9}$$

where s is the spatio-temporal region represented by a given Clifford worm (w) of the input video, H is the template of the action of interest (h is its Fourier transform), and L, M, and N are the dimensions of H in the x (horizontal), y (vertical), and t (time) directions, respectively. Furthermore, l, m, and n correspond to the spatio-temporal locations of the dynamic region. The spatio-temporal overlap cost E_o penalizes regions in the video which contain events of interest that are mapped to new temporal locations which results in some degree of overlap between them. It is given by the total number of spatio-temporal worm collisions, weighted by the likelihood that an event of interest occurs within the respective regions. We minimize the temporal extent energy function efficiently using the multi-label graph cut method described in [57], where labels correspond to time shifts of worms and a cut in the graph represents a specific time shift. The result is an optimal set of time shifts which maximize the number of worms included which contain a desired activity of interest while minimizing the amount of spatio-temporal overlap between worms. The total temporal duration of the action-based summary clip is a parameter which is manually selected during the graph cut energy minimization. In most of our experiments the value ranged from thirty seconds to a

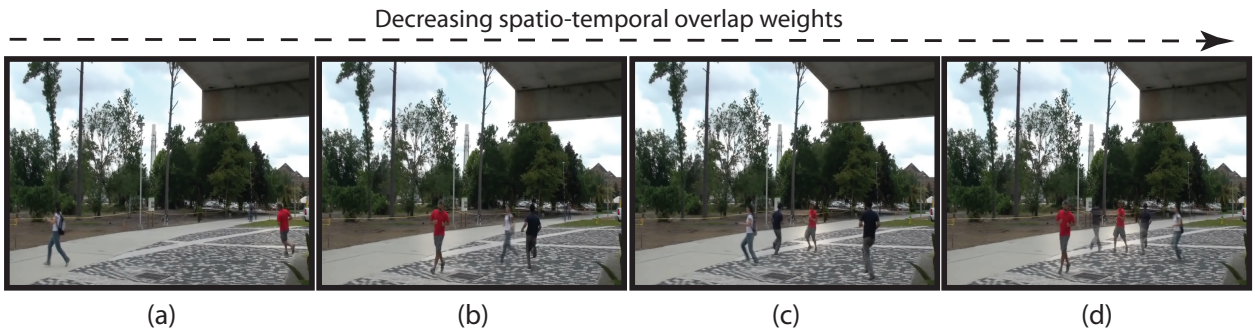


Figure 5.5: Decreasing the weight of the spatio-temporal overlap cost leads to increasingly compact summaries at the cost of additional overlaps. (a) $\alpha = 0.6$, (b) $\alpha = 0.5$, (c) $\alpha = 0.4$, (d) $\alpha = 0.3$

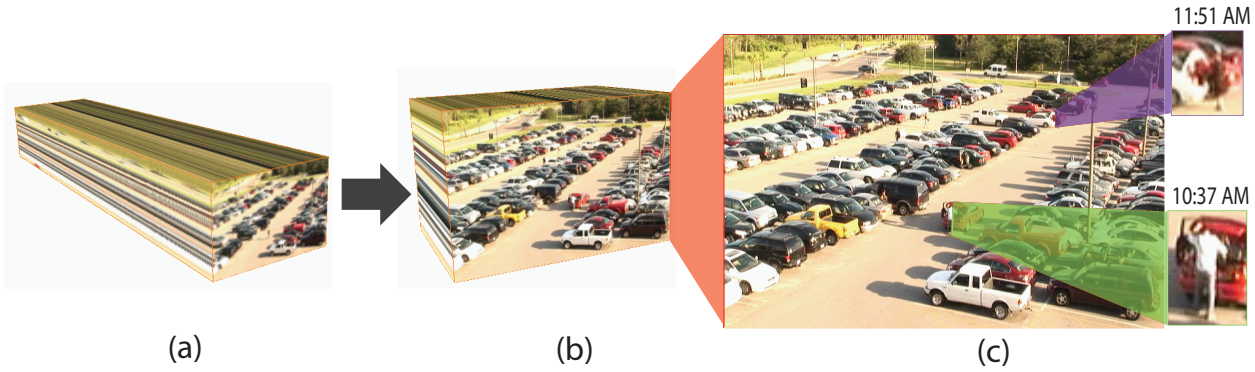


Figure 5.6: A video summary of "opening trunk" events in a parking lot. (a) A two hour long video sequence is summarized in a one minute clip (b) containing most of the instances of the event of interest ("opening trunk"). The video summary displays multiple instances of the event of interest (which may have occurred at different times) concurrently (c).

minute. Once the temporal shifts for each worm are defined we shift the worms in time. In order to minimize visual artifacts and seams we blur the edges of the worm masks.

5.3 Experiments and Results

We performed a number tests to better understand the ability of the proposed method to cope with a range of video sources. Details about the video sources used in generating the action-specific summaries and the experiments performed are given below.

