HUMAN ACTIVITY RECOGNITION BY GAIT ANALYSIS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BURCU KEPENEKCİ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

FEBRUARY 2011

Approval of the thesis:

HUMAN ACTIVITY RECOGNITION BY GAIT ANALYSIS

submitted by **BURCU KEPENEKCİ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen ____________________
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İsmet Erkmen ____________________
Head of Department, **Electrical and Electronics Engineering**

Prof. Dr. Gözde Akar ____________________
Supervisor, **Electrical and Electronics Eng. Dept., METU**

Examining Committee Members:

Prof. Dr. Neşe Yalabık ____________________
Computer Engineering Dept., METU

Prof. Dr. Gözde Akar ____________________
Electrical and Electronics Engineering Dept., METU

Prof. Dr. Uğur Halıcı ____________________
Electrical and Electronics Engineering Dept., METU

Assist. Prof. Dr. İlkay Ulusoy ____________________
Electrical and Electronics Engineering Dept., METU

Assist. Prof. Dr. Hüsrev Taha Sencar ____________________
Computer Engineering Dept., TOBB University

Date: 10.02.2011

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : BURCU KEPENEKCI

Signature :

ABSTRACT

HUMAN ACTIVITY RECOGNITION BY GAIT ANALYSIS

Kepenekci, Burcu

Ph.D., Department of Electrical and Electronics Engineering

Supervisor: Prof. Dr. Gözde Akar

February 2011, 102 pages

This thesis analyzes the human action recognition problem. Human actions are modeled as a time evolving temporal texture. Gabor filters, which are proved to be a robust 2D texture representation tool by detecting spatial points with high variation, is extended to 3D domain to capture motion texture features. A well known filtering algorithm and a recent unsupervised clustering algorithm, the Genetic Chromodynamics, are combined to select salient spatio-temporal features of the temporal texture and to segment the activity sequence into temporal texture primitives. Each activity sequence is represented as a composition of temporal texture primitives with its salient spatio-temporal features, which are also the symbols of our codebook. To overcome temporal variation between different performances of the same action, a Profile Hidden Markov Model is applied with Viterbi Path Counting (ensemble training). Not only parameters and structure but also codebook is learned during training.

Keywords: 3D Gabor Filters, Action Recognition, Feature Selection, Genetic Chromodynamics, Motion Analysis, Profile HMM, spatio-temporal Features, Temporal Texture.

ÖZ

HAREKET ANALİZİ İLE İNSAN AKTİVİTELERİNİ TANIMA

Kepenekci, Burcu

Doktora, Elektrik ve Elektronik Mühendisliği Bölümü

Tez Yöneticisi: Prof. Dr. Gözde Akar

Şubat 2011, 102 sayfa

Bu çalışmada insan activitelerinin tanınması amacıyla zaman-uzamsal analize dayalı yeni bir yöntem önerilmiştir. Video serisinin local zaman-uzamsal betimlemesi 3B Gabor filtreleri kullanılarak elde edilmiştir. Hareket desenini en iyi modelleyen zaman-uzamsal nitelikler iki aşamalı bir nitelik seçme algoritması ile seçilmiştir. Öncelikle iyi bilinen süzgeçleme yöntemiyle bazı nitelikler doğrudan elenmiştir, ikinci aşamada yeni evrimsel bir kümeleme algorithması olan Genetic Kromodinamikler uygulanmıştır. Activiteyi oluşturan hareket desenlerinin her biri zaman uzamsal nitelik kümesi olarak ifade edilmiştir. Aynı aktivitenin farklı performansları arasındaki zamansal değişimlerin etkilerinin azaltılması amacıyla bir Saklı Markov Model yapısı önerilmiştir. Önerilen algoritma ile hem parameter hem de kodkitabı eğitim sırasında öğrenilmiştir.

Anahtar Kelimeler: 3B Gabor Filtreleri, Aktivite Tanıma, Nitelik Seçimi, Genetik Kromodinamik Öbekleme, Hareket Analizi, Profil Saklı Markov Model, Zaman-uzamsal nitelikler, Zamansal Desen.

To my dear nephew, Berk Bayram, for his smile.

ACKNOWLEDGMENTS

It is difficult to overstate my gratitude to my PhD supervisor Prof. Dr. Gözde Akar. This thesis would not have been possible unless her guidance, advice, encouragements and insight. Most of all, I would like to thank for her patience. Beginning from my undergraduate years she has been always a mentor to me, and helped me with the most critical steps of my life. It has been an honor to be her PhD student. I hope she accept me as a lifetime student. I owe her a lot and I am forever grateful.

I would like to give special thanks to Prof. Dr. Mustafa Karaman for his time he devoted to me through my undergraduate years, for triggering and nourishing criticism, for transmitting me his scientific intuition and passion: they have inspired and enriched my growth as a student, a researcher, and as a person. He was the reason why I decided to pursue a career in research.

I would like to record my gratitude to the members of my PhD thesis monitoring committee, Prof. Dr. Neşe Yalabık and Assist. Prof. Dr. İlkay Ulusoy, for their valuable criticism, constructive suggestions and comments.

I am deeply indebted to my friends, İbrahim Demir, Onur Çilingir, Derya and Oktay Algün for helping me to get through the difficult times, motivating me and for all the emotional support, entertainment, and caring they provided.

Last but of course not least, this work would not have been achieved without the support and understanding of my family. They have always supported me in everyway I have chosen in life. I want to thank my parents for their love, understanding and knowing my silent thoughts. I am grateful to my sisters, İlknur and Özgecan, for encouraging me at rough times and keeping me sane; they always showed lot of confidence in me. I appreciate their willingness to be there for me whenever I need.

TABLE OF CONTENTS

ABSTRACT iv
ÖZ v
ACKNOWLEDGMENTS vii
TABLE OF CONTENTS viii
LIST OF TABLES x
LIST OF FIGURES xi
LIST OF ABBREVIATIONS xiii

CHAPTERS

1. INTRODUCTION 1
    1.1. What is Gait? 3
    1.2. Problem Definition 3
    1.3. Our Approach 4
    1.4. Cotributions 5
    1.5. Organization of the Thesis 7
2. LITERATURE SURVEY 8
    2.1. Perceptual Activity Recognition 8
    2.2. Computational Activity Recognition 13
        2.2.1. Pre-processing 15
        2.2.2. Gait Analusis 19
        2.2.3. Activity Modeling and Recognition 28
    2.3. Discussion 31
3. BACKGROUND 34
    3.1. Background/Foregorund Segmentation 35
        3.1.1. Adaptive Gaussian Mixture Model 35
    3.2. Gabor Filters 37
    3.3. Feature Selection 39
        3.3.1. Genetic Cromodynamics 42
    3.4. Hidden Markov Models 42
        3.4.1. Standard Hidden Markov Models 43
        3.4.2. Profile Hidden Markov Models 47

4. PROPOSED HUMAN ACTIVITY RECOGNITION METHOD 51
    4.1. Pre-processing 52
    4.2. Spatio-Temporal Activity Representation 53
        4.2.1. 3D Gabor Filters 53
        4.2.2. Spatio-Temporal Feature Selection 58
    4.3. Human Activity Recognition with Profile HMM 67
        4.3.1. Adapting Profile HMM Procedure for Action Modeling 67
        4.3.2. Parameter Learning 69
        4.3.3. Sequence Scoring and Activity Recognition 73
5. RESULTS 75
    5.1. KTH Action Dataset 75
        5.1.1. Comparative Results 75
    5.2. Weizmann Action Dataset 83
        5.2.1. Comparative Results 84
        5.2.2. Robustness Results 85
    5.3. Discussion 85
6. CONCLUSIONS 88
PREFERENCES 91
CIRRICULUM VITAE 101

LIST OF TABLES

TABLES

Table 1 Notations ……..……………….…………………..…………………………...71

Table 2 Number of sequences in KTH dataset……….………..…………………………...76

Table 3 Comparison of our results with others on KHT dataset…………………………...82

Table 3 Comparison of our results with others on Weizmann dataset………..…………...84

LIST OF FIGURES

FIGURES

Figure 1: Overview of the proposed method 6
Figure 2: General framework of activity recognition 15
Figure 3: 2D Gabor kernels at 5 different scales and 8 different orientations 38
Figure 4: Standard profile HMM structure 46
Figure 5: Preprocessing of a frame a) original frame, b) foreground mask, c) region that will be processed in the later steps of the algorithm. 52
Figure 6: Slice of a 3D Gabor filter 56
Figure 7: Responses of a moving bar to 3D Gabor filters with different orientations. 56
Figure 8: Responses to a 3D Gabor filter frames from t=1 to 42....................................... 57
Figure 9: 3D Gabor responses a) original frames, b) Response samples of the middle frame of (a) to 3D Gabor filters of different spatio-temporal orientations.................................... 59
Figure 10: Exemplary thresholded responses of a frame (5 spatial frequencies, 8 spatial orientations) 60
Figure 11: Spatio-temporal feature locations of a frame after filtering 60
Figure 12: Genetic Chromodynamics algorithm.. 64
Figure 13: Overview of the training algorithm 65
Figure 14: Overview of the recognition algorithm 66
Figure 15: A profile HMM for repeat matches to subsections of the profile model............ 69
Figure 16: Viterbi Count Learning procedure.. 72
Figure 17: Proposed profile HMM recognition algorithm.. 74
Figure 18: Samples from KHT database.. 76
Figure 19: Confusion matrix on KHT database a) performance, b) number of frames 77
Figure 20: Confusion matrix on KHT database a) s-LDA [146], b) MMHCRF [134]........ 78
Figure 21: Confusion matrix on KHT database a) Mid level motion features [115], b) Local trinary patterns [126]... 79
Figure 22: Confusion matrix on KHT database a) Action MACH [133], b) Feature tracking [121].. 80
Figure 23: Confusion matrix on KHT database a) Boosted EigenActions [117], b) Extreme learning machine [58] 81

Figure 24: Weizmann dataset samples ... 83
Figure 25: Samples from the robustness test set of the Weizmann dataset ... 85

LIST OF ABBREVIATIONS

ABBREVIATION	EXPLANATION
GC	Genetic Cromodynamics
DTW	Dynamic Time Wrapping
HMM	Hidden Markov Model
HSV	Human Visual System
kNN	K Nearest Neighbor
MACH	Maximum Average Correlation Height Filter
MMHCRF	Max-Margin Hidden Conditional Random Fileds
PCA	Principle Component Analysis
pHMM	Profile Hidden Markov Model
PL	Point Lights
s-LDA	Semi-Latent Dirichlet Allocation
SVM	Support Vector Machine
VC	Viterbi Counting
GPU	Graphical Processor Unit

CHAPTER 1

INTRODUCTION

As the number of cameras grows exponentially, it becomes impossible for a security operator to monitor several video streams for extended periods of time. This leads to the need for the systems that are able to detect, categorize and recognize human actions, requesting human attention only when necessary. Detection and tracking systems have progressed significantly in the past few years, and human motion and behavior interpretation have naturally become the following step. In all of the surveillance systems, tracking is first step and behavior recognition is the final goal [44, 100].

Study of gait is a relatively new area for computer vision research. Gait analysis is inspired by the human ability to recognize other people from their movements, especially from their walking patterns, even at very coarse resolutions [3, 39, 40, 112].

Activity recognition can be described as the analysis and recognition of human motion (gait) patterns. The aim of the analysis is to represent the human motion effectively for recognition. The main challenges of analyzing and comparing activities are the temporal and spatial variations of the human motion. Temporal variation is caused by the difference in duration of performing actions and spatial variation is due to the fact that people have different ways of performing activities and different physics [2].

Approaches to human activity recognition can be classified as top-down methods based on geometric body reconstruction [62-76], and bottom-up methods based on low level features such as image differences, edges, corners or local motion [100-146].

Most systems based on the top-down approach employ a geometrical model of the human body; model parameters are determined from images. One approach is to placing markers on the moving body; however, this approach is generally impractical and moreover impossible for some applications [2, 76]. Marker free approaches mainly determine body/part orientation by tracking the edges or uniform color regions [62, 74, 70]. However, those methods suffer from the inaccuracy of the body model. More sophisticated model based methods, having higher computational complexity, matches 2D image sequences to a model to automatically find feature correspondences [63, 64, 66, 67]. Reconstructing the human body shape yields rich and useful representation such as joint angle parameters if the reconstruction is successful, then pattern classification techniques can be used to recognize a sequence of model parameters extracted from image sequences. However, the reconstruction procedures are neither robust nor reliable for real images. Since the purpose is to recognize human actions from time sequential images, not obtaining geometric representation of human bodies, geometric reconstruction is not essential.

Bottom-up approaches which utilize low level features extracted from image sequences have been the subject of various studies [102, 103, 104, 105]. Those methods are inspired by approaches to the image based object recognition that relies on spatial features; however, video and images have distinct properties. In video analysis, compared to the spatial features, local spatio-temporal features capture both characteristic shape and motion in video and provide relatively robust representation of events [97].

The main challenge about the feature based approaches is that the relations between low level features and the high level action description category are not explicit.

The problem of describing action categories with low level features can be eliminated with a learning procedure. The goal is to classify observed spatio-temporal features into human activity categories. Majority of the methods based on the spatio-temporal features follow a similar trajectory: detect local interest points, compute a representation of a window of pixels around each of these points using a descriptor. A commonly used approach is to quantize the local space-time features to model each sequence as a spatio-temporal “bag of features” [130, 125, 153, 136], and then feed to a classifier. Most of the effort has been made in designing space-time feature detectors [138, 3, 4, 102, 124] and descriptors [5,125], rather than modeling the spatial and temporal relations between them.

In most of the methods based on spatio-temporal features, activity video is modeled as a whole (with histograms of visual words, word co-occurrence matrix, etc.) [81, 87, 114, 116, 119, 122, 123, 127, 128, 131, 133, 137, 139]. Although some recent works proved that temporal relations between spatio-temporal features are also valuable information for activity recognition, instead of modeling temporal order [147], those works mainly focused on extracting additional features based on temporal relations [126]. Other methods based on state-space representation use fixed number of states by clustering or sampling, or introduce end point constrains to align different length sequences [112, 120, 125, 126, 130, 132, 134, 135, 140, 142, 143, 145].

In this work an activity is considered as a time evolving structure composed of temporal texture primitives. A novel state-space representation of the activity is achieved by 3D Gabor filtering and a two step feature selection algorithm. Proposed feature selection algorithm both eliminates redundancy in space and time, and segments the video sequence in time into temporal texture primitives of the underlying activity, number of those segments are not predetermined. Activities are modeled and recognized by a Profile Hidden Markov Model structure, which provides a flexible tool to learn and compare unaligned activity sequences with different lengths. Codebook of each activity class is also learned incrementally though the Profile HMM training.

1.1. What is Human Gait?

Since the gait analysis studies are covered by a range of disciplines (medical, mechanical, computer vision etc.) for a various applications (training, treatment, simulation, human identification, activity recognition, etc.), gait has some different definitions from different aspects, and those definitions are generally based on the analyzed parameters. Therefore it is found to be required to state the definition of the gait on which this work is built up.

Human Gait is the way locomotion is achieved using human limbs. In other words, Gait is the way in which human move their body from one point to another. Most often, this is done by walking, although it may also be run, skip, hop etc.

1.2. Problem Definition

A general statement of the problem can be formulated as follows: automatically recognizing a set of human activities such as walking, running, hand waving, using gait analysis, in a video stream taken from a natural scene by a fixed camera, invariant of the human subject.

General activity recognition, a task that is done by humans in daily activities, comes from virtually uncontrolled environment. Systems which automatically recognizes actions from uncontrolled environment, must detect moving humans in the video. Moving person segmentation from the scene is considered as a preprocessing step of activity recognition studies and mostly considered as a different problem with respect to activity recognition in the literature. Some recent works based on spatio temporal features claim moving subject segmentation is not required; however those methods also considers while the process of action modeling there is only the probe motion in the scene.

Human activity recognition is a difficult problem due to the variation in the way people perform actions: temporal and spatial variation. Temporal variation is caused by the difference in duration of performing actions and it is usually effectively compensated by using Hidden Markov Models or Dynamic Time Warping. Spatial variation in the other hand is due to the fact that people have different ways of performing activities and different physics, and more difficult to compensate. It must be noted that even the same person never performs an action exactly the same again. On the other hand recognition of activities from an uncontrolled environment is a very complex task; lighting conditions may vary tremendously, activities may appear at different orientations and subject can be occluded.

In this thesis, moving subject is segmented with a preprocessing step. We aim to provide the correct label associated with the activity in case of occlusions and illumination changes.

1.3. Our Approach

Our approach is based on modeling the activity as a pattern composed of time sequential temporal texture primitives. Therefore our first objective is to analyze texture of the motion. This is achieved by 3D Gabor filtering followed by a two step feature selection algorithm to eliminate redundancy and extract relevant features. 3D Gabor filters are inspired from the human visual system [93, 98, 147]; they give high responses at spatio-temporal locations with high variation at different scales and orientation.

The proposed feature selection algorithm locates the temporal texture primitives in both spatial and temporal domains, and represents them with local spatio-temporal features. In the first step, local maximums of 3D Gabor kernel responses of each frame are detected as the interest points to eliminate spatial redundancy. Then a feature vector is extracted at each interest point as a composition of 3D Gabor kernel responses at that point. In the second step, Genetic Chromodynamics, a recent evolutionary unsupervised clustering algorithm, is adopted to cluster local spatio-temporal features in time to extract relevant features of time ordered temporal texture primitives. Elimination of temporal redundancy and irrelevant spatio-temporal features are continued throughout the clustering stage.

To model and recognize the activities, an HMM approach, which is one of the most successful algorithms on sequential data modeling [2], is proposed. However in this work the number of the local spatial-temporal features is not predetermined, and since they are found through a fully automatic feature selection method, can be alter for different sequences. This is a handicap to apply the standard HMM approach [70, 106]. On the other hand another well known sequential data modeling method, Dynamic Time Wrapping (DTW) can handle such differences between sequences by operations of deletion-insertion, in exchange for the consideration of inter actions between nearby subsequences occurring in time [69, 154]. However DTW requires the beginning and the end of each sequence is rigidly fixed, which makes it inapplicable for our problem since labeling the end points of the activity from a given sequence is not an easy task. Inspired from the success of DTW handling different length sequences we proposed a Profile HMM structure which also has inserting and deleting states to align different length sequences, while still considering correlation between adjacent temporal instances. Proposed structure also enables to recognize sub-sequences and does not require end point constrains. Moreover applying Viterbi Path Counting algorithm for training is lead to the incremental learning of the underlying common pattern of each activity class as well as its codebook. Overview of the proposed method is given in figure 1.

1.4. Contributions

The following are our main contributions:

1. A novel spatio-temporal feature detector and descriptor is proposed: spatio-temporal texture representation is achieved through 3D Gabor filters.
2. A novel state-space based representation is proposed: activity is segmented into time ordered temporal texture primitives by a two step unsupervised feature selection algorithm, and each segment is represented by a set of spatio-temporal features. The number of temporal texture primitives involved in a sequence is not predetermined.
3. A novel action modeling and recognition method based on a Profile Hidden Markov Model with Viterbi Path Counting algorithm is proposed: at the training stage of the Profile HMM, the common pattern of each activity class is learned from different length sample sequences and an HMM structure is constructed accordingly. Codebook is also learned incrementally for each activity class, at the training stage. Proposed Profile HMM structure does not require end point constrains and can handle and missing data by accepting sub-sequences.