5.3.1 Ground Camera Videos

In the first round of experiments a collection of videos obtained from ground cameras which included parking lot scenes and street scenes wwas used to generate video summaries of activities of interest. The video corpus contained a total of five hours of video divided across six different clips. Activities of interest in these experiments were defined to be "running," "picking up an object," "entering vehicle," and "loading/unloading trunk."

Each activity of interest occurs multiple times at different points within the collection of long video sequences. "Running" occurs 28 times, "picking up an object" occurs 19 times, "entering vehicle" occurs 38 times, and "loading/unloadding trunk" occurs 23 times.

Figure 5.1 demonstrates the effect of generating a video summary based on the "picking up object" activity of interest. A six hour long video which contains only five instances of the action of interest is represented by a short, one minute clip, containing most of the instances of the event of interest. In this example of our results we see how four different instances of the "picking up object" action are displayed concurrently, despite the fact that they have occurred over an extended time.

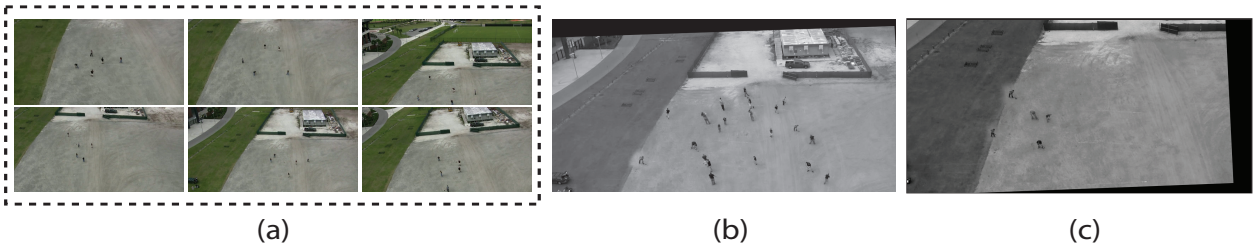


Figure 5.7: (a) Frames of a 12 minute aerial video sequence shot from an R/C helicopter flying at 400 feet. (b) A non-action based video summary.(c) A video summary of the "digging" action.

The value of an action-specific video summary is evident in Figure 5.3. In this experiment we generated a 10 second video summary based on the most dynamic spatio-temporal regions (worms) which results in a short yet cluttered video clip (Figure 5.3-b) given that all of the spatio-temporal dynamic regions are treated equally. When we employ the action-specific video summary framework using the "picking up" action of interest the resulting clip consists of relevant events (two people picking up an object) and is considerably less cluttered. In long videos of crowded scenes where moving objects abound action-specific video summaries provide a means of distilling the long sequence into a short clip that clearly depicts events of interest that occurred over a period of time.

A more challenging scenario is seen in Figure 5.6, where we have a busy parking lot scene which contains many motions which can potentially be irrelevant to a given user. Therefore, it may not be appropriate to generate a synopsis based on all moving objects in the scene. In this experiment, we generated a video synopsis of a three hour long video clip based on the "open vehicle trunk" event of interest. Despite the fact that instances of the event of interest are relatively small as compared to the rest of the scene (typically no greater than 15×15 pixels), our summary

includes seven out of the total eight instances of the event of interest in a one one minute clip. Searching for this particular event manually would require careful observation as the video is fast-forwarded, a time consuming and inefficient process.

A similar cluttered scenario we consider is depicted in Figure 5.10, where we condense all of the instances of the running action which occur throughout a long traffic sequence into a one minute clip. Given that in this particular scene, running pedestrians tend to occur in one particular region in the video (the crosswalk), we increase the weight of the spatio-temporal overlap cost term (by setting α to 0.3) in order to avoid artifacts caused by multiple running activities which are mapped to the same spatio-temporal region.

The effect of varying the spatio-temporal overlap cost is depicted in our experiment in Figure 5.5, in which a ten second summary of the "running" action is obtained from a 21 minute input video. As we lower the weight (α) of the spatio-temporal overlap cost we observe how additional instances of the running action are included in the video summary resulting in additional clutter.

5.3.2 Aerial Videos

A second round of experiments was based on aerial video sequences obtained using a UAV equipped with an HD camera mounted on a gimbal. Videos were recorded at a flying altitude of over 400 feet. The collection contains a diverse pool of events such as people getting into vehicles, and people running, which occur over the course of one hour. These videos are divided into sequences which typically average 12 minutes in length. In these experiments our goal is to evaluate the ability to generate video synopses based on moving aerial camera video sequences. Given an