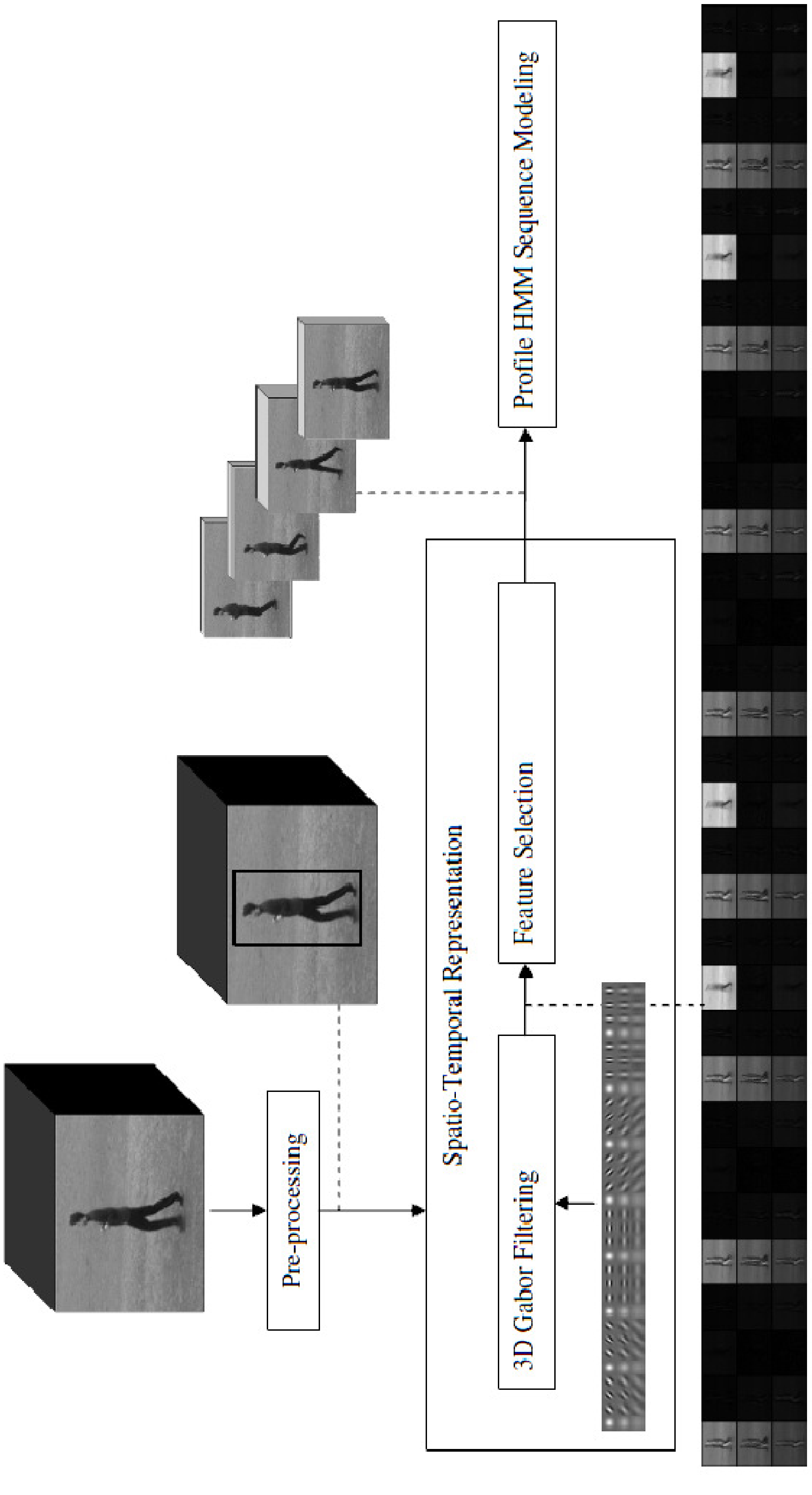


Figure 1: Overview of the proposed method

1.5. Organization of the thesis

Understanding human gait is the starting point of understanding human activities. Studies on the analysis of the human gait were started by the psychophysics community, and then various computational methods were proposed as the result of medical, human recognition and finally activity recognition studies on human gait. In the next chapter first we summarize the literature on both human and machine analysis of human gait, and we survey some main approaches on human activity recognition.

In chapter 3, background about the algorithms used in the proposed method is given.

Chapter 4 introduces the proposed approach based on spatio-temporal representation of human actions with 3D Gabor filters and profile HMM learning. Our main contributions on feature selection and action learning are given. The algorithm is explained in detail.

Performance of our method is examined on two different publicly available human action datasets. Simulation results and their comparisons to well known action recognition methods are presented in Chapter 5.

In Chapter 6, concluding remarks are stated. Future works, which may follow this study, are also presented.

CHAPTER 2

LITERATURE SURVEY

Studies on the ability of humans to recognize others style of walking were started by the psychophysics community. Medical studies on gait were arising by a need for an improved understanding of locomotion for the treatment of World War II veterans. Inspired from the medical studies computer vision society focused on the people recognition problem using gait. With the increasing number of surveillance systems need for more intelligent systems emerged, which results in the recent studies of activity recognition by human gait analysis.

2.1. Perceptual Activity Recognition

Human observers exhibit an impressive level of visual sensitivity to human movement. Johansson [3, 4] created displays of human movement by attaching point lights to the major joints of human models. The experiments of Johansson then extended by other researchers to analyze different aspects of perceptual activity recognition [159].

In [3] the models were filmed so that only these point lights were visible to observers. Although observers rarely perceived a human form when these displays were static, when the displays were set in motion, observers were rapidly able to detect and identify various human movements. Moreover, with stimulus duration of only 200 ms, observers could perceptually organize and accurately identify the particular action, such as walking or running.

As the result of the early studies it is stated that although a person’s affective state can be discerned from static pictures, motion provides even more reliable and compelling information.

Observers can easily identifying what an actor is doing in a given PL display [5, 6], even when the number of possible activities is quite large. People can also easily discern activities (e.g., dancing) involving two or more individuals [7], and they can judge the emotional implication of an action when viewing PL animations of the whole body [8, 9, 10] or even the movements of individual limbs [11].

Sensitivity to human motion increases with the number of illuminated joints as well as with the exposure duration of the animation. But even under potentially ambiguous conditions, perception of human motion is remarkably robust. Observers can recognize human activity when a PL animation is presented for less than one-tenth of a second [3], when the dots are blurred or randomized in contrast polarity over time [12], or when stereoscopic depths of the dots marking the joint positions of a PL walker are altered such that the 3D locations of the dots are unrelated to their implied depth orderings for the human figure [13].

Observers can easily recognize a walking person when the PL animation is embedded in an array of dynamic noise dots that far outnumber the dozen or so dots defining the person [14]. At least for PL walkers, the points defining the wrists and ankles are crucially important when judging the direction of walking [12], while the points defining the mid-limb joints (elbow and knees) and the torso (shoulder and hips) contribute significantly to detection of PL walkers embedded within noise [15]. Although human action can be recognized most easily when the PL tokens are placed on the joints of the body, observers can still detect PL human figures when tokens are placed on positions other than the joints, such as intermediate positions on the limbs [16].

Perception of human motion is seriously disrupted by perturbations in the temporal relations of the PL tokens [17]. Introducing spatiotemporal jitter into the phase relations of the moving dots disturbs the quality of human motion [19]. Moreover, PL animations depicting abnormally slow walking movements produce perception of rotation in depth about a vertical axis, not perception of slow human gait [18].

One of the important characteristics of human motion perception is its vulnerability to inversion: Human action is difficult to perceive in inverted PL animations. In this respect, bodily motion perception resembles face perception, which is also highly susceptible to inversion [20]. This orientation dependence operates in egocentric, not environmental, coordinates: PL animations shown upright with respect to gravity are not difficult to perceive when the observer’s head is turned so that the retinal image of those animations is no longer upright with respect to head position [21]. Prior knowledge cannot counteract this inversion effect: Informing observers ahead of time that they’ll be seeing upside-down people does not help them identify what they’ve seen, which implies that they cannot mentally rotate the images [22]. With practice, observers can learn to detect inverted human motion [23], but in so doing observers are relying on detection of conspicuous clusters of dots, not on global impression of a human figure.

Visual sensitivity to human motion is also affected when a PL figure is imaged in the visual periphery, and this impairment is not simply attributable to the periphery's reduced visual resolution; increasing the sizes of the PL dots and the overall size of the human figure cannot compensate for this loss in sensitivity [14].

Perception of human motion is also impaired when PL animations are viewed under dim light conditions [19]. Observers can use kinematics information to infer properties of objects with which PL actors are interacting. For example, people can accurately estimate the weight of a lifted object from observing the lifting motion alone [24], and they can judge the elasticity of a support surface by watching a PL person walking on that surface [25]. There is disagreement whether observers are directly perceiving kinetic object properties from human kinematics information or, instead, are deploying heuristics to infer object properties from kinematics. In either event, however, there is no doubt that kinematics can accurately specify the act of lifting and, moreover, the effort required to do so [26].

Accurate perception of PL figures is not limited to human activity. Mather *et al.* [27] demonstrated that people could identify animals, such as a camel, goat, baboon, horse, and elephant, whose movements were represented by PL animations. People found this task impossible, however, when viewing a single, static frame from the PL sequence.

The ability to perceive PL depictions of human motion arises early in life. Infants four months old will stare at human motion sequences for longer durations than they will at the same number of dots undergoing random motions, a preference not exhibited when infants view an inverted PL person [17]. Using behavioral testing, Pavlova *et al.* [28] have shown that young children between the ages of three and five steadily improve in their ability to identify human and nonhuman forms portrayed by PL animations, with adult levels of performance achieved by age five. At the other end, observers older than 60 years are quite good at discriminating among various forms of human motion even when the PL sequences are brief in duration or the dots are partially occluded [29]. Moreover the ability to perceive human motion stands in contrast to age-related deficits in speed discrimination [30], coherent motion detection, detection of low-contrast moving contours, and perception of self-motion from optic flow.

The decreased sensitivity to inverted displays of human movement suggests that low-level visual mechanisms may not be sufficient to account for action perception. There is debate in the literature about the involvement of top-down influences in perception of human motion, where top down means conceptually driven processing.

As with so many of these kinds of debates, the emerging resolution entails a synthesis of bottom-up and top-down determinants [31].

Evidence for the role of low-level visual processes in the perception of human movement comes from several lines of research. Johansson [3] thought that perceptual process underlying human motion sequences involved vector analysis of the component body parts, with those vectors then incorporated into a single structured percept.

To determine whether low-level motion analyses underlie the perception of human movement, [12] inserted blank intervals between successive frames of a PL animation, reasoning that low-level motion analysis is restricted to very brief interstimulus intervals (ISIs), whereas high-level motion favors longer ISIs. On direction discrimination tasks involving a PL walker, Mather *et al.* [16] found that perception of human gait was best at the shortest ISIs and deteriorated with longer ISIs. Mather *et al.* concluded that the perception of human motion relies on signals arising within low-level visual mechanisms whose response properties are constrained to operate over short spatial and temporal intervals.

Subsequent work showed, however, that higher-level visual processes also support perception of human motion. Thornton *et al.* [32] found that human gait can be perceived with PL animations over a range of temporal display rates that exceed the value typically associated with low-level, local motion analysis.

Thornton *et al.* [32] found that when a PL sequence is embedded among dynamic noise dots, attention is crucial for perceiving the human figure at long ISIs but not at short ISIs. The same pattern of results was found when Thornton *et al.* used a very brief ISI but varied the type of masking noise. To perceive a human figure in a mask of scrambled dots (i.e., dot motions with vectors identical to the PL dots), observers had to attend to the animation. Conversely, in random noise mask (i.e., dots randomly replaced each frame), distracted attention had no effect on walker detection.

Thus, both bottom-up and top-down processes are employed during the perceptual analysis of PL animations of human motion. Human motion itself can also exert a top-down influence on other aspects of perception. For example, Watson *et al.* [33] found that while viewing PL walkers differing in color and in heading direction, observers saw one PL walker or the other over time. Based on this and related results, Watson *et al.* [160] concluded that dominance during rivalry resulted from the integration of high-level perceptual organization (responsible for perception of human motion) with lower-level inhibition between cortical representations of input from the two eyes. Human motion can also influence the perceived direction of translational motion. Thus, when a coherent PL person walks in front of a counter phase flickering grating with no net directional energy, the grating appears to translate in the direction opposite the walker's heading, just as the physical environment flows past us when we walk [34]. In a similar study, the global motion engendered by a PL walker provides an effective reference frame for judging whether or not local dot motions are coherent [35]. All of these studies imply that human motion exerts a significant influence on putatively low-level motion processing.

PL animations of human activity seemingly contain little form information about the human body, yet people can easily detect PL figures appearing within a cloud of moving "noise" dots whose local motions are identical to those defining the PL figures. Based on motion alone, detection of a PL figure should be extraordinarily difficult under these conditions. Evidently, the visual analysis of human motion is constrained by the hierarchical structure of the human body. When PL animations contain motion vectors that violate the hierarchical structure of the human body, observers experience great difficulty detecting the presence of a PL body part [15]. Several lines of evidence underscore the importance of bodily form. Sequential presentation of two static pictures of a person performing some action is sufficient for the perception of human action, even though such displays contain minimal motion information. On the other hand, changing positions of the moving points in a PL animation convey sufficient form cues for the detection of a PL walker. To dissociate position and motion, Beintema & Lappe [36] designed a variant of the point-light animation in which the positions of the dots are not confined to the joint but instead can appear anywhere along the limbs. Moreover, the dots change their positions along the limbs unpredictably from frame to frame. While these two manipulations should not disrupt specification of body shape, they make it virtually impossible to perceive coherent motion of the dots defining an activity. Nonetheless, observers viewing these displays can judge with reasonable accuracy the direction (left versus right) in which a PL figure is walking. In fact, detection performance measured using these special PL animations is well predicted by the total number of points seen in a trial, irrespective of the distribution of these points over time [18].

Additional evidence for the importance of bodily form is from a work by Hiris *et al.* [23]. They created "arbitrary" motion sequences by relocating the dots from a PL walker. Thus, for example, a wrist dot might be placed at the location of the shoulder dot and the shoulder dot relocated to the position of the knee, and so on. The resulting arbitrary figure comprised the same dot motions as the walker in the absence of a human form. Following the design of studies with masked PL walkers, the arbitrary figures were presented within a mask of dynamic dots. With practice, observers learned to discriminate whether or not the arbitrary figure was presented in a mask, with performance eventually approaching that achieved with an ordinary PL walker. However, observers described performing this task by looking for a characteristic cluster of dots at a given location, a strategy very dissimilar from that used with upright PL walkers in noise. Moreover, inverting the arbitrary figure had no effect on detection performance. When a display change forced observers to detect the global pattern of motion in the arbitrary figures, performance hovered near chance levels regardless of practice.

The human form is important for the perception of human motion, especially when it is impossible to rely on local motion regularities. Such results further suggest that different processes are employed during the analyses of human motion and object motion [37]. Taken together, current research indicates that both form and motion play critical roles in the perception of human action.

2.2. Computational Activity Recognition

Activity recognition is a relatively new area to computer vision researchers. In the general framework of the activity recognition systems motion detection and tracking is the first step. Once the subject of the motion is segmented the next step is the gait analysis to achieve a good representation to differentiate different activities. Final step is the activity modeling and recognition with the gait analysis results (figure 2).

Since there is a considerable background on motion detection and tracking algorithms in the literature, leaving it as a preprocessing step, current activity recognition studies mostly focused on gait analysis.

Early studies, inspired from the medical gait analysis with markers, proved that joint angles are sufficient for gait recognition [2]. However reliably recovering joint angles from a monocular video without markers is a hard problem.

One approach to markerless gait analysis is body part tracking. Body-part tracking based gait analysis can be broadly classified as being either model based or model free. Various approaches for tracking the whole body have been proposed in the literature using a variety of 2D and 3D shape models and image models [2, 76]. Other approaches determine body/part orientation by tracking the edges or the same color regions [62, 70, 74]. Both methodologies follow the general framework of feature extraction, feature correspondence and high-level processing. The major difference is with regard to feature correspondence between two consecutive frames. Methods based on a priori models match the 2D image sequences to the model data. Feature correspondence is automatically achieved once matching between the images and the model data is established. Model free methods establish correspondence between successive frames based upon the prediction or estimation of features related to position, velocity, shape, texture and color; they assume some implicit notion of what is being observed.

Although model free approaches reduce the computational complexity of the model based approaches, body part tracking still suffers from the low resolution of the surveillance video.

On the other hand, recent gait analysis studies for activity recognition are shifted to holistic approaches, mainly because of their ease of application and simplicity. Holistic approaches does not focused on individual body parts, instead they analyze the whole frame to find some low level features, such as corners, edges, temporal differences, binary silhouettes and optical flow [97]. Spatio-temporal feature detectors and descriptors are the main building blocks of the holistic approaches.

Many different spatio-temporal feature detectors and descriptors have been proposed in the past few years [100-105]. Feature detectors usually select spatio-temporal locations in video by maximizing specific saliency functions. Feature descriptors capture shape and motion in the neighborhoods of selected points using image measurements such as spatial or spatio-temporal image gradients and optical flow. A spatio-temporal feature is a short, local video sequence such as a knee bending.

Regardless of the gait analysis approach there are some common tools in the literature used to model and recognize the activity. The most common tools employed by the methods rely on sequential representation, are Hidden Markov Model and Dynamic Time Warping [69, 70, 78]. Principle Component Analysis and Support Vector Machines are the common tools used by the methods rely on time holistic representation [116, 117, 137].

In the next section a brief summary of the approaches to segment motion from the rest of the video is given. In section 2.2.2 gait analysis methods are reviewed under two main categories; body-part based approaches and holistic approaches, and for the sake of completeness activity classification algorithms used by those methods are also mentioned. In section 2.2.3 activity modeling and recognition algorithms are reviewed with the emphasis on sequential representation.

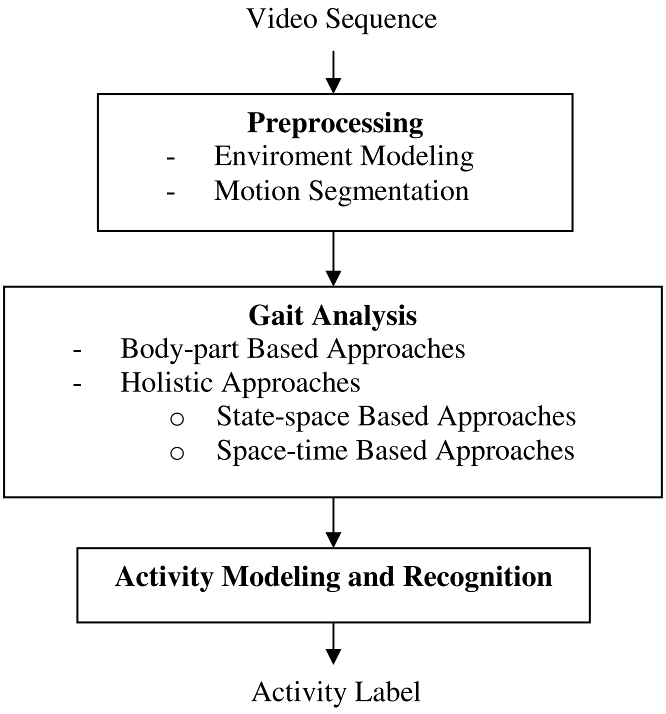

Video Sequence
Preprocessing
- Enviroment Modeling
- Motion Segmentation
Gait Analysis
- Body-part Based Approaches
- Holistic Approaches
o State-space Based Approaches
o Space-time Based Approaches
Activity Modeling and Recognition
Activity Label

Figure 2: General framework of activity recognition

2.2.1. Preprocessing

Nearly every activity analysis system starts with motion detection. The aim of the motion detection is to segment regions corresponding to the moving objects from the rest of the image. Subsequent processes such as tracking and behavior recognition are dependent on it. Motion detection generally involves environment modeling and motion segmentation, which intersect each other during processing.

2.2.1.1. Environment Modeling

Environment models can be classified into two: 2D models in the image plane and 3D models in the world coordinates. Due to their simplicity 2D models have more applications.

For fixed cameras the key problem is to automatically recover and update background images from a dynamic sequence. Illumination variance, shadows and shaking branches, bring many difficulties to the acquirement and updating of background images.

The roots of background segmentation go back to 19th century. It was shown that the background image could be obtained simply by exposing a film for a period of time much longer than the time required for the moving objects to traverse the field of view. Thus in its simplest form, the background image is the long-term average image:

$$B(x,y,t)=\frac{1}{t}\sum_{t'=1}^{t}I(x,y,t') \tag{2.1}$$

$I(x,y,t)$ is the instantaneous pixel value for the *(x,y)* pixel at time *t*. This can also be computed incrementally:

$$B(x,y,t)=\frac{t-1}{t}B(x,y,t-1)+\frac{1}{t}I(x,y,t) \tag{2.2}$$

One obvious problem of this approach is that lighting conditions change over time. This is handled using a moving window average or more efficiently using exponential forgetting. Each image‘s contribution to the background image is weighted so as to decrease exponentially as it recedes into the past.

$$B(x,y,t)=(1-\alpha)B(x,y,t-1)+\alpha I(x,y,t) \tag{2.3}$$

$1/\alpha$ is the time constraint of the forgetting process.