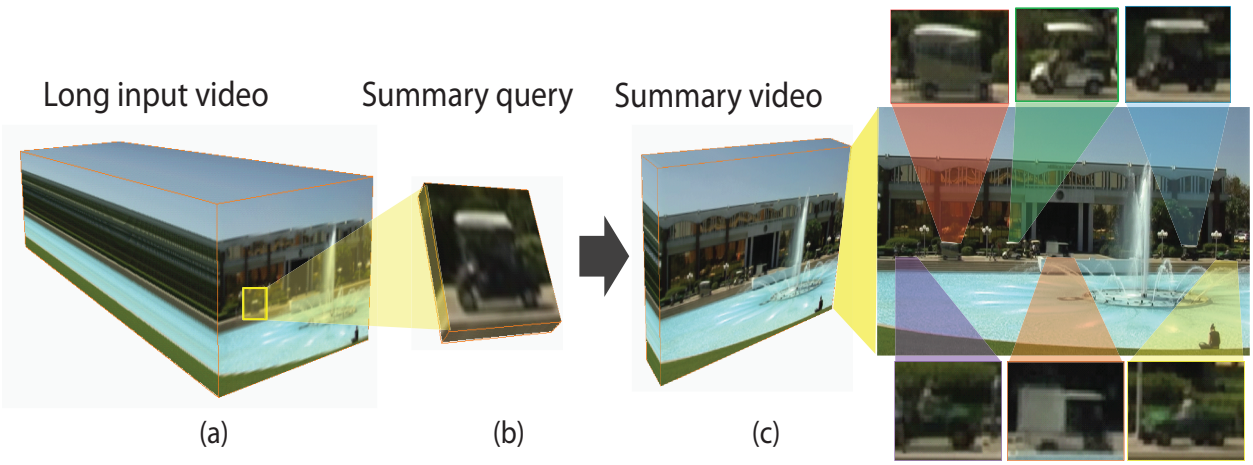


Figure 5.8: Summary by example: given a long video sequence (a), we specify a spatio-temporal region as a query (b) which contains an event of interest. A video summary (c) which includes events in the scene which match the query is then automatically generated.

aerial video sequence obtained by a UAV hovering over a region of interest, we begin the summary process by performing frame-to-frame registration across the sequence. Subsequently, we identify dynamic regions within the registered aerial video, identify events of interest, and perform temporal optimization using the methods described in Sections 5.2.3.1, and 5.2.4 respectively.

Figure 5.11 depicts an example of a 30 second video summary generated from an aerial video sequence (which is 9 minutes long) based on the "running" action of interest. The short summary clip contains four out of the seven running events which occur in the scene over the entire duration of the longer clip.

Another aerial action video summary is depicted in Figure 5.7. In this experiment two video summaries are generated from a 12 minute aerial video. The first is a 30 second non-action video

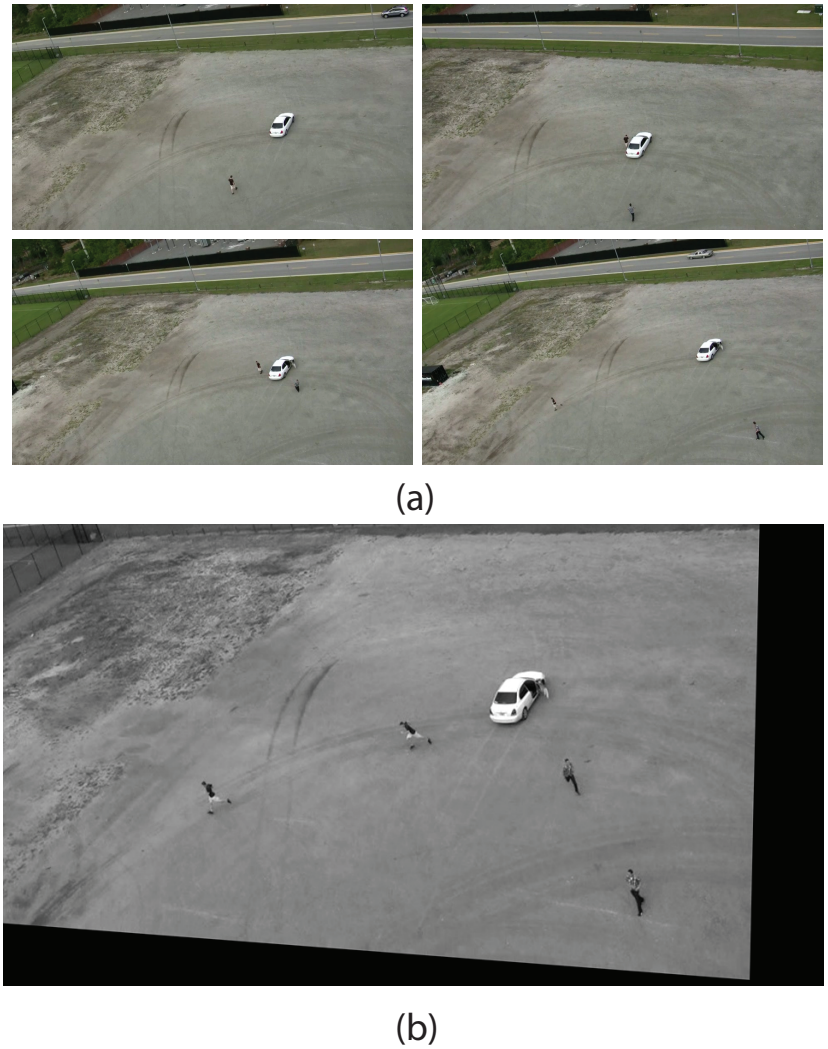


Figure 5.9: (a) Frames of a 13 minute aerial video sequence. (b) A video summary of the "running" action.