Another serious problem arises when objects are slow moving or temporary stationary. The solution proposed by Koller *et al.* [38] updates the background image with only those pixels not identified as moving objects. This approach becomes effective but still suffers from very slow motion. Adaptive background updating approach is studied at [41]. Presented method uses Gaussian distribution for background generation. Instead of using all the pixels in the image actively, they divide the pixels into active and inactive ones. Gaussian distributions are used to model the history of active pixels and to state whether they belong to background or foreground. According to the classification of the previous active pixel, the inactive pixels are also classified as a part of the background or foreground. They also reduce the frame frequency and use only every n^{th} frame in the image sequence to construct the adaptive background.

In [42] the proposed method is robust to noise and the intensity change of the image, which can be affected by the illumination or the function of a camera. The method is based on the assumption that pixels of the same intensity in the original background image will keep the same intensity even with a change in intensity of the whole image.

Toyama *et al.* [43] propose the Wallflower algorithm in which background maintenance and background subtraction are carried out at three levels: the pixel level, the region level, and the frame level. Haritaoglu *et al.* [44] build a statistical model by representing each pixel with three values: its minimum and maximum intensity values, and the maximum intensity difference between consecutive frames observed during the training period. These three values are updated periodically. McKenna *et al.* [45] use an adaptive background model with color and gradient information to reduce the influences of shadows and unreliable color cues.

Regarding 3-D environmental models [46, 47], current work is mostly limited to indoor scenes because of the difficulty of 3-D reconstructions of outdoor scenes.

2.2.1.2. Motion Segmentation

Motion segmentation in image sequences aims to detect regions corresponding to moving objects. Detecting moving regions provides a focus of attention for later processes such as tracking and behavior analysis because only these regions are required to be considered in the later processes [48].

2.2.1.2.1. Background subtraction

Background subtraction is a popular method for motion segmentation, especially under those situations with a relatively static background.

It detects moving regions in an image by taking the difference between the current image and the reference background image in a pixel-by-pixel fashion. It is simple, but extremely sensitive to changes in dynamic scenes derived from lighting and extraneous events etc. Therefore, it is highly dependent on a good background model to reduce the influence of these changes.

In [48] moving object detection is handled by the use of an adaptive background subtraction scheme which reliably works both in indoor and outdoor environments.

2.2.1.2.2. Temporal differencing

Temporal differencing makes use of the pixel-wise differences between two or three consecutive frames in an image sequence to extract moving regions. Temporal differencing is very adaptive to dynamic environments, but generally does a poor job of extracting all the relevant pixels. In [49] after the absolute difference between the current and the previous frame is obtained, a threshold function is used to determine changes. By using a connected component analysis, the extracted moving sections are clustered into motion regions. An improved version uses three-frame instead of two-frame differencing. Temporal differencing is carried on edges detected from the observed image sequence to find object boundaries.

In [50], motion is segmented in an entire sequence. Rather than segmenting the image plane into a set of disjoint regions, they propose to segment the spatio-temporal volume into a set of disjoint phases of homogeneous motion. Compared to the iteration of the two-frame model, this introduces an additional temporal regularity of the estimated motion boundaries.

2.2.1.2.3. Optical flow

Optical-flow-based motion segmentation uses characteristics of flow vectors of moving objects over time to detect moving regions in an image sequence. [51, 52] computes the displacement vector field to initialize a contour based tracking algorithm, called active rays, for the extraction of articulated objects. The results are used for gait analysis.

Optical-flow-based methods can be used to detect independently moving objects even in the presence of camera motion. However, most flow computation methods are computationally complex and very sensitive to noise, and cannot be applied to video streams in real time without specialized hardware [53].

Besides the basic methods described above, there are some other approaches for motion segmentation. Using the extended expectation maximization (EM) algorithm, Friedman *et al.* [54] implement a mixed Gaussian classification model for each pixel. This model classifies the pixel values into three separate predetermined distributions corresponding to background.

In [55] the key of the motion segmentation is to find the same moving region in each frame, and the projection transforms mapping this region between frames. They used a layered approach for motion segmentation; the goal is finding 2D layers corresponding to different non-rigid motion of the human body parts. Their method is to model the apparent motions in video based on optical flow and random sample consensus (RANSAC). They are using affine and homography models to describe motions of body parts. They combine tracking with motion segmentation, in order to obtain better motion features.

2.2.2. Gait Analysis

Here we summarize the gait analysis and motion representation approaches in the literature by diving them into two main categories according to their activity representation strategies; body-part based approaches and holistic approaches.

2.2.2.1. Body-Part Based Approaches

Body-part based approaches generally track human body parts from one frame to another in a video stream. These methods represents underlying action by computing some properties of moving body parts such as joint trajectory, body part orientation, joint angles, velocities. The tracking algorithms usually have considerable intersection with motion detection during processing. Tracking over time typically involves matching objects in consecutive frames using features such as points, lines or blobs. Useful mathematical tools for tracking include the Kalman filter, the Condensation algorithm, the dynamic Bayesian network, etc.

Some approaches followed the methods detecting silhouettes and body parts, while others used a motion capture system to circumvent the problems in view-invariant action recognition [76].

Niyogi *et al.* [62] use the spatio-temporal pattern in XYT space to track, analyze and recognize walking figures. They examine the characteristic pattern produced by the lower limbs of a walking human, the projections of head movements are then located in the spatio-temporal domain, followed by the identification of the joint trajectories; The contour of a walking figure is outlined by utilizing these joint trajectories, and a more accurate gait analysis is carried out using the outlined 2-D contour for the recognition of the specific human.

Motion models of human limbs and joints are widely used in tracking. They are effective because the movements of the limbs are strongly constrained. These motion models serve as prior knowledge to predict motion parameters [64, 64], to interpret and recognize human behaviors [65], or to constrain the estimation of low-level image measurements [66].

Bregler [65] decomposes a human behavior into multiple abstractions, and represents the high-level abstraction by HMMs built from phases of simple movements. This representation is used for both tracking and recognition.

Zhao *et al.* [63] learn a highly structured motion model for ballet dancing under the minimum description length (MDL) paradigm. This motion model is similar to a finite-state machine (FSM). Ong *et al.* [66] employ the hierarchical PCA to learn their motion model which is based on the matrices of transition probabilities between different subspaces in a global eigenspace and the matrix of transition probabilities between global eigenspaces. Ning *et al.* [68] learn a motion model from semi-automatically acquired training examples and represent it using Gaussian distributions.

Wang *et al.* [74] computed a mean contour to represent the static silhouette information. Fourteen rigid body parts are used to construct a dynamic model; each part is represented by a truncated cone. Particle filter and K-nearest neighbor classifier are then applied for pose estimation and action classification. Davis and Taylor [124] proposed a human action recognition system to distinguish walking from nonwalking. Several body parts are detected from the silhouettes and four motion properties are then extracted based on feet locations.

Ren and Xu [70] proposed a similar approach to compute a binary silhouette and detect the head, torso, hands and elbow angles. Then a coupled Hidden Markov Model is used to recognize predefined actions.

Some researchers recently performed human activity recognition based on manifold learning. Wang and Suter [72, 73] adopted Locality Preserving Projections (LPP) to achieve the low dimensional embedding of silhouette data. The spatio-temporal property and geometric structure in the low dimensional space are analyzed. Three methods namely

Hausdorff nearest manifold, Gaussian mixture model and Hidden Markov Model are then proposed to recognize motions in the lower dimensional feature space.

Souvenir and Babbs [75] proposed a framework for learning a viewpoint-invariant manifold of primitive actions. The manifold provides a compact representation of human actions, action recognition and viewpoint estimation are then simultaneously performed.

Jia and Yeung [71] proposed a local spatio-temporal discriminant embedding method for human action recognition. This method is able to find an optimal embedding in both local temporal and local spatial domains.

In [69] Duygulu *et al.* proposed a “bag-of-rectangles” method for representing and recognizing human actions in videos, where each human pose in an action sequence is represented by oriented rectangular patches extracted over the whole body. Spatial oriented histograms are formed to represent the distribution of these rectangular patches. To recognize an action four different methods are proposed; frame by frame voting, global histogramming, a classifier based approach using SVMs, and Dynamic Time Warping on the temporal representation of the descriptor. Above all it is reported that the DTW methods achieved the best success.

2.2.2.2. Holistic Approaches

Holistic approaches were introduced to recognize human actions, with less computational complexity. These approaches does not focused on individual body parts, instead they analyze the whole frame to find some low level features, such as corners, edges, temporal differences, binary silhouettes and optical flow [97].

Referring to the success of interest point detectors in object recognition from images, several spatio-temporal interest point (region and/or volume) detectors have been proposed to encode the contents of videos compactly. Laptev *et al.* [128] proposed a 3D interest point detector to detect local corner-like structures in the space-time dimensions by adding a temporal constraint to the Harris-Laplace 2D interest point detector. Oikonomopoulos *et al.* [113] proposed an approach to detect spatiotemporal regions with high entropy, which extended the Kadir and Brady saliency detector [107]. Recently, Willems *et al.* [108] proposed a detector based on the Hessian matrix rather than a spatiotemporal second-moment matrix so that the interest points detected can be scale-invariant and densely cover the video contents. Instead of using interest point detectors, the dense sampling method can also be adopted to generate features in the spatio-temporal volume, as in [109].

Dollar *et al.* [139] proposed a detector to locate periodic frequency components in videos by applying a 2D Gaussian smoothing kernel along the spatial dimensions and a quadrature pair of 1D Gabor filters along the time dimension.

Regarding descriptors based on spatio-temporal information of videos, there have been several approaches. Efros *et al.* [111] proposed a motion descriptor based on the optical flow in different frames. Ke *et al.* [109] proposed 3D volumetric features calculated on *x* and *y* optical flow channels. Scovanner *et al.* [110] applied sub-histograms to encode local temporal and spatial information to generate a 3D version of SIFT. Recently, Klaeser *et al.* [104] proposed a new descriptor based on the histogram of oriented 3D spatio-temporal gradients.

Holistic methods can be divided into the space-time based approaches and the state space based approaches, according to their activity modeling strategies with spatio temporal features.

2.2.2.2.1. Space-Time Based Approaches

In [147] Chomat *et al.* proposed a probabilistic recognition method by using Gabor filters. A multi-dimensional histogram is computed from the outputs of the filter bank at each pixel *(x,y,t)* as a feature vector. Then probability of each action-feature vector is computed by Bayes rule. Final decision is according to the spatial average probability over a given frame. Although the authors report effective representation of the motion with Gabor filter responses, their method was suffering from redundancy, which results in inefficient representation of local features, and lack of a reliable recognition scheme.

Laptev [127] proposed a space–time interest point detector, which locates local salient pixels in space–time volume with significant local variations in both spatial and temporal domain. The local saliency maxima are detected based on the Harris operator. However, this method detects a small number of stable interest points which may not be sufficient to characterize complex events.

Niebles *et al.* [137] represented a video sequence as a collection of spatio-temporal words by extracting space–time interest points. By using probabilistic latent semantic analysis and latent dirichlet allocation, distributions of spatio-temporal words and intermediate topics are learned corresponding to human action categories in an unsupervised manner. The algorithm is able to localize multiple actions in complex motion sequences.

Uemura *et al.* [121] proposed a local feature tracking based method. Multiple interest point detectors (MSER, FAST, Harris Laplace, Hessian Laplace) are used to provide large number of features for every frame. The motion vectors for the features are estimated using optical flow and SIFT based matching. Motion compensated SIFT descriptors extracted and vocabulary trees are constructed. Their action recognition approach follows the standard paradigm using local features, vocabulary based representation and voting. Each frame is labeled by a vocabulary tree search, and final activity label is achieved by a voting of 5 successive frames.

In [117] a boosting EigenActions algorithm for human action recognition is proposed. A spatio-temporal Information Saliency Map (ISM) is calculated from a video sequence by estimating pixel density function. A continuous human action is segmented into a set of primitive periodic motion cycles from information saliency curve. Each cycle of motion is represented by a Salient Action Unit (SAU), which is used to determine the EigenAction using principle component analysis. A human action classifier is developed using multi-class AdaBoost algorithm with Bayesian hypothesis as the weak classifier. Given a human action video sequence, method locates the SAUs in the video, and recognizes the human actions by categorizing the SAUs.

Gorelick *et al.* [131] proposed a spatio-temporal patch correlation based method for human action recognition. Small spatio-temporal reference volumes are correlated against the entire video sequences in the target volume. The overall peak correlation value shows the matched actions. They utilized the properties of the poisson equation solution to analyze the spatio-temporal volume. Three dimensional space–time shapes are generated from the silhouettes of the spatio-temporal volume, and space–time salient features namely local space–time saliency, action dynamics, shape structure and orientation, are then extracted.

Schuldt *et al.* [116] compute local space-time features at locations selected in a scale-space representation. These features are used in an SVM classification scheme.

Yilmaz and Shah [103] extracted differential geometry features from the 3D contour of an action volume. The 3D contour is then projected to a 2D surface. The projection on the time axis forms the new spatio-temporal volume. Then the human moving speed, moving direction and human shape are extracted from the volume.

Niebles and Li [129] proposed a mixture hierarchical model for human action recognition based on spatial and spatio-temporal features. They showed that static shape features can improve the recognition performance when using spatio-temporal features. Liu *et al.* [161] proposed to use two types of feature sets for human action recognition. The first feature set is generated from the quantized vocabulary of local spatio-temporal volumes, while the second feature set is generated from spin images which capture the shape deformation of actions. It shows that action recognition accuracy can be improved by discovering relationships among the combined feature sets in the embedded Euclidian space.

In [146] Wang *et al.* presented a hierarchical probabilistic model (semi-latent Dirichlet allocation) for action recognition based on motion words, where each word corresponds to a frame in the video sequence, rather than a collection of words from vector quantization of space-time interest points. Their model is trained in a semi-supervised fashion. A new topic model called Semi-Latent Dirichlet Allocation (S-LDA) is proposed. The major difference to the Latent Dirichlet Allocation (LDA) model is that some of the latent variables in LDA are observed during the training stage in S-LDA.

In [118] a method for recognizing human actions based on pose primitives is proposed. In learning mode, the parameters representing poses and activities are estimated from videos. They obtain pose primitives by a Histogram of Oriented Gradient (HOG) based descriptor to better cope with articulated poses and cluttered background. Action classes are represented by histograms of poses primitives. Instead of having a look at individual pose primitives, to benefit from the temporal context they provide a sub-sequencing by means of n-gram expressions. Action recognition is based on a simple histogram comparison.

There are also a large number of works addressing the human action categorization problem by either spatial or temporal template matching.

Polana and Nelson [136], have 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. Efros *et al.* [111] 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 [123] 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.

Bobick and Davis [119] proposed to use temporal templates for representing and recognizing human actions. They constructed a binary motion-energy image (MEI) which represents the motion occurred in a video and generated a scalar-valued motion-history image (MHI) where intensity is a function of the motion history at each pixel. MEI and MHI together can be considered as a temporal template which is matched against the learned models of known movements.

Recently, improved temporal template [138] methods based on MEI and MHI are proposed. Yi *et al.* [122] proposed a pixel change ratio map (PCRM) which is similar to the concept of MHI, but PCRM is computed based on motion histogram. Babu and Ramakrishnanb [141] computed a motion flow history (MFH) from motion information in the compressed domain. Then they classified human activities by using the MHI.

In [115] Fathi *et al.* presented a method for human action recognition using mid-level motion features. Features are computed on a figure-centric representation, in which the human figure is stabilized inside a spatio-temporal volume. Mid-level motion features were constructed from low-level optical flow features computed on a local volume. For some small cuboids (of fixed size) inside the figure-centric volume, AdaBoost is applied to select a subset of the weak classifiers (low-level features) inside each figure-centric volume to construct better classifiers. Each mid-level feature is focused on a small cuboid inside the figure-centric volume, and is built from the low level features which best discriminate between pairs of action classes. AdaBoost is also used for a second time to train a final classifier from the mid-level motion features. This time AdaBoost will choose the best subset of mid-level motion features that can separate the two action classes.

In [144] authors extended the 2D Convolutional Neural Network approach to temporal domain by applying 3D convolutions. Their model extracts features from both spatial and temporal dimensions by performing 3D convolutions, capturing the motion information encoded in multiple adjacent frames. The developed architecture generates multiple channels, namely gray, gradient-x, gradient-y, opticalflow-x, and opticalflow-y, of information from adjacent input frames and perform convolution and subsampling separately in each channel. The final feature representation is computed by combining information from all channels. The output layer consists of the same number of units as the number of actions. For action recognition model parameters are trained by online error back-propagation algorithm.

In a recent work Rodriguez *et al.* [133], introduced 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. 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 (ACE), the Average Similarity Measure (ASM), and the Output Noise Variance (ONV). For activity recognition responses to the MACH filter is analyzed by thresholding.

2.2.2.2.2. State-space based approaches

The state–space based approach is a popular approach for human action recognition. An action is modeled as a set of states and connections in the state space using a Dynamic Probabilistic Network (DPN). Hidden Markov Model (HMM) is the most commonly used DPN has the advantages in modeling the time varying feature data.

Kale *et al.* [112] used an HMM to model human gait. Each individual is trained with one HMM; five representative binary silhouettes are used as hidden states for HMM training. In the recognition phase, the HMM which gives the largest probability is identified as the individual. Because the HMM training process requires the training dataset to be very large and representative, the model learning is very dependent on the training data.

Brand and Kettnaker proposed a Multi-Observation-Mixture + Counter Hidden Markov Model (MOMC-HMM) [132] to factorize the observation space. To factorize both the state and the observation space, Oliver *et al.* [135] proposed a Coupled Hidden Markov Model (CHMM) to model the temporal and causal correlations among hidden states. Recently, Xiang and Gong [145] developed a Dynamically Multi-Linked Hidden Markov Model (DML-HMM) for modeling human activities. The number of temporal processes in the DML-HMM is the same with the number of actions detected in the scene. So the DML-HMM has unsupervised manner and both the structure and the parameters are learned from training data. Ahmad and Lee [130] addressed action recognition by using combined shape flow and local-global motion flow. Based on the combined features, a set of multidimensional Hidden Markov Models were built to represent each action from multiple views. Shi *et al.* [140] proposed a semi-Markov discriminative approach to human action segmentation and recognition, by employing a Viterbi-like column generation algorithm. This approach gives efficient feature representation from motion segmentation. Ohet *et al.* [143] proposed a data-driven Markov chain Monte Carlo (MCMC) sampling method to enhance the duration modeling capabilities of switching linear dynamic systems. The proposed framework is robust to interpret complex human motions. Xiang *et al.* [142] constructed a contour graph from human contour sequences. The motion in the video is considered as an instance of random walks on the graph and sequence Monte Carlo method is used to estimate the random walks on the graph from the video.

In [120] for realizing event detection, the video is initially segmented into shots and for every resulting shot appropriate motion features are extracted at fixed time intervals, to form a motion observation sequence. Then, Hidden Markov Models (HMMs) are employed for associating each shot with a semantic event based on its formed observation sequence. Regarding the motion feature extraction procedure, a new representation for providing local-level motion information to HMMs is presented, while motion characteristics from previous frames are also exploited.

Yeffet *et al.* [126] proposed an activity recognition method inspired by local self similarity approach; every pixel at every frame is encoded as a short string of ternary digits (trits) by a process which compares this frame to the previous and to the next frame. The frame is then divided into *(mxn)* regions and the histograms of the trinary strings are computed for each of the *mn* region. These histograms are accumulated every few frames and the vector which contains all concatenated histograms serves as a video descriptor for a section of the video. They divide video to k equal time slices and compute the accumulated histograms for each region among the frames of each time slice. All *mn* region histograms for the *k* time slices are accumulated to one vector of length *512mnk* which is used to represent the entire video. In order to recognize an action, they apply a linear SVM classifier on the square-root values of the vectors. One of their contributions is to introduce the two additional unknowns (time shift and scale) by running several detectors in parallel, each observing different starting points and different scales of action length, which enables the system to label activity sequences with different end points and lengths.

In [134] a max-margin learning framework applied for training classifiers with structured latent variables. Mori *et al.* introduces a learning algorithm based on the cutting plane method and decomposed dual optimization. Motion features based on optical flows are used. A frame in a video is represented by a global motion feature extracted from the whole frame and a set of salient local patches. Their model consists of a root filter and a constellation of several hidden parts. The root filter models the compatibility of the action label and the global motion feature of the whole frame. A hidden part assigns a latent "part label" to a local patch. Intuitively, those "part labels" correspond to local motion patterns that are useful for discriminating different actions. Given a learned model, the classification is achieved by first finding the best labeling of the hidden parts for each action, then picking the action label with the highest score. The learning algorithm aims to set the model parameters so that the scores of correct action labels on the training data are higher than the scores of incorrect action labels by a large margin.

In [125] a framework for modeling motion by exploiting the temporal structure of the human activities is presented. Their main argument is that it is critical to incorporate temporal context information, particularly the temporal ordering of the movements. Information from motion segments are considered both for their visual features as well as their temporal composition. They represent activities as temporal compositions of motion segments. 3-D Harris corner detector is used to find interest points and each interest point is described by HoG (Histogram of Gradients) and HoF (Histogram of Flow) descriptors. A descriptor codebook is obtained by k-means clustering of the descriptors in the training set. During model learning and matching, histograms of codebook memberships over particular temporal ranges of a given video are computed. A video sequence is first decomposed into many temporal segments of variable length (including the degenerate case of the full sequence itself). Each video segment is matched against one of the motion segment classifiers by measuring image-based similarities as well as the temporal location of the segment with respect to the full sequence. The best matching scores from each motion segment classifier are accumulated to obtain a measure of the matching quality between the full action model and the query video.

2.2.3. Activity Modeling and Recognition

In this section some basic activity modeling and recognition methods are summarized. Since proposed method is based on the state space representation of the activity, we focused on only the methods that can be applied to the state space based approaches.

Understanding of activities may simply be thought as the classification of time varying feature data, i.e., matching an unknown test sequence with a group of labeled reference sequences representing typical activities. It is then obvious that a fundamental problem of activity recognition is to learn the reference activity sequences from training samples, and to devise both training and matching methods for coping effectively with small variations of the feature data within each class of motion patterns.

Dynamic time warping is a method for computing nonlinear time normalization between a template vector sequence and a test vector sequence. These two sequences could be of differing lengths. Forner-Cordero *et al.* [153] show experiments that indicate that the intrapersonal variations in gait of a single individual can be better captured by DTW rather than by linear warping. The DTW algorithm which is based on dynamic programming computes the best nonlinear time normalization of the test sequence in order to match the template sequence by performing a search over the space of all allowed time normalizations [154]. The space of all time normalizations allowed is cleverly constructed using certain temporal consistency constraints. Some of the temporal consistency constraints are given below:

- End point constraints. The beginning and the end of each sequence is rigidly fixed. For example, if the template sequence is of length N and the test sequence is of length M, then only time normalizations that map the first frame of the template to the first frame of the test sequence and also map the N^{th} frame of the template sequence to the M^{th} frame of the test sequence are allowed.
- The warping function (mapping function between the test sequence time to the template sequence time) should be monotonically increasing. In other words, the sequence of “events” in both the template and the test sequences should be the same.
- The warping function should be continuous.

Finite-state machine (FSM) is another tool that is used to model activities with its state-transition function. The states are used to decide which reference sequence matches with the test sequence. Wilson *et al.* [90] analyze the explicit structure of natural gestures where the structure is implemented by an equivalent of a FSM but with no learning involved. State-machine representations of behaviors have also been employed in higher level description; Bremond *et al.* [77] use handcrafted deterministic automata to recognize airborne surveillance scenarios describing vehicle behaviors in aerial imagery.

A HMM is a kind of stochastic state machines [78]. It allows a more sophisticated analysis of data with spatio-temporal variability. The use of HMMs consists of two stages: training and classification. In the training stage, the number of states of a HMM must be specified, and the corresponding state transition and output probabilities are optimized in order that the generated symbols can correspond to the observed image features of the examples within a specific movement class. In the matching stage, the probability with which a particular HMM generates the test symbol sequence corresponding to the observed image features is computed. HMMs generally outperform DTW for undivided time series data, and are therefore extensively applied to behavior understanding. Starner *et al.* [79] propose HMMs for the recognition of sign language. Oliver *et al.* [80] propose and compare two different state-based learning architectures, namely, HMMs and coupled Hidden Markov Models (CHMMs) for modeling people behaviors and interactions such as following and meeting. The CHMMs are shown to work much more efficiently and accurately than HMMs [70, 78]. Brand *et al.* [81] show that, by the use of the entropy of the joint distribution to learn the HMM, a HMM's internal state machine can be made to organize observed behaviors into meaningful states. This technique has found applications in video monitoring and annotation, in low bit-rate coding of scene behaviors, and in anomaly detection.

In [82] delay units are added to a general static neural network, and some of the preceding values in a time-varying sequence are used to predict the next value for the hand gesture recognition. Similar structure is used successfully for lip-reading in [83].

Brand [132] uses a simple nonprobabilistic grammar to recognize sequences of discrete behaviors. Ivanov *et al.* [84] describe a probabilistic syntactic approach to the detection and recognition of temporally extended behaviors and interactions between multiple agents. The fundamental idea is to divide the recognition problem into two levels. The lower level is performed using standard independent probabilistic temporal behavior detectors, such as

HMMs, to output possible low-level temporal features. These outputs provide the input stream for a stochastic context-free parser. The grammar and parser provide longer range temporal constraints, disambiguate uncertain low-level detection, and allow the inclusion of a priori knowledge about the structure of temporal behaviors in a given domain.

Wada *et al* [85] employ non-deterministic finite automata (NFA) as a sequence analyzer, because it is a simple example satisfying the following properties: instantaneousness and pure-nondeterminism. They present an approach for multiobject behavior recognition based on behavior driven selective attention.

Johnson *et al.* [86] describe the movement of an object in terms of a sequence of flow vectors, each of which consists of 4 components representing the positions and velocities of the object in the image plane. A statistical model of object trajectories is formed with two competitive learning neural networks that are connected with leaky neurons. Sumpter *et al.* [87] introduce feedback to the second competitive network giving a more efficient prediction of object behaviors. Hu *et al.* [88] introduced a new neural network structure that has smaller scale and faster learning speed, and is thus more effective. Owens *et al.* [89] apply the Kohonen self-organizing feature map to find the flow vector distribution patterns. These patterns are used to determine whether a point on a trajectory is normal or abnormal.

2.3. Discussion

Understanding human activities is one of the key challenges of the new generation intelligent systems.

Despite the fact that good results were achieved by traditional activity recognition approaches, they still have some limitations. Many of them involve computation of optical flow, whose estimation is difficult due to, e.g., aperture problems, smooth surfaces, and discontinuities. Others employ feature tracking and face difficulties in cases of self-occlusions, change of appearance, and problems of reinitialization. Methods that rely on key frames or eigenshapes of foreground silhouettes lack information about the motion. Some approaches are based on periodicity analysis and are thus limited to cyclic actions. Some of the recent successful work done in the area of activity recognition have shown that it is useful to analyze actions by looking at a video sequence as a space-time volume (of intensities, gradients, optical flow, or other local features), they result in a very compact representation but are too few to build activity models robust to background clutter, occlusion, and lack information about the temporal order of the data.

There is a main discussion on whether the dynamics or the shape features have better representation, and researches have shown a common conclusion that shape has rich information about the underlying action however introducing dynamics has lead to better recognition [2]. This conclusion is also supported by the studies on human activity perception.

To summarize approaches to the activity recognition problem; early body-part based methods suffer from the computational complexity, and lack of reliable segmentation of body parts in low resolution surveillance videos. More recent holistic methods achieve better results on unconstrained environments. These holistic methods process each frame (or segmented part of each frame as the moving subject) without focusing on the individual body parts.

Moreover some methods also supports holism on time domain by modeling each action as a “bag of features” where others have frame based time sequential representation or state sequential representation based on time segmented video. Time holistic representation has a main drawback that different time sequential data can produce the same feature set. On the other hand frame by frame representations have redundancy and comparing different length high dimensional continuous data is a complex task. State sequential representation is more appropriate for the action recognition problem, however segmenting video into states and getting a good representation of those states are the two main problems one must overcome to achieve a good performance. Frame sequential and state sequential representations must also handle unaligned video sequences, while this is not an issue in time holistic representation.

To recognize an action time holistic representation can be combined with either discriminative [155, 128] classifiers, semi-latent topic models [156] or unsupervised generative [137] models. Again such holistic representation of video sequences ignores temporal ordering and arrangement of features in the sequence.

To handle high dimensionality of the video data, most methods apply a simple feature selection based on filtering and/or dimensionality reduction (PCA) algorithm. Other methods employ clustering (SVM, kNN are the two common algorithms) or sampling data in fixed intervals. However there is not much work that is focused to eliminate redundant, irrelevant features which will help better representation.

Many approaches based on sequential representation of the activity, use the methods of *Dynamic Time Warping (DTW)* [154] or *Hidden Markov Model (HMM)* [2] to exploit temporal constraints. DTW deals with differences between sequences by operations of deletion-insertion, compression expansion, and substitution, of subsequences. By defining a metric of how much the sequences differ before and after these operations, DTW classifies the sequences. DTW lacks, however, the consideration of inter actions between nearby subsequences occurring in time. HMM considers this correlation between adjacent time instances by formulating a Markov process. The success of HMM models in dealing with speech data motivated vision researchers to apply HMMs to visual recognition problems. In contrast to speech recognition, computer vision lacks a general underlying modeling unit, i.e. how to map the images into symbols. Therefore, in order to recognize complex actions and interactions, researchers combine various HMM structures. Markov models have shown promise but require manual design by experts.

There are several outcomes of the state of the art literature:

- Considering both dynamics and shape features will lead to better representation.
- High dimensionality of the activity data must be handled properly.
- Temporal ordering of the features is also discriminative information.
- Training and recognition algorithm should handle unaligned samples of activity video of different lengths.

As the conclusion of our literature survey on activity recognition there are two main questions: which features will be used to represent the motion? And, how the underlying motion will be represented with extracted features?

CHAPTER 3

BACKGROUND

Nearly every activity recognition method starts with background/foreground segmentation. The aim of background/foreground segmentation is to segment moving parts from the rest of the image. Common approaches to background/foreground segmentation problem are summarized in section 2.1.1. Since our main contributions are on activity representation and modeling, algorithm applied at the preprocessing step is given in this chapter as a background.

Using spatio-temporal features is a popular approach to human activity recognition problem. One of the main motivations of spatio-temporal feature based methods is due to: representation of motion in a very compact way and application simplicity. Although model based approaches provide a very rich data for recognition of actions, model matching and tracking trough the video is a difficult and computationally complex task. Implicitly extracting body-part orientation or joint angles also lack of accuracy for low resolution video which is the most common case when considering surveillance data. Spatio-temporal feature based methods are based on finding interest points of the motion and representing corresponding information in an efficient way. However, choosing suitable interest points and representing spatio-temporal properties at those points are extremely critical for the performance of a recognition system.

Searching nature for finding an answer has lead researchers to examine behavior of human visual system. Physiological studies found simple cells, in human visual cortex, that are selectively tuned to orientation as well as to spatial frequency. It was suggested that the response of a simple cell could be approximated by 2D Gabor filters [103, 91, 92]. As the result of their success to represent spatial texture properties of images [93], their application to temporal domain emerges. In section 3.2. a brief background about the Gabor filters is given.

After achieving an initial spatio-temporal representation, next step is to find salient features, to effectively and compactly model each activity class. Feature selection algorithms are the tools to eliminate redundant and noisy features, and to choose an elite set of features that are necessary and sufficient to describe the concept. Feature selection provides great advantage over classification accuracies, as well as training and classification times [99]. However optimal feature selection can be seen as a search in a set of possible solutions and requires an exponentially large search space. In this work an evolutionary clustering method, Genetic Chromodynamics [94, 95], is adopted for activity feature selection problem. Original Genetic Chromodynamics procedure is given in section 3.3.1.

For activity recognition, evaluating the temporal ordering of the representation provides better discrimination between different activity classes. Hidden Markov Models are successful tools to model sequential data. Because of their applicability to wide range of applications there are many extended structures derived from the standard HMM structure [106]. One of such structures is profile HMM [96]. Profile HMM is originally proposed to model distant protein structures of the same family. Their sequence aligning ability and flexible learning strategy make them a suitable choice for the activity modeling problem. Standard HMM algorithm and its profile HMM extension is explained in sections 3.4.1 and 3.4.2 respectively.

3.1. Background/Foreground Segmentation

In this work as a preprocessing step, Adaptive Gaussian Mixture Model is used to model the background of the scene and pixels do not fit the model are assigned as the foreground [41].

3.1.1. Adaptive Gaussian Mixture Model

This method is based on modeling each background pixel by a mixture of K Gaussian distributions (K is a small number from 3 to 5). Different Gaussians are assumed to represent different colors. The weight parameters of the mixture represent the time proportions that those colors stay in the scene. The background components are determined by assuming that the background contains B highest probable colors. The probable background colors are the ones which stay longer and more static. Static single-color objects trend to form tight clusters in the color space while moving ones form widen clusters due to different reflecting surfaces during the movement. The measure of this was called the *fitness* value in their papers. To allow the model to adapt to changes in illumination and run in real-time, an update scheme was applied. It is based upon selective updating. Every new pixel value is checked against existing model components in order of fitness. The first matched model component is updated. If it finds no match, a new Gaussian component is added with the mean at that point and a large covariance matrix and a small value of weighting parameter.

Each pixel in the scene is modeled by a mixture of K Gaussian distributions. The probability that a certain pixel has a value of x_N at time N can be written as

$$p(x_N) = \sum_{j=1}^{K} w_j \eta(x_N; \theta_j), \tag{3.1}$$

where w_k is the weight parameter of the k^{th} Gaussian component. $\eta(x_N; \theta_j)$ is the Normal distribution of k^{th} component represented by

$$\eta(x; \theta_k) = \eta(x; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}, \tag{3.2}$$

where μ_k is the mean and $\Sigma_k = \sigma_k^2 I$ is the covariance of the k^{th} component.

The K distributions are ordered based on the fitness value w_k/σ_k and the first B distributions are used as a model of the background of the scene where B is estimated as

$$B = argmin_b\left(\sum_{j=1}^{b} w_j > T\right), \tag{3.3}$$

the threshold T is the minimum prior probability that the background is in the scene. Background subtraction is performed by marking a foreground pixel any pixel that is more than 2.5 standard deviations away from any of the B distributions. The first Gaussian component that matches the test value will be updated by the following update equations.

$$\hat{w}_k^{N+1} = (1-\alpha)\hat{w}_k^N + \alpha \hat{p}(w_k | x_{N+1}), \tag{3.4}$$

$$\hat{\mu}_k^{N+1} = (1-\alpha)\hat{\mu}_k^N + \rho x_{N+1}, \tag{3.5}$$

$$\hat{\Sigma}_k^{N+1} = (1-\alpha)\hat{\Sigma}_k^N + \rho(x_{N+1} - \hat{\mu}_k^{N+1})(x_{N+1} - \hat{\mu}_k^{N+1})^T, \tag{3.6}$$

$$\rho = \alpha \eta(x_{N+1}; \hat{\mu}_k^N, \hat{\Sigma}_k^N),$$

$$\hat{P}(w_k | x_{N+1}) = \begin{cases} 1\,; if\ w_k\ is\ the\ first\ match\ Gaussian\ component \\ 0\ ; otherwise \end{cases}, \tag{3.7}$$

where w_k is the k^{th} Gaussian component. $1/\alpha$ defines the time constant which determines change.

If none of the K distributions match the pixel value, the least probable component is replaced by a distribution with the current value as its mean, an initially high variance, and a low weight parameter.

Algorithm begins estimating of the Gaussian mixture model by expected sufficient statistics update equations then switch to *L*-recent window version when the first L samples are processed. The expected sufficient statistics update equations provide a good estimate at the beginning before all L samples can be collected. This initial estimate improves the accuracy of the estimate and also the performance of the tracker allowing fast convergence on a stable background model. The *L*-recent window update equations gives priority over recent data therefore the tracker can adapt to changes in the environment.

$$\widehat{w}_k^{N+1} = \widehat{w}_k^{N} + \frac{1}{L}\left(\hat{p}(w_k|x_{N+1}) - \widehat{w}_k^{N}\right), \tag{3.8}$$

$$\hat{\mu}_k^{N+1} = \hat{\mu}_k^{N} + \frac{1}{L}\left(\frac{\hat{p}(w_k|x_{N+1})x_{N+1}}{\widehat{w}_k^{N+1}}\right), \tag{3.9}$$

$$\widehat{\Sigma}_k^{N+1} = \widehat{\Sigma}_k^{N} + \frac{1}{L}\left(\frac{\hat{p}(w_k|x_{N+1})(x_{N+1}-\hat{\mu}_k^{N+1})(x_{N+1}-\hat{\mu}_k^{N+1})^T}{\widehat{w}_k^{N+1}} - \widehat{\Sigma}_k^{N}\right), \tag{3.10}$$

3.2. Gabor Filters

Since the discovery of crystalline organization of the primary visual cortex in mammalian brains forty years ago by Hubel and Wiesel [90], an enormous amount of experimental and theoretical research has greatly advanced our understanding of this area and the response properties of its cells. On the theoretical side, an important insight has been advanced by Marcelja [91] and Daugman [92] that simple cells in the visual cortex can be modeled by Gabor functions. The Gabor functions proposed by Daugman are local spatial bandpass filters that achieve the theoretical limit for conjoint resolution of information in the 2D spatial and 2D Fourier domains.

Gabor functions fist proposed by Denis Gabor as a tool for signal detection in noise. Gabor [152] showed that there exists a “quantum principle” for information; the conjoint time-frequency domain for 1D signal must necessarily be quantized so that no signal or filter can occupy less than certain minimal area in it. However, there is a trade of between time resolution and frequency resolution. Gabor discovered that Gaussian modulated complex exponentials provide the best trade off. For such a case, original Gabor elementary functions are generated with a fixed Gaussian, while the frequency of the modulating wave varies.

Gabor filters, rediscovered and generalized to 2D, are now being used extensively in various computer vision applications. Daugman generalized the Gabor function to the following 2D form in order to model the receptive fields of orientation selective simple cells:

$$g(x,y) = \mathrm{K}\exp(-\pi(\mathrm{a}^2(\mathrm{x}-\mathrm{x}_0)_\mathrm{r}^2 + \mathrm{b}^2(\mathrm{y}-\mathrm{y}_0)_\mathrm{r}^2))\exp\big(j(2\pi(u_0x+v_0y)+P)\big), \quad (3.11)$$

where (u_0, v_0) and P define the spatial frequency and the phase of the sinusoid respectively. (x_0, y_0) is the peak of the Gaussian function, *a* and *b* are scaling parameters of the Gaussian, and the *r* subscript stands for a rotation operation such that:

$$(x-x_0)_r = (x-x_0)cos\theta + (y-y_0)sin\theta, \quad (3.12)$$

$$(y-y_0)_r = -(x-x_0)sin\theta + (y-y_0)cos\theta, \quad (3.13)$$

Recent neurophysiological evidence suggests that the spatial structure of the receptive cells having different sizes virtually invariant. Daugman [92] have proposed that ensemble of simple cells best modeled as a family of 2D Gabor filters sampling frequency domain in a log polar manner. This class is equivalent to a family of affine coherent states that generated by rotation and dilation.

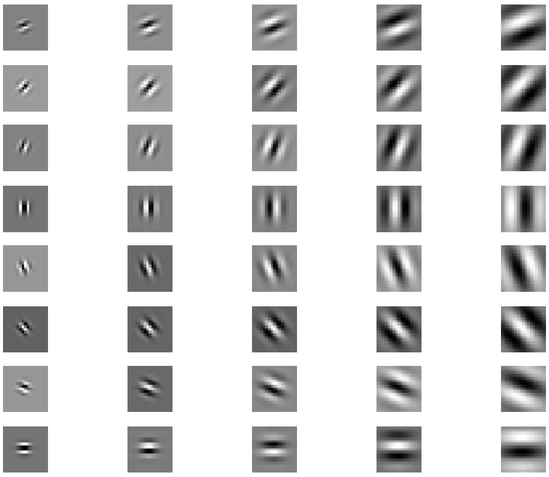


Figure 3: 2D Gabor kernels at 5 different scales and 8 different orientations

Each member of this family of Gabor filters models the spatial receptive field structure of a simple cell in the primary visual cortex. The Gabor decomposition can be considered as a directional microphone with an orientation and scaling sensitivity. Due to the end-inhibition property of these cells, they respond to short lines, line ending and sharp changes in curvature. Since such curves correspond to some low level salient features in an image, these cells can be assumed to form a low level feature map of the intensity image.