summary. As can be seen in Figure 5.7-b, this results in a cluttered video clip that contains various movers that originate from different spatio-temporal regions. Figure 5.7-c depicts an action-based video summary of the same length, that contains five instances of the digging action which occur at different point in time in the original video. Due to small out-of-plane parallax errors which are

propagated over time, a modest amount of drift in alignment is accumulated which results in some visible artifacts around some of the shifted action instances.

In our experiments we observed that the main issues related to generating video summaries of aerial sequences are noise in the flow field which is caused by slight errors in the motion compensation. This leads to noisy dynamic regions which are sometimes included in the final summary video. This effect can be observed in Figure 5.9 in which a ten second video summary of the running action is generated from a 13 minute aerial video. Due to nosy worms individual instances of an action have been segmented into disjoint events which are then shifted in time independently. As can be seen in Figure 5.9-b two seperate running instances have been segmented into four small running segments which are depicted concurrently in the short summary clip.

5.3.3 Summary by Example

It is not always possible to obtain a large training set for a collection of events of interest. Nor is it feasible to assume that we are only interested in video summaries of a static set of pre-defined events (such as running, opening car door, etc). Therefore, in our last round of experiments we introduce the concept of "summary by example." That is, given a long video sequence, a user can select any instance of a particular event of interest by specifying a spatial region in the video and the temporal extent of the event. Subsequently, a short video summary which contains all the events that match the selected query is generated for the rest of the long sequence.

Summary by example can be accomplished without any major changes to the overall approach described above. This is due to the fact that we treat the example of the event of interest as a

single instance spatio-temporal template. In order to account for the possibility of observing the event of interest at different scales across the long video, we synthesize templates at three scales by resizing the original example. Aside from employing this special case of the spatio-temporal template, the remaining steps of our approach remain the same. Figure 5.8 depicts a summary by example, in which an event of interest consisting of a moving golf cart is selected within a long video sequence. Based on this example of an event of interest, our method generates a short thirty second video summary which condenses six separate instances of a moving golf cart event which occurs at different times within the long video.

5.4 Conclusion

In this chapter we have explored the role of template-based action recognition methods in generating short video summaries of long ground camera videos and aerial videos. We do not consider all moving objects in the long video sequence to be of equal importance when generating a given video summary. Instead we focussed on generating video summaries based on a set of events of interest which can be specified when generating a summary. We found that these activity-specific video summaries provide us with a more meaningful way of quickly reviewing a long video for particular events of interest in the form of a short video clip which condenses all activities of interest that have occurred across some time span. Furthermore, by focusing on events of interest instead of moving objects we were able to generate meaningful summaries of crowded scenes. As future work we intend to explore multi-agent events with long-range spatio-temporal dependen-

cies. We also intend to use confidence values of action detection to draw attention to specific areas in the summary.

(a)

(b)

Figure 5.10: (a)Frames from a long cityscape video. (b) A frame from a short clip generated by our system which captures instances of running in the scene over an extended period of time.

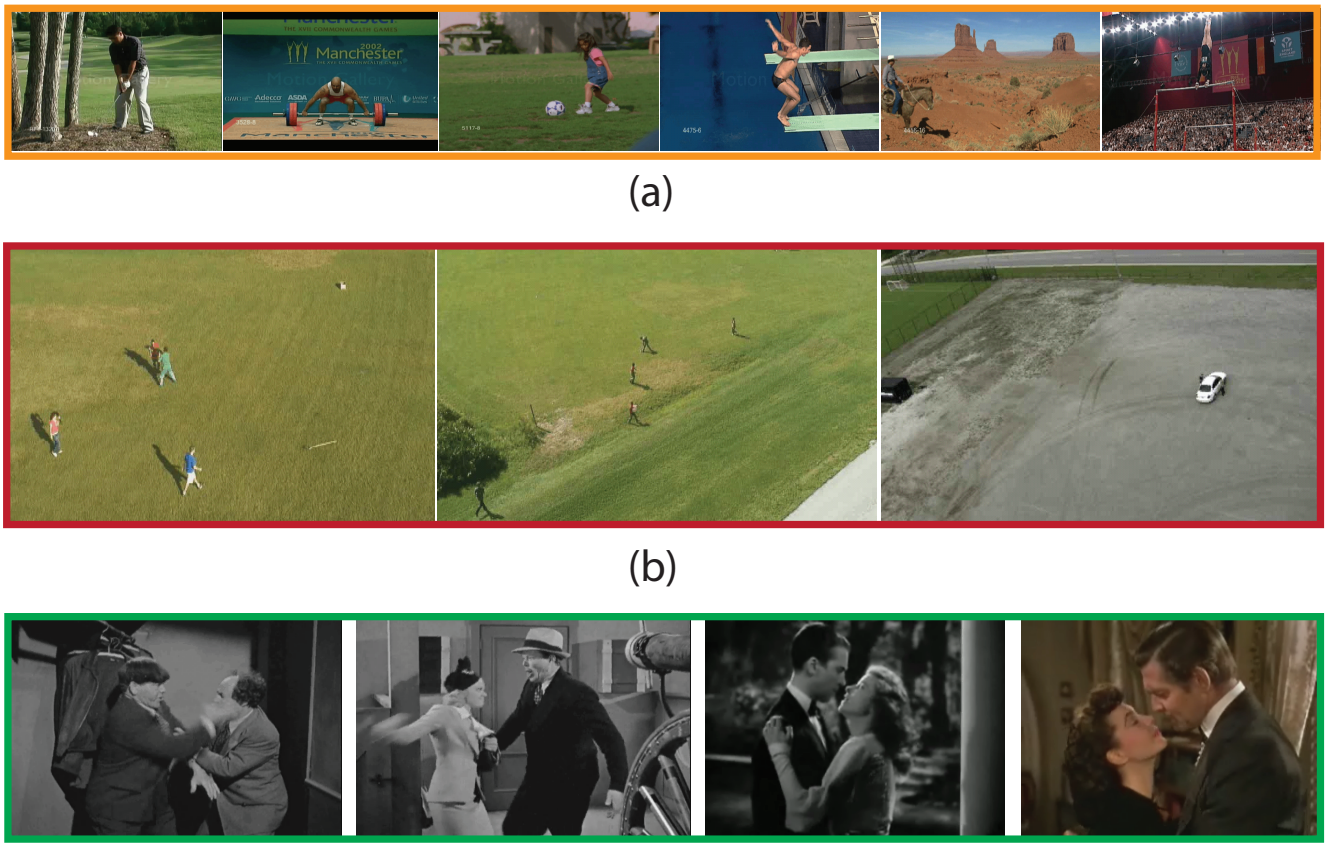


Figure 5.11: UAV aerial video summary containing the “running” action. Four instances of the running action which occur at different time instances across a long video are displayed concurrently.

CHAPTER 6: CONCLUSION AND FUTURE WORK

In this dissertation we have addressed three unexplored challenges in template-based action recognition. The first problem we addressed was the ability to generalize from positive exemplars of an action class in order to effectively generate a single action template which captures the intra-class variability of an action using a collection of examples. We showed that by extending traditional maximum average correlation height filters to include actions one can effectively classify actions in unconstrained environments.

The second problem we addressed was an exploration of correlation filter paradigms that are capable of explicitly incorporating not only positive samples of an action class but also negative exemplars. We demonstrated how a quadratic spatio-temporal action template framework is capable of generalizing from multiple examples of positive instances of an action class as well as a set of negative examples present in a dataset. By employing both positive and negative samples of the training set we were able to achieve significant increases in classification accuracy.

The third problem we addressed was the generation of video summaries that are specific to an action of interest. We showed how template-based action recognition methods can provide as basis for generating compact video representation based on a set of activities of interest, while preserving the scene dynamics of the original video. Through an extensive set of experiments

we demonstrated how action-specific video summaries could be generated from both ground and aerial videos.

Further discussion and future directions are discussed in Section 6.2. We briefly summarize the key contributions of this dissertation in the next section.

6.1 Summary of Contributions

The main contributions of this research to the literature include:

1. Action Recognition Using Temporal Templates

    (a) A new template-based action recognition approach that is capable of incorporating multiple labeled examples to generate a single template that is capable of capturing the variability associated with an action class.

    (b) Extended traditional MACH filters to include vector-valued data by employing the Clifford Fourier Transform, which generalizes the traditional Fourier transform by including vector-valued data.

    (c) Addressed the incorporation of multiple positive and negative examples of a target action class when generating an action template by employing the Fukunaga-Koontz Transform.

    (d) Explored a range of low-level and mid-level motion features and assessed their effectiveness within the context of aerial action recognition.

2. Video Summarization

   (a) Introduced activity-specific video summaries which provide an effective means of browsing and indexing video based on a set of events of interest. The method that was introduced automatically generates a compact video representation of a long sequence, which features only activities of interest while preserving the general dynamics of the original video.

   (b) Demonstrated that video summaries of cluttered scenes can be generated effectively using action templates to select specific moving regions.

   (c) Explored the generation of non-chronological video summaries of aerial videos.

6.2 Future Work

In this section we present some possibilities for future directions to further the research that was carried out in this dissertation.