3.3. Feature Selection

Feature extraction (or dimensionality reduction) is an important research topic in computer vision and pattern recognition fields. Because the curse of high dimensionality is usually a major cause of limitations of many practical technologies and the large quantities of features may even degrade the performances of the classifiers when the size of the training set is small compared to the number of features.

Many feature extraction methods have been proposed, in which the most well-known ones are *Principal Component Analysis* (*PCA*) [148] and *Linear Discriminant Analysis* (*LDA*) [149]. However, there are still some limitations for directly applying them to solve vision problems.

Firstly, although *PCA* is a popular unsupervised method which aims at extracting a subspace in which the variance of the projected data is maximized (or, equivalently, the reconstruction error is minimized), it does not take the class information into account and thus may not be reliable for classification. On the contrary, *LDA* is a supervised technique which has been shown to be more effective than *PCA* in many applications. It aims to maximize the between class scatter and simultaneously minimize the within-class scatter.

These methods do not deal directly with eliminating irrelevant and redundant variables, but are rather concerned about transforming the observed variables into a small number of “projections”, or “dimensions”. The linear methods are not able to reduce the number of original features as long as all the variables have non-zero weights in the linear combination.

Feature selection methods are used to find the set of features that yield the best classification accuracy for a given data set. There are many feature selection algorithms in the literature based on two main approaches: filter and wrapper [99].

Filter methods are independent of the learning algorithm and can be seen as a pre-processing step: elimination of non-useful features. On the other hand wrapper methods, which use a learning algorithm as a subroutine, have higher computational complexity but better classification performance. Recent studies on feature selection also suggest using embedded algorithms which combines wrapper methods with a preprocessing step of filtering.

In most wrapper methods, a prototypical vector (the cluster center) identifies a cluster. The problem of cluster optimization can be divided into two: optimization of cluster centers and determination of number of clusters. The determination of number of clusters has often been neglected in standard approaches (static clustering methods, i.e. the c-means algorithm, Kohonenmaps, elastic nets and fuzzy c-means), as these typically fix the number of clusters a priori.

Genetic algorithms (GA), a form of inductive learning strategy, are adaptive search technique [150]. Genetic Algorithms are designed to simulate A Hybrid Feature Selection Algorithm the evolutionary processes that occur in nature. The basic idea is derived from the Darwinian theory of survival of the fittest.

There are three fundamental operators in GA: selection, crossover and mutation within chromosomes. As in nature, each operator occurs with a certain probability. There must be a fitness function to evaluate individuals' fitness. The evaluation function is a very important component of the selection process since offspring for the next generation are determined by the fitness values of the present population. Crossover and mutation are used to generate new individuals (offspring) for the next generation. Crossover operates by randomly selecting a point in the two selected parents and exchanging the remaining segments of the parents to create new individuals. Mutation operates by randomly changing one or more components of a selected individual.

Evolutionary algorithms represent ideal tools for solving difficult optimization problems. Several optimization problems for which classical methods do not work very well or are simply inapplicable can be solved with evolutionary techniques. However standard evolutionary algorithms find only one solution, even if the search space is a highly multimodal domain. Genetic Chromodynamics (GC) is a recent evolutionary unsupervised clustering procedure that maintains population diversity and forces the formation of stable sub-populations [94].

3.3.1. Genetic Chromodynamics

The GC framework has demonstrated success in application to function optimization, clustering and classification [95].

The main idea of the GC strategy is the formation and maintenance of stable subpopulations. In particular, GC is able to concentrate search on many basins of attraction in parallel, so that several optima are found simultaneously.

Each subpopulation will evolve towards the optimum of the region. A characteristic of the method is that the number of individuals decreases: very similar/closed individuals are unified, thus, dimension of the population decreases. Using a local scheme of chromosomal crossover and a unification mechanism of similar individuals, the strategy makes every subpopulation to contain finally just one chromosome, which represents the corresponding local optimum. The advantage of this method is the fact that finally both multiple optima and their number in the searching space could be determined [94].

In the beginning, the population size is high and, as the algorithm progresses, the number of solutions (chromosomes) can be reduced by each generation. In GC algorithm, each individual participates in the forming of the new generation: its mate is searched by applying a local selection scheme; if a second chromosome is found within its local range (called the mating region), they recombine and the competition for survival of the fittest is held between the resulting offspring and the first parent only. If no chromosome can be found in the local range of the considered chromosome, it is mutated. The removal of less fit solutions is performed by introducing a special operator that merges very similar chromosomes into one chromosome; this chromosome is usually considered to be the fittest one of the chromosomes to be merged.

Algorithm 1: Merging Procedure

```
Repeat
    a chromosome $F$ is considered as the current chromosome;
    select all $m$ chromosomes in the merging region of $F$, including itself;
    combine chromosomes
    remove all but the best chromosome from the selection;
until merging cannot be applied at all
```

Algorithm 2: GC algorithm

```
k = 0;
initialize population P(k);
Repeat
    for all chromosomes F in the population do
        if mating region of F is empty then
            apply mutation to F;
            if obtained F' is fitter than F then
                replace F;
            end if
        Else
            choose one chromosome from the mating region of F for
            crossover;
            obtain and evaluate one offspring;
            if the offspring has better fitness than F then
                replace F
            end if
        end if
    end for
merging;
k = k + 1;
until there is no change
```

3.4. Hidden Markov Models

There are many different HMM structures in literature proposed for modeling different kinds of data. HMMs are well suited for recognition of complex human actions because they can efficiently characterize motion profiles in spite of the broad variation in the space and time domains in which actions are performed.

Variations between different performances of the same activity, with the possible effects of unaligned sequences, make it hard to select an initial model and maintain a static structure throughout the training. The nature of the data analyzed in this work also requires sequence aligning.

In traditional HMM approaches structure learning is independent of parameter learning and learning is first performed off-line and then system switches to a utilization stage where no further learning is performed [1].

3.4.1. Standard Hidden Markov Models

Hidden Markov Model is a statistical model which characterizes the statistical properties of the signal [1]. The underlying assumption is that the signal can be well characterized as a parametric random process, and that the parameters of the stochastic process can be determined in a precise, well defined manner.

An HMM can be characterized by the following:

1) N, the number of states in the model

$$S = \{S_1, S_2, \dots, S_N\} \tag{3.14}$$

2) M, the number of distinct observation symbols per state

$$V = \{V_1, V_2, \dots, V_M\} \tag{3.15}$$

3) The state transition probability distribution $A = \{a_{ij}\}$, where

$$a_{ij} = P\left[q_{t+1} = S_j | q_t = S_i\right] \qquad 1 \leq i, j \leq N \tag{3.16}$$

4) The observation probability distribution in state j, $B = \{b_j(k)\}$, where

$$b_j(k) = P\left[V_k \; at \; t | q_t = S_j\right] \quad 1 \leq j \leq N, 1 \leq k \leq M \tag{3.17}$$

5) The initial state distribution $\pi = \{\pi_i\}$ where

$$\pi_i = P\{q_1 = S_i\} \;\; 1 \leq i \leq N \tag{3.18}$$

Process moves from one state to another generating a sequence of states. According to the Marcov chain property, probability of each subsequent state depends only on what was the previous state. To define Markov model, the transition probabilities and initial probabilities have to be specified.

There are three basic problems of interest that must be solved for the model to be useful in real-world applications.

Evaluation problem. Given the HMM $\lambda = (A, B, \pi)$ and the observation sequence $O = O_1 O_2 O_3 \dots O_t$, calculate the probability that observed sequence was produced by the model $P(O|\lambda)$. Scoring how well a given model matches a given observation sequence. To solve this problem efficiently forward backward procedure is used.

Consider the variable $\alpha_t(i)$ defined as the probability of the partial observation sequence until time t and state S$_i$ at time t, given the model λ.

$$\alpha_t(i) = P(O_1 O_2 O_3 \dots O_t, q_t = S_i|\lambda) \tag{3.19}$$

$\alpha_t(i)$, can be solved inductively with forward recursion.

Forward recursion

Initialization:

$$\alpha_1(i) = \pi_i b_i(O_1) \quad 1 \leq i \leq N \tag{3.20}$$

Recursion:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}) \quad 1 \leq t \leq T-1, \; 1 \leq j \leq N \tag{3.21}$$

Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{3.22}$$

In a similar manner backward variable $\beta_t(i)$, which will be used in the solution of the 3rd problem, is defined as the probability of the partial observation sequence from t+1 to the end, given state S$_i$ at time t and the model λ.

$$\beta_t(i) = P(O_{t+1} O_{t+2} O_{t+3} \dots O_T, q_t = S_i|\lambda) \tag{3.23}$$

$\beta_t(i)$, can be solved inductively with backward recursion.

Backward recursion

Initialization:

$$\beta_T(i) = 1 \quad 1 \leq i \leq N \tag{3.24}$$

Recursion:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(O_{t+1}) \beta_{t+1}(i) \quad t = T-1, T-2, \dots, 1 \quad 1 \leq i \leq N \tag{3.25}$$

Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \beta_1(i)\, b_i(O_1)\, \pi_i \tag{3.26}$$

Decoding problem. Given the HMM $\lambda = (A, B, \pi)$ and the observation sequence$O = O_1 O_2 O_3 \dots O_t$, calculate the most likely sequence of hidden states S$_i$ that produced this observation sequence O. The solution to this problem depends upon the way "most likely state sequence" is defined. One approach is to find the most likely state q_t at *t=t* and to concatenate all such q_t. But some times this method does not give a physically meaningful state sequence. In a better method, commonly known as *Viterbi algorithm*, the whole state sequence with the maximum likelihood is found. In order to facilitate the computation we define an auxiliary variable, $\delta_k(i)$ as the maximum probability of producing observation sequence O when moving along any hidden state sequence $q_1 \ldots q_{k-1}$ and getting into $q_k = S_i$.

$$\delta_k(i) = \max P(q_1 \ldots q_{k-1}, q_k = S_i,\ O_1 O_2 \ldots O_k) \tag{3.27}$$

where max is taken over all possible paths $q_1 \ldots q_{k-1}$.

Viterbi Algorithm

Initialization:

$$\delta_1(i) = \max P(\,q_1 = S_i,\ O_1) = \pi_i b_i(O_1) \qquad 1 \le i \le N \tag{3.28}$$

Forward recursion:

$$\begin{aligned} \delta_k(j) &= \max P(q_1 \ldots q_{k-1}, q_k = S_j,\ O_1 O_2 \ldots O_k) = \\ &\max\nolimits_{\mathrm{i}}\left[a_{ij} b_j(O_k) \max P(q_1 \ldots q_{k-1} = S_j,\ O_1 O_2 \ldots O_{k-1})\right] = \\ &\max\nolimits_{\mathrm{i}}\left[a_{ij} b_j(O_k) \delta_{k-1}(j)\right] \qquad 1 \le j \le N, 2 \le k \le K \end{aligned} \tag{3.29}$$

Termination:

choose best path ending at time K

$$\max\nolimits_{\mathrm{i}}[\delta_K(i)] \tag{3.30}$$

Back-track best path.

Learning problem. Given some training observation sequence $O = O_1 O_2 O_3 \ldots O_t$, and general structure of HMM (numbers of hidden and visible states), determine HMM parameters $\lambda = (A, B, \pi)$ that best fit training data.

Define $\xi_t(i,j) = P\big(q_t = S_i, q_{t+1} = S_j \big| O, \lambda\big)$, from the definitions of the forward-backward variables

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)} = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)} \tag{3.31}$$

The probability of being in state S_i at time t, $\gamma_t(i)$, can be related to $\xi_t(i,j)$ by summing over j,

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j), \tag{3.32}$$

$$\sum_{t=1}^{T-1}\gamma_t(i) = expected\ number\ of\ transitions\ from\ S_i, \tag{3.33}$$

$$\sum_{t=1}^{T-1}\xi_t(i,j) = expected\ number\ of\ transitions\ from\ S_i\ \ to\ S_j, \tag{3.34}$$

Using above formulas a set of reasonable reestimation formulas for π, A and B are

$$\bar{\pi}_i = expected\ frequency\ (number\ of\ times) in\ state\ S_i\ at\ time\ (t=1) = \gamma_1(i),$$

$$\bar{a}_{ij} = \frac{expected\ number\ of\ transitions\ from\ state\ S_i\ to\ S_j}{expected\ number\ of\ transitions\ from\ state\ S_i} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{3.35}$$

$$\bar{b}_j(k) = \frac{expected\ number\ of\ times\ in\ state\ j\ and\ observing\ symbol\ V_k}{expected\ number\ of\ times\ in\ state\ j} = \frac{\substack{\sum_{t=1}^{T-1}\gamma_t(i) \\ s,t,O_t=V_k}}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{3.36}$$

If current model is defined as $\lambda = (A, B, \pi)$ and the reestimated model is defined as $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$. According to the Baum-Welch method if we iteratively use $\bar{\lambda}$ in place of λ and repeat the reestimation calculation, we then improve the probability of O being observed from the model until some limiting point is reached. Because either the initial model λ defines a critical point of the likelihood function ($\lambda = \bar{\lambda}$) or model $\bar{\lambda}$ is more likely than model λ. The final result of this reestimation procedure is called a maximum likelihood estimate of the HMM. This method can be derived using simple ``occurrence counting" arguments or using calculus to maximize the auxiliary quantity

$$Q(\lambda, \bar{\lambda}) = \sum_Q P(Q|O,\lambda) log\left[P(Q|O,\bar{\lambda})\right] \tag{3.37}$$

over $\bar{\lambda}$.

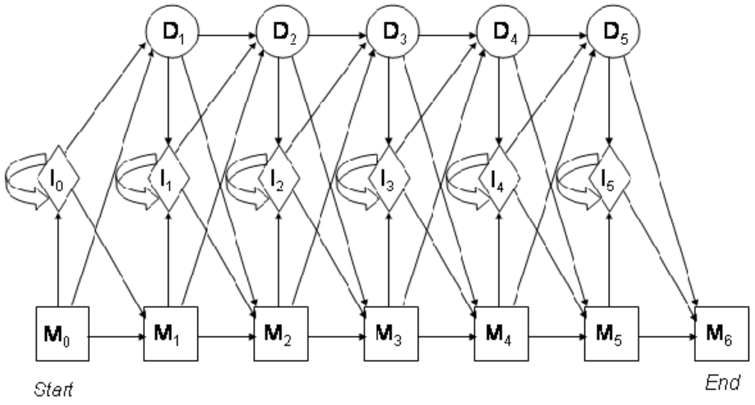


Figure 4: Standard profile HMM structure

3.4.2. Profile Hidden Markov Models

In order to model real sequences one must need to consider the possibility that, gaps might occur when a model is aligned to a sequence. Two types of gaps may arise. The first type occurs when the sequence contains a region that is not present in the model (an insertion in the sequence). The second type occurs when there is a region in the model that is not present in the sequence (a deletion in the sequence).

A Profile HMM is a standard topology for modeling sequence motifs [96]. Aligning a sequence to a profile HMM is done by a Viterbi algorithm that finds the most probable path that the sequence may take through the model, using the transition and emissions probabilities to score each possible path. In general, if the sequence is equivalent to the consensus of the original alignment, the path through the model will pass from match state to match state in a linear fashion. If the sequence contains a deletion relative to the consensus, the path passes through one or more delete states before transitioning to the next match state; if the sequence contains an insertion relative to the consensus, the path passes through an insert state between two match states.

There are three kinds of states in profile HMM (figure 4): *matching states (M), deletion states (D)* and *insertion states (I).* Matching states are the ones that emit codebook symbols. When there is a region in the model that is not present in the sequence model transits to the Deletion state. *Insertion states (I)* handle the cases when the sequence contains a region that is not present in the model.

Constructing a Profile HMM from an unlabeled data:

Profile HMM formalism can be used to model a shared pattern in sequences. If the sequences are already aligned, then we have labeled data. In other words, one can determine from the alignment which state is associated with each symbol in each sequence. In that case, required initialization is to determine the number of match states in the Profile HMM, set up the topology, and determines the parameters from the labeled data.

Given unlabeled sequences that are known to share a pattern, Profile HMM can be used to discover the pattern, label the data, and construct a multiple sequence alignment.

To discover a pattern in unlabeled data requires the following steps:

1. Estimating the length: Given a set of unaligned sequences, where each sequence is an instance of the pattern, let L, the length of HMM (i.e., the number of match

states) be the average length of sequences. If the given sequences contain a pattern but are much longer than the pattern, then some approach to estimating the length is required.

2. The topology: Construct a Profile HMM with L+2 match states. M_0 and M_{L+1} are silent states corresponding to the start state and the end state.

3. Learn parameters Guess "good" initial parameters The model should be encouraged to use 'sensible' transitions; for instance transitions into match states should be large compared to other transition probabilities.

4. Train model using Baum Welch.

5. Determining the motif Use the Viterbi algorithm or posterior decoding to find path most likely to produce each sequence. The Viterbi recurrence can be greatly simplified and expressed in terms of log odds for the special case of Profile HMMs. The log odds formulation avoids underflow and to reduces length effects.

6. Multiple Sequence Alignment: The most paths for each sequence obtained from decoding can be used to obtain a multiple alignment of the input sequences.

7. Model surgery: The topology of the model can be iteratively refined. If more than half of the sequences enter the delete state at a particular position remove that match state from the topology. If more than half of the sequences enter the insert state at a given position, add match states (number equal to average length of the insertion).

8. Re-estimate the parameters: If the states change due to model surgery, re-estimation of the parameters would be required. Label the multiple alignment with the new states and calculate the transition and emission probabilities as described above for labeled data. If the number of states that are changed is a significant percentage of the entire HMM, structure can be retrained.

Compared with the exact dynamic programming algorithm for multiple sequence alignment, which runs in exponential time, this approach can align many sequences quickly. Note that this method doesn't say how to align sequences of different length. Correspond to unconserved portions, not meaningfully alignable; often just left-justified and shaded.

Model surgery

After training a model we can analyze the alignment it produces:

- From counts estimated by the forward-backward procedure how much a certain transition is used by the training sequences can be seen.
- The usage of a match state is the sum of counts for all letters emitted in the state.
- If a certain match state is used by less than half the number of given sequences, the corresponding module (triplet of match, insert, delete states) should be deleted.
- Similarly, if more than half (or some other predefined fraction) of the sequences use the transitions into a certain insert state, this should be expanded to some number of new modules (usually the average number of insertions).

Profile HMM training from unaligned sequences

Initialization:

- Choose the length of the profile HMM (i.e., the number of match states), and initialize the transition and emission parameters.
- A commonly used rule is to set the profile HMM's length to be the average length of the training sequences.

Training:

- Estimate the model using the Baum-Welch algorithm or its Viterbi alternative.
- Start Baum-Welch from multiple different points to see if it all con verges to approximately the same optimum.
- If necessary, use a heuristic method for avoiding local optima.

Multiple alignment:

Align all sequences to the final model using the Viterbi algorithm and build a multiple alignment.