In Chapters 3 and 4 we explored the generation of action templates which focussed on single-agent activities such as ”running” and ”digging”. Future efforts in template-based action recognition can focus on extending action-templates to include multi-agent activities such as ”meeting”, ”hugging”, ”shaking hands”, and ”fighting”. Furthermore, crowd behaviors such as ”panic”, ”bottlenecks”, and ”queuing” behaviors can also be explored using the paradigm of action templates.

Future efforts in action-based video summarization can include moving beyond simply generating short summary clips of long videos. Instead, future work on this problem can focus on

the role of video summaries in content-based video retrieval. Most of the content-based video retrieval community has focussed on modeling long video sequences via a set of low-level features which are expensive to compute for very long video sequences. Continued effort can focus on using video summaries of long videos as compact signatures that can be used to efficiency retrieve similar video sequences.

LIST OF REFERENCES

[1] J. Aggarwal and Q. Cai. Human motion analysis: A review. *CVIU*, 73(3), 1999.

[2] J. Aggarwal and S. Park. Human motion: modeling and recognition of actions and interactions. *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pages 640–647, 2004.

[3] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Articulated and elastic non-rigid motion: A review. *Workshop on Motion of Non-Rigid and Articulated Objects*, 1994.

[4] O. Alatas, P. Yan, and M. Shah. Spatio-Temporal Regularity Flow (SPREF): Its Estimation and Applications. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(5):584–589, 2007.

[5] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. *IEEE Conference on Computer Vision and Pattern Recognition*, 2, 2005.

[6] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto. Recognition of human gaits. *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[7] M. Black, D. Fleet, and Y. Yacoob. Robustly estimating changes in image appearance. *Computer Vision and Image Understanding*, 78(1):8–31, 2000.

[8] M. J. Black, Y. Yacoob, A. D. Jepson, and D. Fleet. Learning parameterized models of image motion. *IEEE Conference on Computer Vision and Pattern Recognition*, 1997.

[9] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE International Conference on Computer Vision*, 2005.

[10] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. *IEEE International Conference on Computer Vision*, 2, 2005.

[11] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. In *IEEE International Conference on Computer Vision*, volume 2, 2005.

[12] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *IEEE International Conference on Computer Vision*, volume 2, 2005.

[13] A. Bobick and J. Davis. Real-time recognition of activity using temporal templates. *IEEE Workshop on Applications of Computer Vision (WAVC)*, 1996.

[14] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *PAMI*, 23(3):257–267, 2001.

[15] A. Bobick and J. Davis. The Recognition of Human Movement Using Temporal Templates. *TPAMI*, 2001.

[16] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *TPAMI*, 23(3), 2001.

[17] A. F. Bobick and A. D. Wilson. A state-based technique for the summarization and recognition of gesture. *IEEE Conference on Computer Vision and Pattern Recognition*, 1995.

[18] P. Bone, R. Young, and C. Chatwin. Position-, rotation-, scale-, and orientation-invariant multiple object recognition from cluttered scenes. *Optical Engineering*, 45:077203, 2006.

[19] L. W. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. *IEEE Conference on Computer Vision and Pattern Recognition*, 1995.

[20] S. Carlsson and J. Sullivan. Action recognition by shape matching to key frames. *Workshop on Models Versus Exemplars in Computer Vision*, 2001.

[21] C. Cedras and M. Shah. Motion based recognition: A survey. *Image and Vision Computing*, 1995.

[22] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2009.

[23] J. Chen, M. Kim, Y. Wang, and Q. Ji. Switching gaussian process dynamic models for simultaneous composite motion tracking and recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2009.

[24] K. M. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[25] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. Huang. Facial expression recognition from video sequences: temporal and static modeling. *CVIU*, 91(1-2):160–187, 2003.

[26] T. Darrell and A. Pentland. Classifying hand gestures with a view-based distributed representation. *NIPS*, 1993.

[27] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *IEEE International Workshop on VS-PETS*, 2005.

[28] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72, 2005.

[29] J. Ebling and G. Scheuermann. Clifford Fourier transform on vector fields. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):469–479, 2005.

[30] J. Ebling and G. Scheuermann. Clifford Fourier transform on vector fields. *VCG*, 11(4), 2005.

[31] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *IEEE International Conference on Computer Vision*, pages 726–733, 2003.

[32] I. Essa and A. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *PAMI*, 19(7):757–763, 1997.

[33] W. T. Freeman and M. Roth. Orientation histogram for hand gesture recognition. *International Workshop on Automatic Face and Gesture Recognition*, 1995.

[34] D. M. Gavrila. The visual analysis of human movement: A survey. *CVIU*, 1999.

[35] P. Hennings-Yeomans, B. Kumar, and M. Savvides. Palmprint Classification Using Multiple Advanced Correlation Filters and Palm-Specific Segmentation. *Information Forensics and Security, IEEE Transactions on*, 2(3 Part 2):613–622, 2007.