Profile HMM Forward Algorithm

Initialization: $\alpha_{M_0}(0) = 1$ (3.38)

Recursion:

$$\alpha_{M_j}(i) = e_{M_j}(x_i)\left[\alpha_{M_{j-1}}(i-1)a_{M_{j-1}M_j} + \alpha_{I_{j-1}}(i-1)a_{I_{j-1}M_j} + \alpha_{D_{j-1}}(i-1)a_{D_{j-1}M_j}\right] \quad (3.39)$$

$$\alpha_{I_j}(i) = e_{I_j}(x_i)\left[\alpha_{M_j}(i-1)a_{M_jI_j} + \alpha_{I_j}(i-1)a_{I_jI_j} + \alpha_{D_j}(i-1)a_{D_jI_j}\right] \quad (3.40)$$

$$\alpha_{D_j}(i) = \alpha_{M_{j-1}}(i)a_{M_{j-1}D_j} + \alpha_{I_{j-1}}(i)a_{I_{j-1}D_j} + \alpha_{D_{j-1}}(i)a_{D_{j-1}D_j} \tag{3.41}$$

Termination:

$$\alpha_{M_{L+1}}(n+1) = \alpha_{M_L}(n)a_{M_L M_{L+1}} + \alpha_{I_L}(n)a_{I_L M_{L+1}} + \alpha_{D_L}(n)a_{D_L M_{L+1}} \tag{3.42}$$

Profile HMM Backward Algorithm

Initialization:

$$\beta_{M_{L+1}}(n+1) = 1 \tag{3.43}$$

$$\beta_{M_L}(n) = a_{M_L M_{L+1}},\ \beta_{I_L}(n) = a_{I_L M_{L+1}},\ \beta_{D_L}(n) = a_{D_L M_{L+1}} \tag{3.44}$$

Recursion:

$\beta_{M_j}(i) =$

$$\beta_{M_{j+1}}(i+1)a_{M_j M_{j+1}}e_{M_{j+1}}(x_{i+1}) + \beta_{I_j}(i+1)a_{M_j I_j}e_{M_j}(x_{i+1}) + \beta_{D_{j+1}}(i)a_{M_j D_{j+1}} \tag{3.45}$$

$$\beta_{I_j}(i) = \beta_{M_{j+1}}(i+1)a_{I_j M_{j+1}}e_{M_{j+1}}(x_{i+1}) + \beta_{I_j}(i+1)a_{I_j I_j}e_{M_j}(x_{i+1}) + \beta_{D_{j+1}}(i)a_{I_j D_{j+1}} \tag{3.46}$$

$$\beta_{D_j}(i) = \beta_{M_{j+1}}(i+1)a_{D_j M_{j+1}}e_{M_{j+1}}(x_{i+1}) + \beta_{I_j}(i+1)a_{D_j I_j}e_{M_j}(x_{i+1}) + \beta_{D_{j+1}}(i)a_{D_j D_{j+1}} \tag{3.47}$$

Profile HMM Viterbi Algorithm

Initialization: $v_{M_0}(0) = 1$ (3.48)

Recursion:

$$v_{M_j}(i) = e_{M_j}(x_i)\ max \begin{cases} v_{M_{j-1}}(i-1)\ a_{M_{j-1}M_j} \\ v_{I_{j-1}}(i-1)\ a_{I_{j-1}M_j} \\ v_{D_{j-1}}(i-1)\ a_{D_{j-1}M_j} \end{cases} \tag{3.49}$$

$$v_{I_j}(i) = e_{I_j}(x_i)\ max \begin{cases} v_{M_{j-1}}(i-1)\ a_{M_{j-1}I_j} \\ v_{I_{j-1}}(i-1)\ a_{I_{j-1}I_j} \\ v_{D_{j-1}}(i-1)\ a_{D_{j-1}I_j} \end{cases} \tag{3.50}$$

$$v_{D_j}(i) = max \begin{cases} v_{M_{j-1}}(i-1)\ a_{M_{j-1}D_j} \\ v_{I_{j-1}}(i-1)\ a_{I_{j-1}D_j} \\ v_{D_{j-1}}(i-1)\ a_{D_{j-1}D_j} \end{cases} \tag{3.51}$$

Termination:

Final score is $v_{M_{L+1}}(n)$, calculated using the top recurtion relation.

Track back the optimal state sequence.

CHAPTER 4

PROPOSED HUMAN ACTIVITY RECOGNITON METHOD

Compared to the spatio-temporal feature based approaches, which models the activity as the "bag of features", proposed algorithm also considers their temporal order. Our 3D Gabor based feature extraction maintains locality on both space and time. However to recognize an activity, frame by frame modeling is not an adequate solution, mainly because of the huge redundancy and high dimensionality of the data. Instead a feature selection algorithm is proposed to achieve state based representation.

In this thesis a recent evolutionary feature selection method, genetic Chromodynamics, is adopted to obtain state-space based representation of the underlying human activity by local spatio-temporal feature clusters. Spatio temporal features and state representation of the action sample is not though a predefined structure; i.e. number of spatio temporal features and number of states are not fixed.

A profile HMM, is used to learn the underlying common pattern for each action class. In the literature, Profile HMM formalism is used to model a shared pattern in biomolecular sequences. Profile HMM structure provides an effective tool for action modeling to overcome the restrictions of common HMM methods. Standard profile HMM further adopted to handle repeating subsequences and unlabeled end points of the activity in a given sequence.

At the end of the learning proposed method automatically builds an activity specific model for each class of activity. Recognition is done by evaluating the probe activity by each model.

In the following sections, the proposed activity recognition method will be explained and details about building blocks of our algorithm namely: 3D Gabor filters, Genetic Chromodynamics and Profile HMM will be given.

4.1. Preprocessing

In this work as a preprocessing step, regions of probe action are segmented, and in the following sections only those regions are processed. We used the approach in [41] which improves the adaptive background mixture model, since it learns faster and more accurately as well as adapts effectively to changing environments. Details of the algorithm are given in section 3.1. The reason for the segmentation of the moving region is to obtain person centered coordinate frame in the later processes.

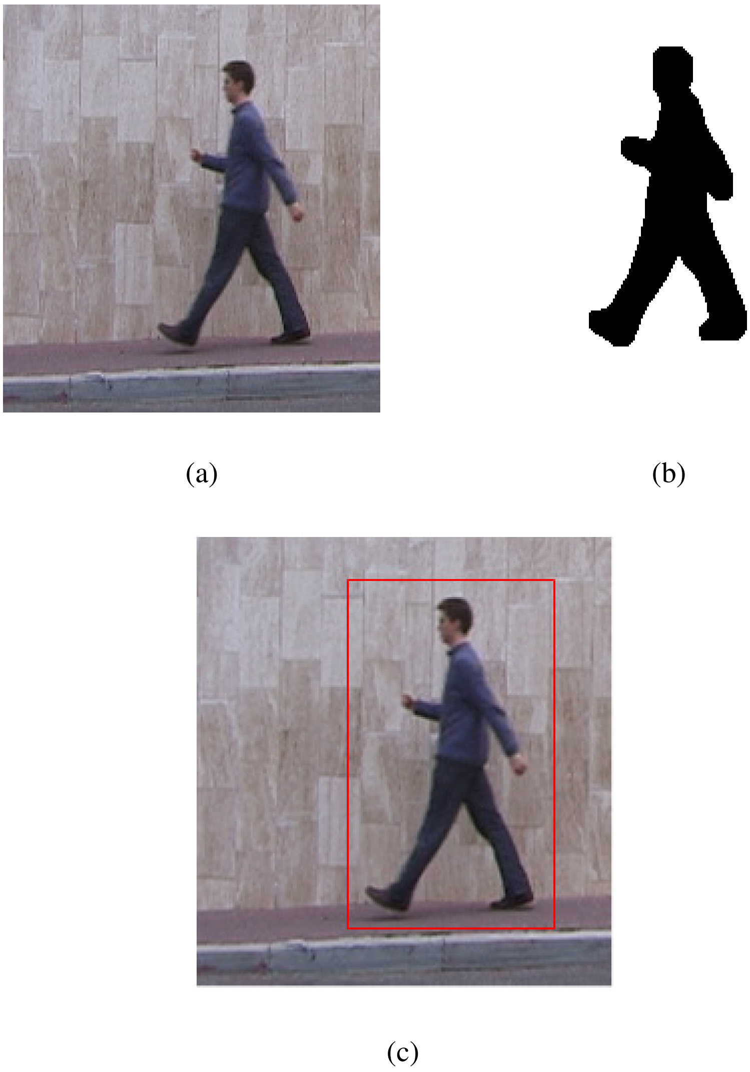


Figure 5: Preprocessing of a frame a) original frame, b) foreground mask, c) region that will be processed in the later steps of the algorithm.

4.2. Spatio-Temporal Action Representation

4.2.1. 3D Gabor Filters

Using 3D Gabor filters are similar to enhancing edge contours, as well as valleys and ridge contours in both spatial and temporal domain. Such an approach enables finding interest points as well as extracting their local representation.

3D Gabor filters has very similar aspects with 2D Gabor filters. Studies on the spatio-temporal model of the human visual system (HVS) suggest that the 3D Gabor filters unify the treatment of spatial and spatiotemporal aspects of the response selectivity of the V1 cells.

Our motivation is to simulate the behavior of the HVS with a 3D Gabor filter bank which decomposes the data into perceptual channels, each one being tuned to a specific spatial frequency, orientation and temporal frequency. 3D Gabor filters allows the description of temporal frequency structure and temporal relations, in addition to the description of the spatial frequency structure and spatial relations.

A 3D Gabor is the product of a 3D Gaussian and a 3D harmonic function. The length of the axes is controlled by the Gaussian and the frequency is controlled by the harmonic function.

$$w(x) = \exp(-\pi x^T x), \tag{4.1}$$

$$s(x) = exp\ (j2\pi u_0^T\, x), \tag{4.2}$$

$$g(x) = Kexp(jP)w\big(A(x - x_0)\big)s(x), \tag{4.3}$$

where K scales the magnitude of the Gaussian envelope, A is the height of the peak of the Gaussian, $\vec{x}_0 = (x_0, y_0, t_0)$ is the location of the peak of the Gaussian envelope, $\vec{u}_0 = (u_0, v_0, w_0)$ spatial and temporal frequencies of the sinusoid carrier in Cartesian, P is the phase of the sinusoid carrier. $\vec{u}_0$ can also be expressed in spherical coordinates with magnitude F_0 and directions β, θ:

$$u_0 = F_0 sin\theta cos\beta, \tag{4.4}$$

$$v_0 = F_0 sin\theta sin\beta, \tag{4.5}$$

$$w_0 = F_0 cos\theta, \tag{4.6}$$

The Gabor function as defined above may have a non-zero DC response

$$\hat{g}(0) = \frac{K}{\|A\|}\exp(jP)\exp(j2\pi u_0^T x_0)\, w(A^{-T}u_0), \tag{4.7}$$

where *w(x)=w(-x).* By eliminating the DC response filters will not respond to the absolute intensity of the image. One approach to doing so is to subtract from the original filter the output of a low pass filter.

$$h(x) = g(x) - Cf(x), \tag{4.8}$$

where *C* is a constant and *f(.)* is the low-pass filter. A convenient and popular low-pass filter is as follows

$$f(x) = \frac{K}{\|A\|} w\big(A(x - x_0)\big), \tag{4.9}$$

Using (3.8)

$$f(x) = \frac{K}{\|A\|} w\big(A(x - x_0)\big)(\exp(jP)\exp(j2\pi u_0^T x_0) - C), \tag{4.10}$$

Corresponds to subtracting a complex constant from the complex sinusoid carrier. Note *f* is a Gabor filter with zero phase and zero peak response. Therefore it has the following Fourier Transform;

$$\hat{f}(u) = \frac{K}{\|A\|}\exp(-j2\pi u^T x_0)\, w(A^{-T}u), \tag{4.11}$$

Thus the DC response of the combined filter is as follows

$$\hat{h}(0) = \hat{g}(0) - C\hat{f}(0) = \frac{K}{\|A\|}(\exp(jP)\exp(j2\pi {u_0}^T x_0)\, w(A^{-T}u_0) - C), \tag{4.12}$$

And to get a zero DC response C must be set to;

$$C = \exp\Big(j(P + 2\pi u_0^T x_0)\Big)\, w(A^{-T}u_0), \tag{4.13}$$

In this work we set $x_0 = y_0 = t_0 = 0$, *P*=0, and scales of the Gaussian magnitude are same for each axes. The rotation angle has no effect. $\sigma_x = \sigma_y = \sigma_t = \sigma$

$$g(x, y, t) = Kexp\big(-\pi\sigma^2(x^2 + y^2 + t^2)\big)$$

$$\Big(\exp\big(2\pi j(u_0 x + v_0 y + w_0 t)\big) - \exp\Big(-\frac{\pi}{\sigma^2}({u_0}^2 + {v_0}^2 + {w_0}^2)\Big)\Big), \tag{4.14}$$

3D Gabor function will be:

$$g(x,y,t) = Kexp(-\pi\sigma^2(x^2+y^2+t^2))\left(\exp\left(j2\pi F_0(xsin\theta cos\beta + ysin\theta sin\beta + tcos\theta)\right) - \exp\left(-\frac{\sigma^2}{2}\right)\right), \tag{4.15}$$

$$g_j(\vec{x}\,) = Kexp(-\pi\sigma^2\vec{x}^2)\left(\exp\left(j2\pi\vec{k}_j\vec{x}\right) - exp\left(-\frac{\sigma^2}{2}\right)\right) \tag{4.16}$$

$$\vec{k}_j = \begin{pmatrix} k_v\ sin\theta cos\beta \\ k_v\ sin\theta sin\beta \\ k_v\ cos\theta \end{pmatrix} \tag{4.17}$$

$\vec{k}_j$ is the spatial and temporal frequencies of the sinusoid carrier in spherical coordinates with magnitude k_v and directions β, θ.

To find 3D Gabor responses spatio-temporal convolution of the video sequence and the 3D Gabor kernels is computed.

$$R_j(x_0,y_0,t_0) = \sum_{t=0}^{\frac{w_t}{2}} \sum_{x=0}^{\frac{w_x}{2}} \sum_{y=0}^{\frac{w_y}{2}} g_j(x,y,t)\, V\left(x_0 - \frac{w_x}{2} + x, y_0 - \frac{w_y}{2} + y, t_0 - \frac{w_t}{2} + t\right) \tag{4.18}$$

Where $R_j(x_0,y_0,t_0)$ is response of the video sequence, *V*, to the j^{th} 3D Gabor filter, g_j, at spatio-temporal location $\vec{x} = (x_0,y_0,t_0)$. $\mathrm{w_x}, \mathrm{w_y}, \mathrm{w_t}$ are the window sizes of the 3D Gabor kernel at *x, y* and *t* dimensions respectively.

In figure 6 slice of a 3D Gabor kernel is given. Responses of a simple moving bar is shown in figure 7, it can be seen that responses are maximized when both spatial and temporal orientations of the moving bar and 3D Gabor kernel are matched.

Responses of 42 consecutive frames from a walking sequence to a 3D Gabor filter (spatio-temporal orientation and scale same for all frames) can be seen in figure 8; as the body moves at different time instances 3D Gabor filter matches different parts of the body.

In figure 8 responses of a frame from a walking sequence is given for two 3D Gabor kernels with different spatio-temporal orientations. It can be seen that salient locations of the motion are highlighted by 3D Gabor responses.

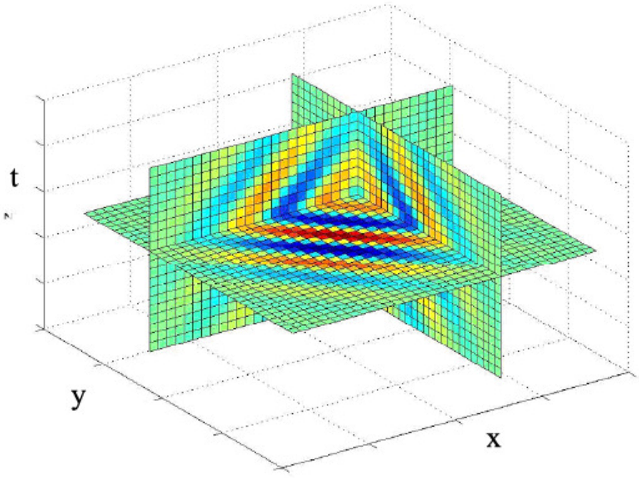


Figure 6: Slice of a 3D Gabor filter

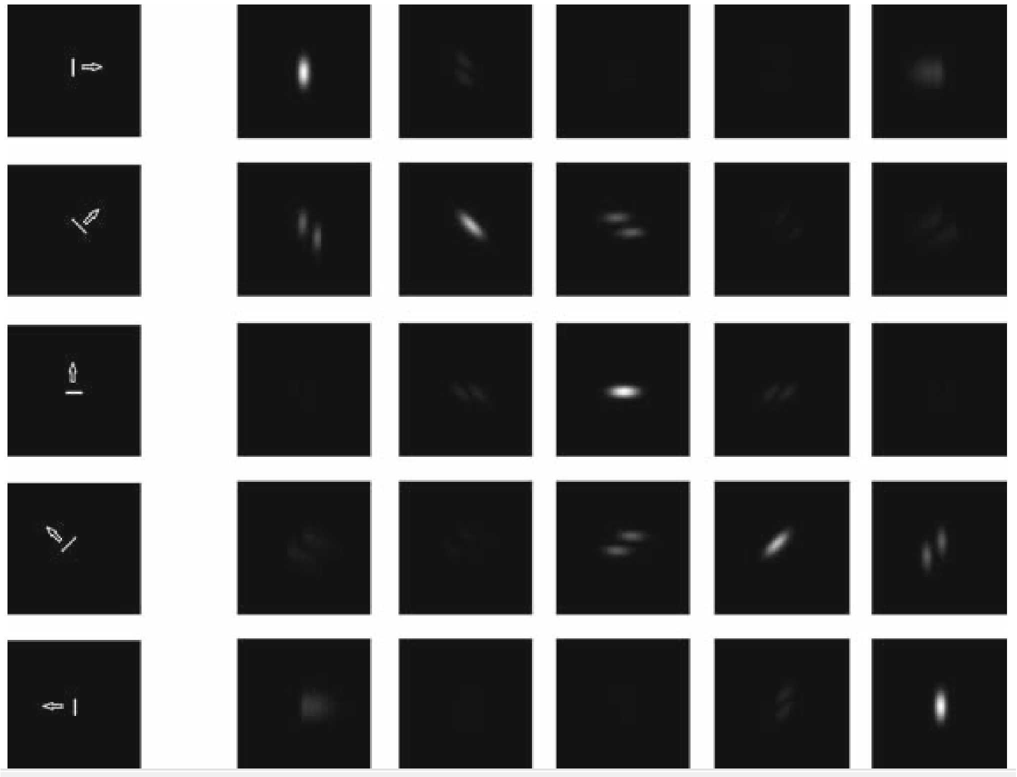


Figure 7: Responses of a moving bar to 3D Gabor filters with different orientations.

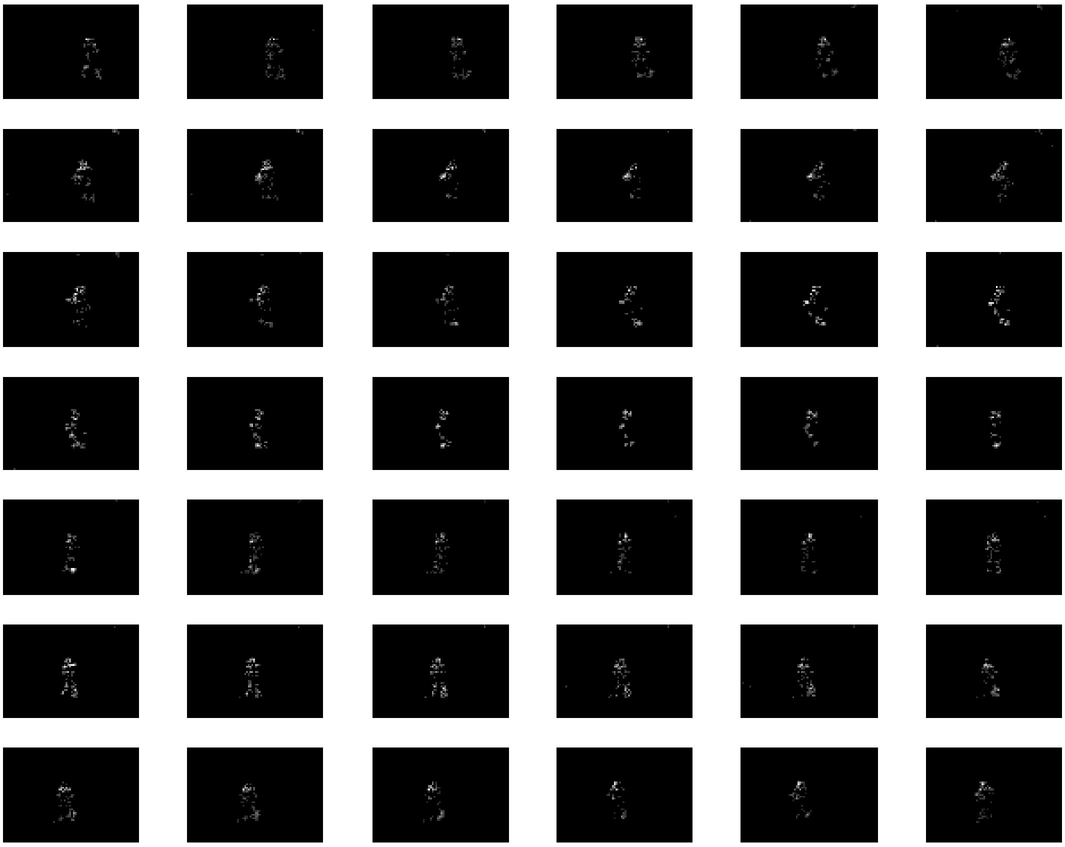


Figure 8: Responses to a 3D Gabor filter frames from t=1 to 42

4.2.2. Spatio-Temporal Feature Selection

In this work a two step feature selection algorithm is applied to find the representative features of the activity: in the first step interest points are found by filtering 3D Gabor responses, in the second step by applying clustering, discriminative spatio-temporal features are found as the symbols of the underlying action. Locality of the features is assured in both spatial and temporal domains.