[36] P. Hong, M. Turk, and T. Huang. Gesture modeling and recognition using finite state machines. In *ICGR*, 2000.

[37] B. Horn and B. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[38] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[39] X. Huo. A statistical analysis of Fukunaga-Koontz transform. *IEEE signal processing letters*, 11(2):123–126, 2004.

[40] P. P. Ivan Laptev. Retrierving actions in movies. *IEEE International Conference on Computer Vision*, 2007.

[41] H. Jiang, M. Drew, and Z. Li. Successive Convex Matching for Action Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[42] H. Jiang, M. Drew, and Z. Li. Successive Convex Matching for Action Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[43] H. Jiang, M. S. Drew, and Z. N. Li. Successive convex matching for action detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[44] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. *International Conference on Automatic Face and Gesture Recognition*, 1996.

[45] T. Kanade and J. Tian. Comprehensive database for facial expression analysis. *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53, 2000.

[46] A. Kapoor, Y. Qi, and R. Picard. Fully automatic upper facial action recognition. *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*, pages 195–202, 2003.

[47] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[48] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *IEEE International Conference on Computer Vision*, 2005.

[49] Y. Ke, R. Sukthankar, and M. Hebert. Efficient Visual Event Detection Using Volumetric Features. *IEEE International Conference on Computer Vision*, 1, 2005.

[50] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. *IEEE International Conference on Computer Vision*, 2007.

[51] Y. Ke, R. Sukthankar, and M. Hebert. Event Detection in Crowded Videos. In *IEEE International Conference on Computer Vision*, 2007.

[52] Y. S. R. H. M. Ke. Spatio-temporal Shape and Flow Correlation for Action Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[53] C. Kim and J. Hwang. An integrated scheme for object-based video abstraction. In *Proceedings of the eighth ACM international conference on Multimedia*, pages 303–311. ACM New York, NY, USA, 2000.

[54] C. Kim and J. Hwang. An integrated scheme for object-based video abstraction. In *ACM Multimedia*, 2000.

[55] Y. Kim, A. Martínez, and A. Kak. Robust motion estimation under varying illumination. *Image and Vision Computing*, 23(4):365–375, 2005.

[56] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 2:1137–1145, 1995.

[57] V. Kolmogorov and R. Zabih. What Energy Functions Can Be Minimized via Graph Cuts? *TPAMI*, 2004.

[58] W. H. L Wang and T. Tan. Recent development in human motion analysis. *Pattern Recognition*, 2003.

[59] I. Laptev. On Space-Time Interest Points. *International Journal of Computer Vision*, 2005.

[60] I. Laptev. On Space-Time Interest Points. *International Journal of Computer Vision*, 2005.

[61] I. Laptev. Space time interest points. *International Journal of Computer Vision*, 2005.

[62] P. Leopardi. The GluCat home page. *HTML document*, 2002.

[63] J. Little and J. E. Boyd. Recognizing people by their gait: The shape of motion. *Journal of Computer Vision Research*, 1998.

[64] A. Mahalanobis, B. Kumar, S. Song, S. Sims, and J. Epperson. Unconstrained correlation filters. *Appl. Opt*, 33(33):3751–3759, 1994.

[65] T. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *CVIU*, 81(3), 2001.

[66] T. B. Moeslund and E. Granum. A survey of computer vision based human motion capture. *CVIU*, 2001.

[67] G. Mori and J. Malik. Estimating human body configurations using shape context matching. *European Conference on Computer Vision*, 2002.

[68] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering human body configurations: Combining segmentation and recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[69] G. Mori, X. Ren, A. Efros, and J. Malik. Recovering Human Body Configurations: Combining Segmentation and Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[70] M. Moslemi Naeini. Clustering and visualizing actions of humans and animals using motion features. *MS Thesis*, 2008.

[71] J. Nam and A. Tewfik. Video abstract of video. In *3rd IEEE Workshop on Multimedia Signal Processing*, pages 117–122, 1999.

[72] J. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[73] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *BMVC*, 6(8), 2006.

[74] A. Oikonomopoulousm, I. Patras, and M. Pantic. Spatiotemporal saliency for human action recognition. *IEEE ICME*, 2005.

[75] C. Pal and N. Jojic. Interactive montages of sprites for indexing and summarizing security video. In *Conference on Computer Vision and Pattern Recognition*, volume 2, 2005.

[76] N. Papenberg, A. Bruhn, and Brox. Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision*, 67(2):141–158, 2006.

[77] V. Parameswaran and R. Chellappa. View invariants for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2003.

[78] V. Parameswaran and R. Chellappa. View invariance for human action recognition. *International Journal of Computer Vision*, 2006.

[79] V. Parameswaran and R. Chellappa. View invariance for human action recognition. *International Journal of Computer Vision*, 66(1):83–101, 2006.

[80] P. Perez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on Graphics*, 22(3):313–318, 2003.