Since proposed feature selection algorithm is fully automatic, numbers of final feature clusters and number of feature vectors of the final feature clusters are not set a priori. With our feature selection algorithm each class of activity is represented according to its specific temporal texture properties effectively.

4.2.2.1. Filtering

Filtering is applied to 3D Gabor responses to obtain a local spatio-temporal feature map of each frame. 3D Gabor filters gives high response at the locations of sharp changes in curvature in spatio-temporal domain. Since such locations correspond to some low level salient features, by filtering, peaks of 3D Gabor responses selected as spatio-temporal interest points.

Local maximums of each 3D Gabor response in a window $(w_x \times w_y)$ are found as the points with high information. A feature point is located at (x_0, y_0, t_0), if

1. $$R_j(x_0, y_0, t_0) = \max_{(x,y)\in W_0} R_j(x, y, t) \tag{4.19}$$

2. $$R_j(x_0, y_0, t_0) > \frac{1}{I_x I_y} \sum_{x=1}^{I_x} \sum_{y=1}^{I_y} R_j(x, y, t) \tag{4.20}$$

where R_j is the response of the video frame to the j^{th} Gabor filter. I_x I_y is the size of the segmented region of the action at $t=t_0$, the center of the window, W_0 is at (x_0, y_0, t_0).

Feature vectors are generated at the feature points as a composition of Gabor filter response coefficients. k^{th} feature vector of i^{th} frame is defined as,

$$v_{i,k} = \{R_j(x_k, y_k, i) | j = 1, \ldots, N_g\} \quad k=1,\ldots,N_i, \tag{4.21}$$

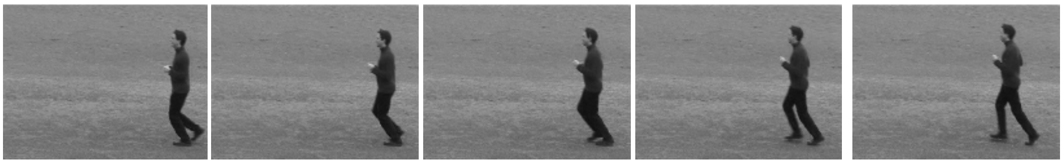
where $(x_k, y_k, t_k=i)$ shows the coordinates of a feature point and R_j are the samples of the Gabor filter responses at that point. N_g is the total number of 3D Gabor filters; in this work

96 3D Gabor filters are used with 3 different values of k_v, 4 different values of θ, and 8 different values of β (eqn 4.17). N_i is the number of feature vectors for frame i. Another parameter, d_k, associated with each feature vector is computed as the normalized distance of the feature point from the center of the body:

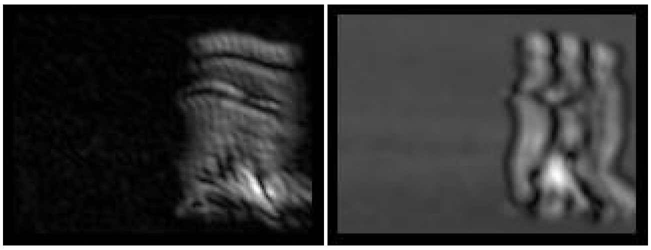
$$d_k = sign\left(y_k - \frac{I_y}{2}\right)\frac{2}{\sqrt{{I_x}^2+{I_y}^2}}\sqrt{\left(x_k - \frac{I_x}{2}\right)^2 + \left(y_k - \frac{I_y}{2}\right)^2}, \tag{4.22}$$

d_k has a sign indicating its location in y axis with respect to the center of the body; this sign enable us to differentiate between upper and lower parts of the body. Therefore arm features could not match to leg features (eqn. 4.27).

Filtering is applied to responses of a frame to each 3D Gabor filter separately (figure 10). Feature points found from each response then combined to achieve a feature point map. Feature vectors (eqn. 4.21) are then extracted at each location on the feature point map (figure 11).



(a)



(b)

Figure 9: 3D Gabor responses a) original frames, b) Response samples of the middle frame of (a) to 3D Gabor filters of different spatio-temporal orientations

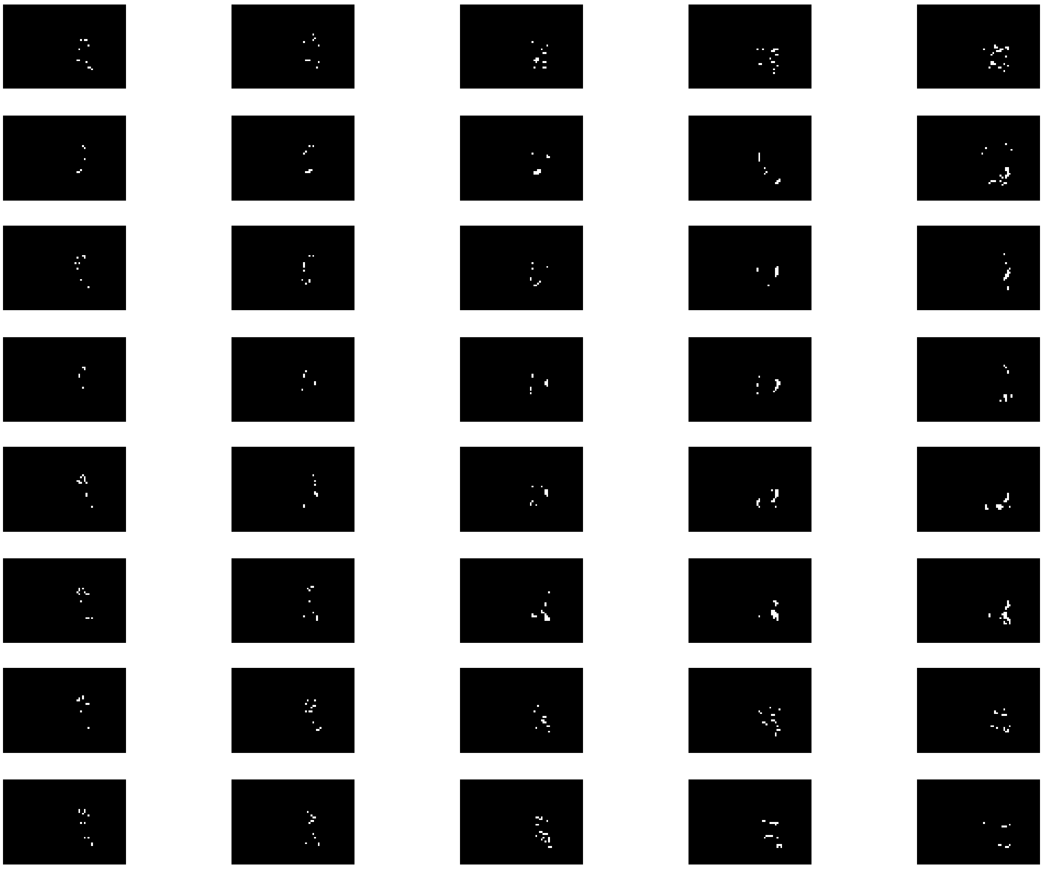


Figure 10: Exemplary thresholded responses of a frame (5 spatial frequencies, 8 spatial orientations)

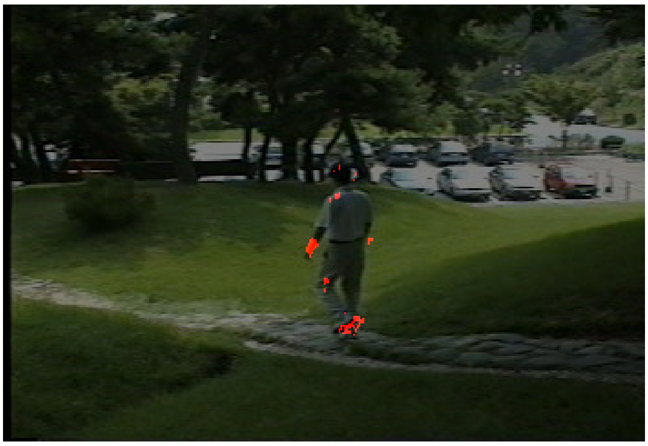


Figure 11: Spatio-temporal feature locations of a frame after filtering

4.2.2.2. Clustering

As a general definition clustering is a tool to divide a data set into regions of high similarity, as defined by a distance metric. In this work a dynamic clustering method is applied, since the number of clusters is not determined a priori.

Data to be clustered (D) is composed of the feature sets (F) of an activity sequence extracted at each frame.

$$D = \{F_t | t = 1, \dots, T\} \tag{4.23}$$

$$F_t = \{v_{t,k} \; | k = 1, \dots, N_t\} \tag{4.24}$$

N_t is the number of feature vectors at time t, T is the length of the activity sequence, v_t is a feature vector. Our aim to cluster D, and finally represent each cluster by a single feature set FC. Although filtering reduces redundancy in space for each frame, there is still a huge redundancy in time at the feature vector level; to obtain FC one must combine feature sets of each cluster by eliminating this redundancy.

$$D^c = \{FC_i \; | i = 1, \dots, T^c \ll T\} \tag{4.25}$$

D^c is the clustered data set and T^c is the number of clusters.

Clustering algorithm here must provide a tool to combine feature sets without redundancy, while maintaining the best discriminative feature vectors and eliminating the spurious ones.

Proposed clustering method utilizes Genetic Chromodynamics procedure providing dimensionality reduction by both feature selection and data partitioning.

4.2.2.2.1. Proposed GC procedure for Codebook Generation

In this work filtered feature set *(F)* of each frame is considered as an individual. Locality is maintained in both time and space. Similarity between two individuals is computed as the mean similarity of its features. Following evaluation function is used to analyze fitness.

$$dist(F_1, F_2) = 1 - Sim(F_1, F_2) \tag{4.26}$$

$$S(v_1, v_2) = \sqrt{\frac{\sum |v_1||v_2|}{\sum |v_1|^2 \sum |v_2|^2}} \, [abs(d_1 - d_2) < \tau_d] \tag{4.27}$$

$$Sim(F_1, F_2) = \text{avg}_{\forall \, v_1 \in F_1, v_2 \in F_2} \, S(v_1, v_2) \tag{4.28}$$

Each spatio-temporal feature set, F, is composed of spatio-temporal feature vectors, v, their normalized distance from the center, d, and their weights, w. τ_d is the distance threshold.

Weight indicating appearance frequency of a feature vector is set to N at the beginning and decreased by 1 at each similarity calculation if there is no match for that feature vector.

Initial population is composed of every feature set F_t (eqn. 4.24)

$$P_0 = \{F_t | t = 1, \dots, T\} \tag{4.29}$$

Mating region, *MT*, is defined by temporal distance and feature similarity:

$$MT(F_t) = \{F_{t+\varepsilon} \mid \varepsilon = \pm 1 \ \& \ Sim(F_t, F_{t+\varepsilon}) > \tau_{MT}\} \tag{4.30}$$

If $MT(F_t)$ is not an empty set, then the member, $F_{t+\varepsilon}$, with the lowest similarity is selected as the mate. Selection of the least similar mate avoids local overfitness. Once a mate is selected for F_t crossover is applied to generate an offspring. A match for each feature vector of F_t is searched in $F_{t+\varepsilon}$ and the one with the higher response magnitude is selected for the offspring $(\acute{F}_t)$. Then $\acute{F}_t$ and F_t is evaluated; fittest one is survive for the next generation. The following evaluating function is used to find the fitness.

$$Eval_{MT}(\hat{F}) = exp\left[-\sum_{F_j \notin MT(F_t)} dist\left(\hat{F}, F(j)\right) + \sum_{F_j \in MT(F_t)} dist\left(\hat{F}, F(j)\right)\right], \tag{4.31}$$

Evaluation function in (eqn. 4.31) supports increased similarity in the mating region and increased distance to the rest of the population.

If $MR(F_t)$ is an empty set, then F_t is mutated. In our approach it is important to maintain actual feature values while extracting representative spatio-temporal feature sets, therefore mutation cannot be directly applied to feature vector values. To mutate, weight of the less frequent feature vectors of an individual (F_t), is set to 0 which indicates that they are eliminated. If weight of a feature is less than a threshold it is considered as noise. Since we compute feature values in a 3D space *(x,y,t)* if a feature is not locally consistent between consecutive frames it can be stated that it is not related to the probe motion.

By mutation some individuals can be completely eliminated. However our motive is not to assign each feature set F_t to a cluster but to find feature clusters that will effectively represents the underlying motion texture.

Above procedure is applied for every individual in the current population. Before iterating for the next generation merging is applied. Merging procedure is introduced by Genetic

Chromodynamics algorithm to remove less fit individuals, to decrease the population size. It must be noted that at the mating region only the first parent and the offspring is compete for the survivor. Without merging, at every generation, while the number of similar individuals increases as the result of mating, total number of individuals would stay the same.

Merging region of F_t is defined similarly as the mating region:

$$MG(F_t) = \{F_{t+\varepsilon} \,|{-1} \leq \varepsilon \leq 1 \ \& \ Sim(F_t, F_{t+\varepsilon}) > \tau_{MG}\} \tag{4.32}$$

To combine individuals in the merging region, first all but the best individual from the selection is removed, and then its feature vectors are compared to every feature vector in the merging region, the one that has higher response magnitude is selected if two feature vectors match. The best individual is selected according to the same evaluation function of survival (eqn. 3.43) but instead of mating region, merging region is considered.

$$Eval_{MR}(\hat{F}) = exp\left[-\sum_{F_j \notin MR(F_t)} dist\left(\hat{F}, F(j)\right) + \sum_{F_j \in MR(F_t)} dist\left(\hat{F}, F(j)\right)\right] \tag{4.33}$$

At the end of the merging a new generation is obtained. If population does not alter from the previous generation by mating and/or merging procedure is finalized, otherwise mating and merging applied for the new generation as well.

At the end of GC clustering video sequence is partitioned into motion texture units and a spatio-temporal feature cluster *(CF)* for each partition is found; each cluster is modeled by a 3D Gabor feature set.

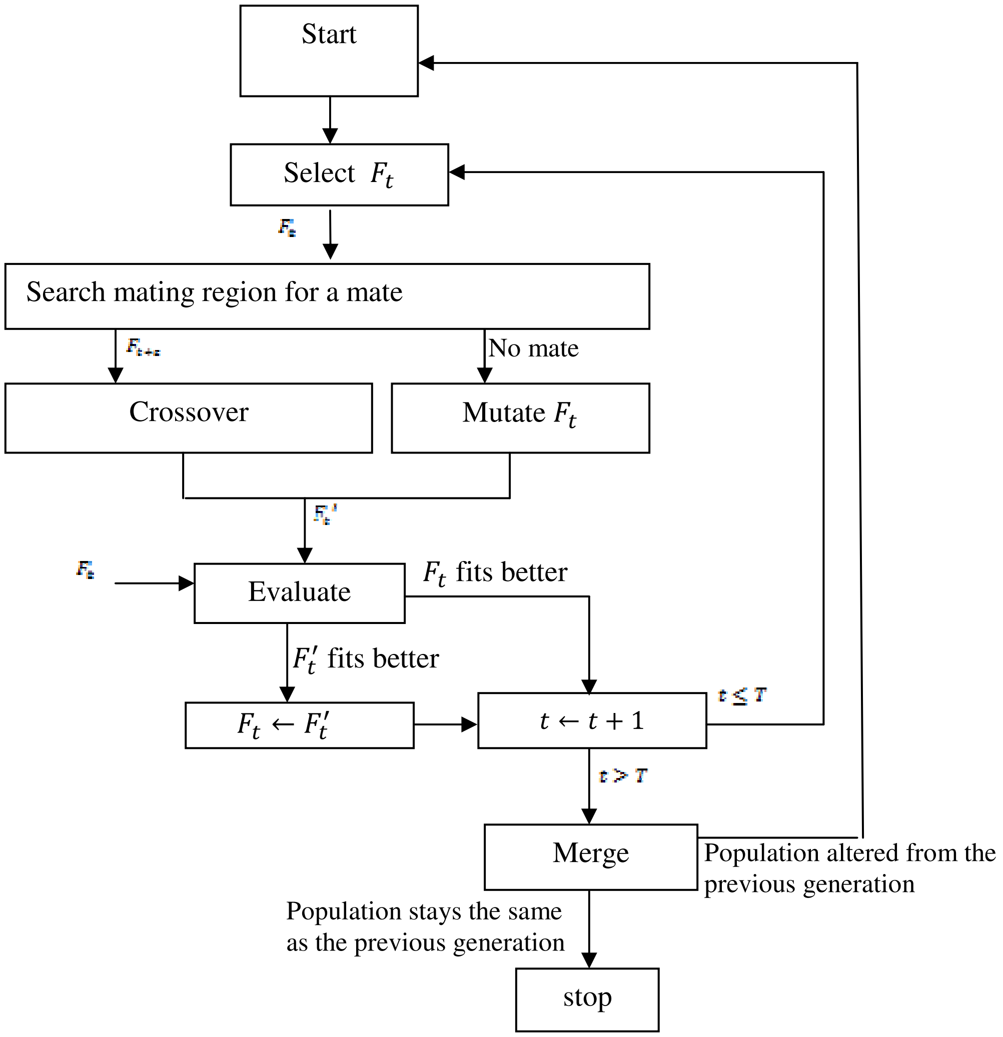


Figure 12: Genetic Chromodynamics algorithm

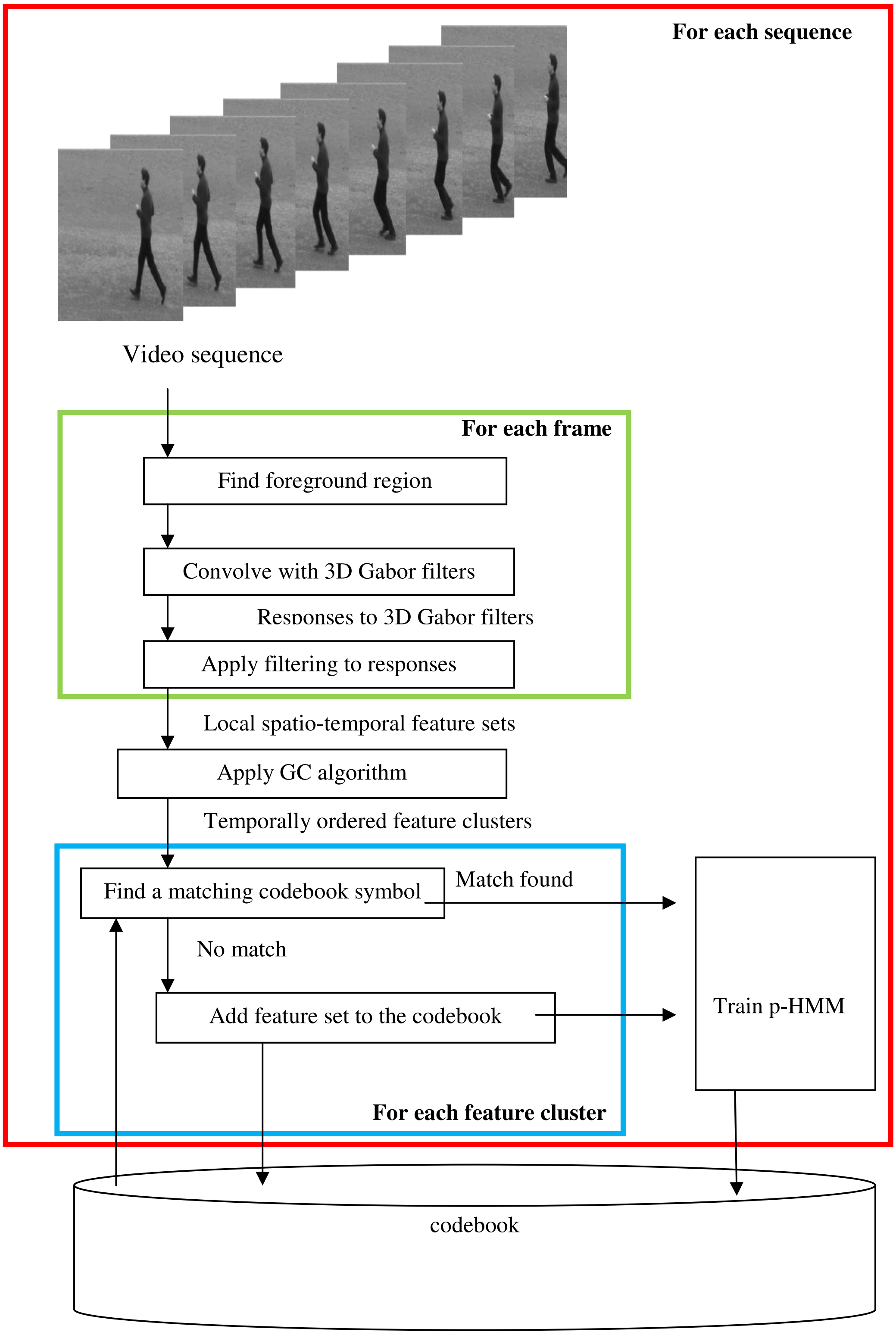

For each sequence
Video sequence
For each frame
Find foreground region
Convolve with 3D Gabor filters
Responses to 3D Gabor filters
Apply filtering to responses
Local spatio-temporal feature sets
Apply GC algorithm
Temporally ordered feature clusters
Find a matching codebook symbol
Match found
No match
Add feature set to the codebook
Train p-HMM
For each feature cluster
codebook