[81] N. Petrovic, N. Jojic, and T. Huang. Adaptive video fast forward. *Multimedia Tools and Applications*, 26(3):327–344, 2005.

[82] N. Petrovic, N. Jojic, and T. Huang. Adaptive Video Fast Forward. *Multimedia Tools and Applications*, 26(3), 2005.

[83] R. Polana and R. Nelson. Low level recognition of human motion (or how to get your manwithout finding his body parts). *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on*, pages 77–82, 1994.

[84] A. Pope, R. Kumar, H. Sawhney, and C. Wan. Video Abstraction. In *ACSS*, 1998.

[85] A. Pope, R. Kumar, H. Sawhney, and C. Wan. Video abstraction: Summarizing video content for retrieval and visualization. In *Asilomar Conference On Signals Systems And Computers*, volume 1, pages 915–919. Computer Society Press, 1998.

[86] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg. Webcam synopsis: Peeking around the world. *IEEE International Conference on Computer Vision*, 2007.

[87] Y. Pritch, A. Rav-Acha, and S. Peleg. Non-Chronological Video Synopsis and Indexing. *TPAMI*, 2008.

[88] A. Rav-Acha, Y. Pritch, and S. Peleg. Making a long video short: Dynamic video synopsis. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 435–441. Citeseer, 2006.

[89] P. Refregier. Optimal trade-off filters for noise robustness, sharpness of the correlation peak, and Horner efficiency. *Optics Letters*, 16(11):829–832, 1991.

[90] M. Rodriguez. CRAM: Compact Representation of Actions in Movies. In *IEEE Conference on Computer Vision and Pattern Recognition, 2010. CVPR 2010*, pages 1–8, 2010.

[91] M. Rodriguez, J. Ahmed, and M. Shah. Action MACH. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[92] J. O. Roukre and N. Badler. Model based image analysis of human motion using constrained propagation. *IEEE PAMI*, 1980.

[93] M. Savvides and B. Kumar. Quad phase minimum average correlation energy filters for reduced memory illumination tolerant face authentication. *Audio and Visual Biometrics based Person Authentication (AVBPA)*, 2003.

[94] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, volume 3, 2004.

[95] L. Schuldt and Caputo. Recognition of human actions. *ICPR*, 2004.

[96] E. Shechtman and M. Irani. Space-time behavior based correlation. *IEEE Conference on Computer Vision and Pattern Recognition*, 1, 2005.

[97] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the Space of a Human Action. In *IEEE International Conference on Computer Vision*, volume 1, 2005.

[98] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *TPAMI*, 2000.

[99] S. Sims and A. Mahalanobis. Performance evaluation of quadratic correlation filters for target detection and discrimination in infrared imagery. *Optical Engineering*, 43:1705, 2004.

[100] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov model. *Computational Imaging and Vision*, 1997.

[101] J. Sullivan and S. Carlsson. Recognizing and Tracking Human Action. In *European Conference on Computer Vision*, 2002.

[102] T. Syeda-Mahmood and Vasilescu. Recognizing action events from multiple viewpoints. In *Events in Video*, 2001.

[103] T. Syeda-Mahmood, A. Vasilescu, and S. Sethi. Recognizing action events from multiple viewpoints. In *IEEE Workshop on Detection and Recognition of Events in Video*, 2001.

[104] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *PAMI*, 27(3):475–480, 2005.

[105] P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[106] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *CVIU*, 2006.

[107] C. Xie, B. Kumar, S. Palanivel, and B. Yegnanarayana. A Still-to-Video Face Verification System Using Advanced Correlation Filters. *International Conference on Biometric Authentication*, pages 102–108, 2004.

[108] Y. Yacoob and M. Black. Parameterized modeling and recognition of activities. In *IEEE International Conference on Computer Vision*, 1998.

[109] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. *IEEE Conference on Computer Vision and Pattern Recognition*, 1998.

[110] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time sequential images using hidden markov model. *IEEE Conference on Computer Vision and Pattern Recognition*, 1992.

[111] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hiddenMarkov model. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–385, 1992.

[112] A. Yilmaz and M. Shah. Actions As Objects: A Novel Action Representation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[113] A. Yilmaz and M. Shah. Actions sketch: A novel action representation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[114] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. *IEEE Conference on Computer Vision and Pattern Recognition*, 1, 2005.

[115] H. Zhou and T. Chao. MACH filter synthesizing for detecting targets in cluttered environment for gray-scale optical correlator. *Optical pattern recognition X*, pages 394–398, 1999.

[116] X. Zhu, X. Wu, J. Fan, A. Elmagarmid, and W. Aref. Exploring video content structure for hierarchical summarization. *Multimedia Systems*, 10(2):98–115, 2004.

[117] X. Zhu, X. Wu, J. Fan, A. Elmagarmid, and W. Aref. Exploring video content structure for hierarchical summarization. *Multimedia Systems*, 10(2), 2004.