Figure 13: Overview of the training algorithm

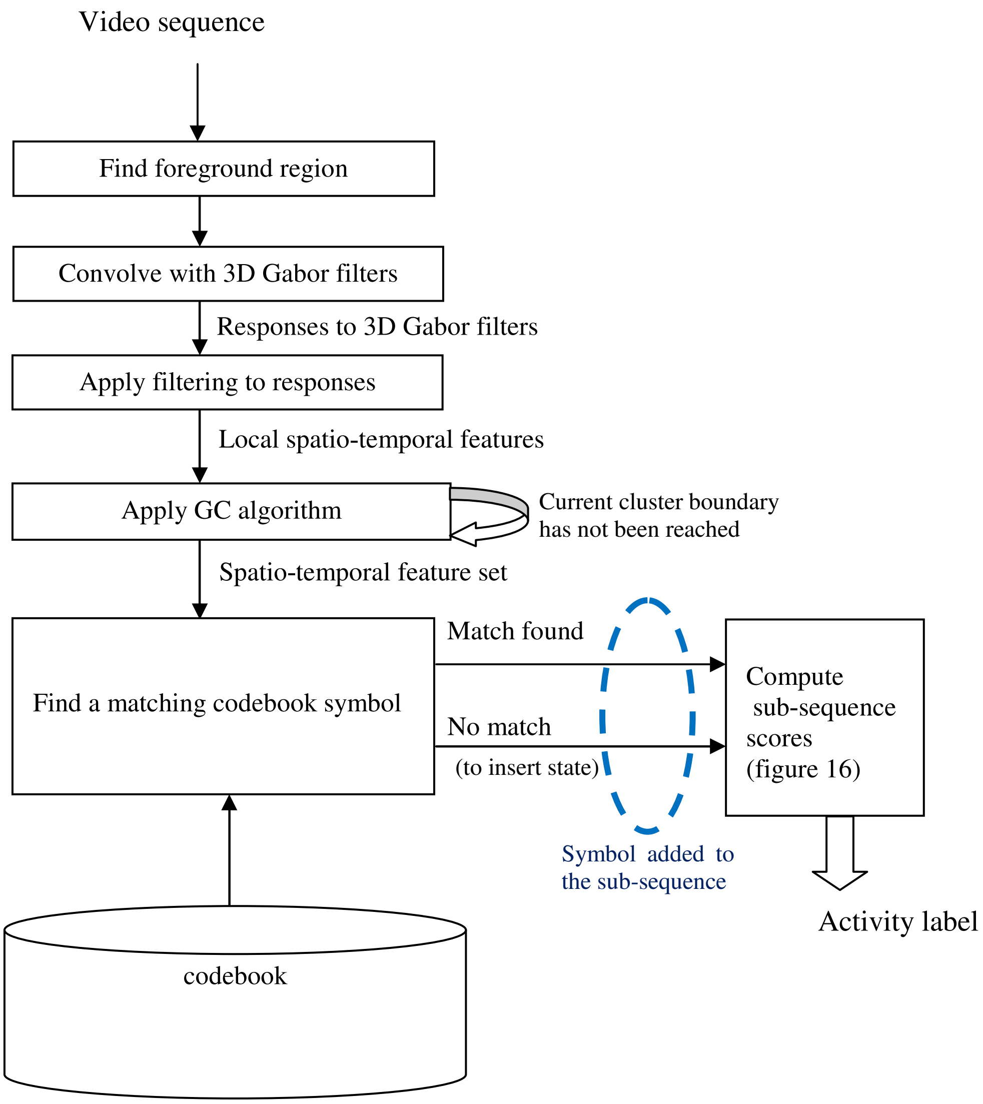

Video sequence
Find foreground region
Convolve with 3D Gabor filters
Responses to 3D Gabor filters
Apply filtering to responses
Local spatio-temporal features
Apply GC algorithm
Current cluster boundary has not been reached
Spatio-temporal feature set
Find a matching codebook symbol
Match found
No match
(to insert state)
Symbol added to the sub-sequence
Compute sub-sequence scores (figure 16)
Activity label
codebook

Figure 14: Overview of the recognition algorithm

4.3. Human Activity Recognition with Profile HMM

In our case, the observations are the spatio-temporal features of the motion texture units. As the body parts moves, 3D Gabor filters matches at some local points during the activity; at a certain instance they produce maximum response showing the points of maximum spatio-temporal variations which characterize the motion texture. Each of such spatio-temporal features is found and clustered into local action units by our feature selection algorithm. Temporally ordered feature clusters correspond to the observation sequence.

$$O = \{CF_1, CF_2, \ldots, CF_N\} \tag{4.34}$$

where N is the number of feature clusters of the sequence. Codebook is obtained as the unique feature clusters (UCF).

$$V = \{V_1, V_2, \ldots, V_M\} = \{UCF_1, UCF_2, \ldots, UCF_M\} \tag{4.35}$$

Feature clusters of the observation sequence are associated to the codebook symbols as:

$$CF_i = V_{j^*} \tag{4.36}$$

$$j^* = argmax_j(Sim(CF_i, UCF_j)) \tag{4.37}$$

Once we have the initial codebook next step is to choose the best HMM structure for our problem.

4.3.1. Adapting Profile HMM Procedure for Action Modeling

Our objective is to build a model representing the consensus sequence for an action not the sequence of any particular member. It must be also noted that the initial observation sequences are considered as unaligned.

Another important issue is; observation sequence can be derived beginning from any instance of the action and can also be ended at an arbitrary instance. That is videos are not labeled and/or segmented according to the start and finish of the activity. Assume we have two observation sequences of the walking action; first sequence can start with a symbol of left foot rising action unit while other starting with a symbol of right foot forward action unit. Moreover some observation sequence can be partial (i.e. half of the walking cycle); while other including repeated subsections (i.e. more than one walking cycle). To handle such cases rather than a standard profile HMM, a profile HMM for repeat matches to subsections of the profile model is proposed in this work (figure 15). Proposed profile HMM structure allows local alignment. Our model incorporates the original profile HMM together with a self-looping model. It behaves very like the insert state that is added to the profile itself. It is called flanking model state [96], because it is used to model the flanking sequences to the actual profile match itself.

The flanking state is shown as shaded diamond. The self transition probability of the flanking is account for long stretches of sequence. There are also silent states, shown as shaded circles, as "switching points" to reduce the total number transitions.

Sequence alignment is achieved by applying Profile HMM Viterbi Algorithm to find the optimal state path with current model parameters. Sequence alignment is embedded into the Viterbi Count Learning procedure. In figure 15 overview of the proposed training algorithm is given.

Another obstacle is the nature of our codebook. Since the words of the codebook are the spatio-temporal features extracted from available action samples, at the beginning of the training codebook is neither universal nor complete. During the training, codebook must also be refined to effectively model the probe action.

In the proposed algorithm there are two kinds of symbols; codebook symbols and unknown symbols. Unknown symbols appear at the stage of learning, as the observation samples that do not match to any codebook sample. If an unknown symbol is observed it is added to the codebook. Here observation symbols are compared to codebook symbols according to the average similarity (eqn. 4.37) of their vectors.

In our approach initially the number of *matching states* is set to the number of clusters in the first sample. When a matching state removed from the model as the result of model surgery, if there exists any symbol exclusive to that matching state it is removed from codebook as well.

Proposed structure can be characterized by the following:

1) N, the number of matching states in the model; initially equals to the number of feature clusters in the first sequence. M, matching states, I insertion states, D deletion states, F flanking state.

   $$S = \{M_1, M_2, \dots, M_N, I_1, I_1, \dots, I_{N-1},\ D_1, D_2, \dots D_{N-2}, F\} \tag{4.38}$$

2) M, the number of distinct observation symbols per state; unique feature clusters *(UCF)*. (eqn. 3.35)

3) The state transition probability distribution

$$A = \left\{a_{M_iM_j}, a_{M_iI_j}, a_{M_iD_j}, a_{I_iM_j}, a_{I_iI_j}, a_{I_iD_j}, a_{D_iM_j}, a_{D_iI_j}, a_{D_iD_j}, a_{FF}, a_{FM_j},\right\}, \quad (4.39)$$

where

$$a_{S_iS_j} = P\left[q_{t+1} = S_j | q_t = S_i\right] \qquad S = \{M, I, D\} \qquad 1 \le i, j \le N \quad (4.40)$$

4) The observation probability distribution in state j, $B = \left\{b_j(k)\right\}$, where

$$b_j(k) = P\left[V_k \ at \ t | q_t = S_j\right] \ \ S = \{M, I\} \ \ 1 \le j \le N, 1 \le k \le M \quad (4.41)$$

5) The initial state distribution $\pi = \{\pi_i\}$ where

$$\pi_i = P\{q_1 = M_i\} \ \ 1 \le i \le N \quad (4.42)$$

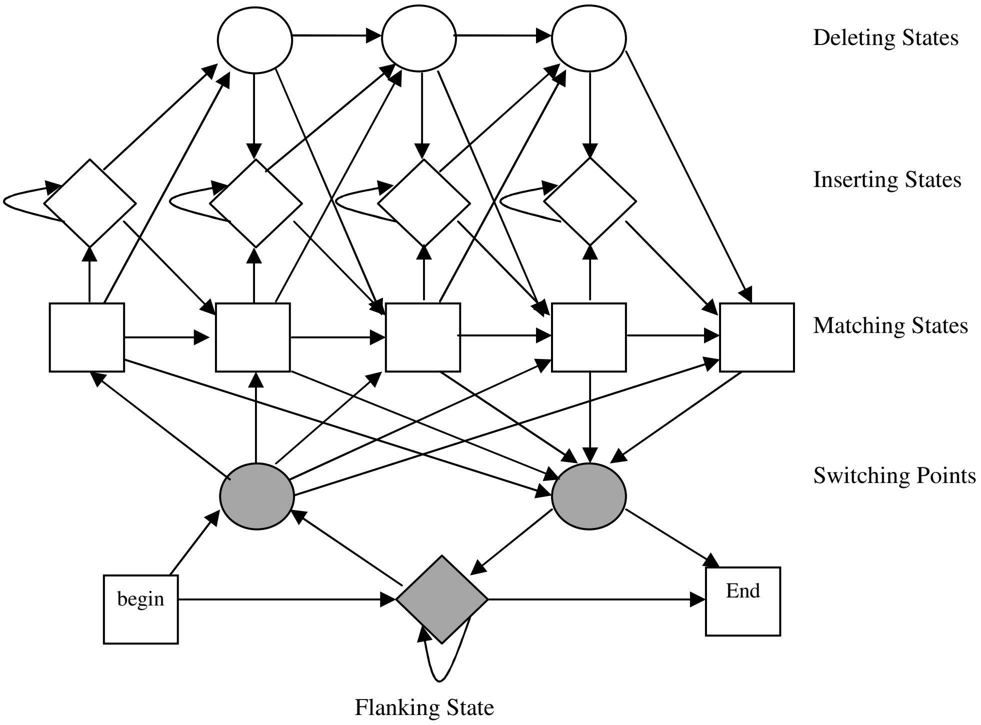


Figure 15: A profile HMM for repeat matches to subsections of the profile model

4.3.2. Parameter Learning

Viterbi Counting Algorithm is especially suitable for profile HMM structure since parameter learning is accordance to the best path. Moreover the Viterbi recurrence can be greatly simplified and expressed in terms of log odds for the case of Profile HMMs. The log odds formulation avoids underflow and reduces length effects.

Viterbi Path Counting method simply counts states and transitions and estimates the model parameters from those total counts. Best path is found by Viterbi algorithm. Training sequences iteratively inserted into the model structure using Viterbi algorithm to select the path to be incremented. The dependence upon the initial value is minimized with this method. The flow of parameter learning algorithm is as follows:

1. Set number of matching states equal to the number of symbols in the initial codebook; initial codebook is composed of symbols from first sequence only.
2. Create profile HMM with parameters A,B,π. All parameters are set to small initial values in this work to ensure each path with limited possibility.

   $$a_{FF} = 1 - \eta \tag{4.43}$$

   $$a_{FM_j} = \frac{\eta}{N} \tag{4.44}$$

   $$a_{M_iM_j} = \psi \quad a_{M_iS_j} = \varepsilon, \pi_i = \omega, \tag{4.45}$$

   $(\omega, \psi, \varepsilon)$ are small values, $\psi \gg \varepsilon$ and $S = \{I, D\}$
3. Create counter matrices and set values according to the symbol occurrences in the first sequence.
4. Set current estimates $\acute{A}, \acute{B}, \acute{\pi}$ to normalized values using counter matrices.
5. Repeat 6 as many times as desired to achieve a good training.
6. Repeat 7-11 for every observation sequence in the training set.
7. If exists add unknown symbols to the codebook, set emission value of the new codebook symbol to a small value for every state.
8. Find best path for the current observation sequence via Viterbi algorithm.
9. Update counter matrices according to the best path; loop the best path add 1 to A_c for all transitions and add 1 to B_c for all emitted symbols appear in the current state.
10. Set current estimates $\acute{A}, \acute{B}, \acute{\pi}$ to normalized values using counter matrices.
11. Apply model surgery if needed.

While updating counter matrixes, and model parameters, there could be paths that are not visited by any of the training sequence, which are resulted zero transition and emission probabilities. This is not allowed in proposed structure, whenever a counter matrix element becomes zero a small value is assigned to assure non-zero transition and/or emission probabilities.

Table 1: Notations

Notations	
$v_{M_j}(t)$	the probability of the best path matching the subsequence $x_{1,\dots,t}$ to the profile submodel up to the column j, ending with x_t being emitted by state *M_j*.
$v_{I_j}(t)$	the probability of the best path in x_t being emitted by state I_j.
$v_{D_j}(t)$	the probability of the best path ending in state D_j. (x_t is the last symbol emitted before D_j)
$V_{M_j}(t), V_{I_j}(t), V_{D_j}(t)$	the log-odds scores corresponding respectively to $v_{M_j}(t), v_{I_j}(t), v_{D_j}(t)$

Profile HMM Viterbi Algorithm To Find The Optimal State Path

Initialization: $V_{M_0}(0) = 0$

Recursion:

$$V_{M_j}(i) = \log e_{M_j}(x_i) + max \begin{cases} V_{M_{j-1}}(i-1) + \log a_{M_{j-1}M_j} \\ V_{I_{j-1}}(i-1) + \log a_{I_{j-1}M_j} \\ V_{D_{j-1}}(i-1) + \log a_{D_{j-1}M_j} \\ V_F(i-1) + \log a_{FM_j} \end{cases} \tag{4.46}$$

$$V_{I_j}(i) = \log e_{I_j}(x_i) + max \begin{cases} V_{M_{j-1}}(i-1) + \log a_{M_{j-1}I_j} \\ V_{I_{j-1}}(i-1) + \log a_{I_{j-1}I_j} \\ V_{D_{j-1}}(i-1) + \log a_{D_{j-1}I_j} \end{cases} \tag{4.47}$$

$$V_{D_j}(i) = max \begin{cases} V_{M_{j-1}}(i-1) + \log a_{M_{j-1}D_j} \\ V_{I_{j-1}}(i-1) + \log a_{I_{j-1}D_j} \\ V_{D_{j-1}}(i-1) + \log a_{D_{j-1}D_j} \end{cases} \tag{4.48}$$

$$V_F(i) = max \begin{cases} V_{M_{j-1}}(i-1) + \log a_{M_{j-1}F} \\ V_F(i-1) + \log a_{FF} \end{cases} \tag{4.48}$$

Termination:

Final score is $V_{M_{L+1}}(n)$, calculated using the top recurtion relation.

Track back the optimal state sequence.

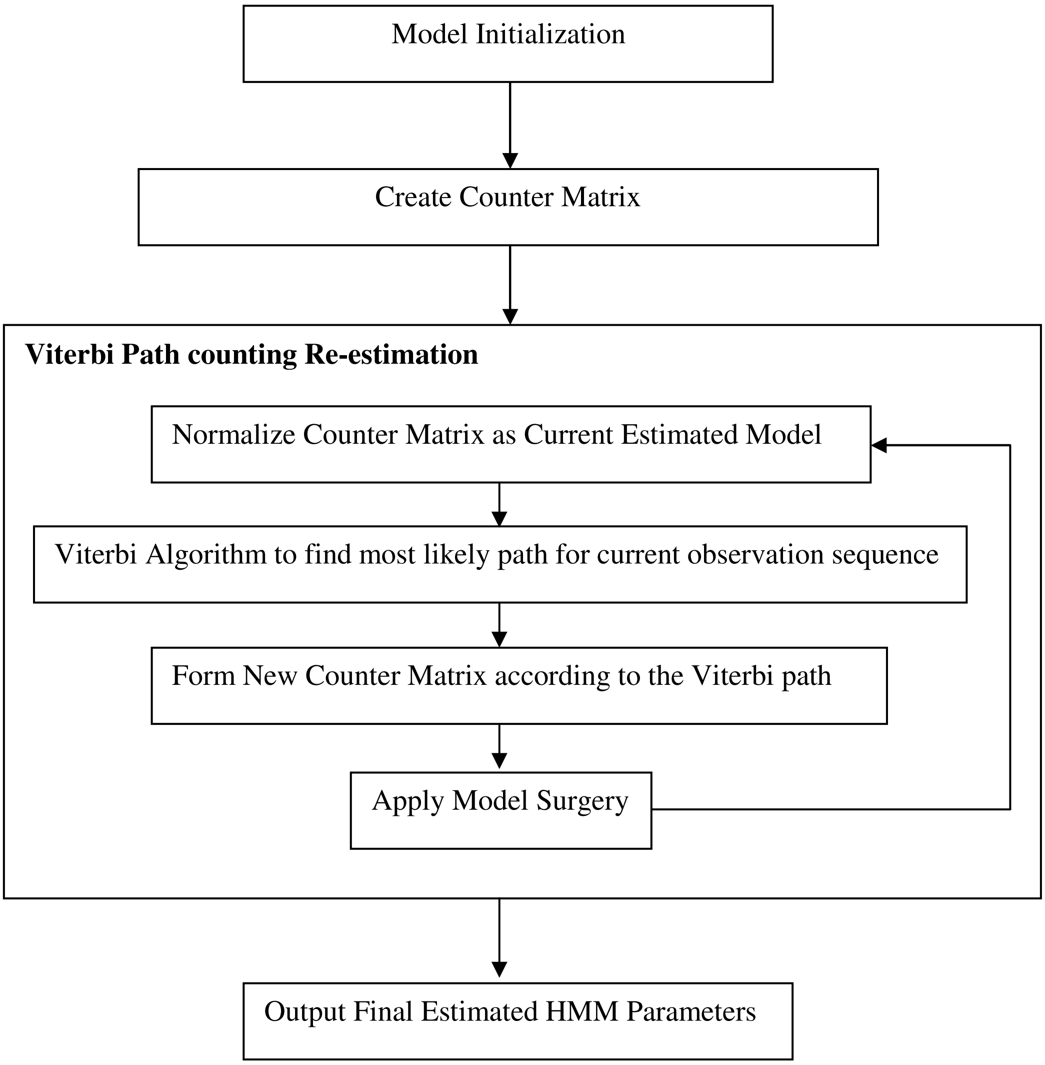

Model Initialization
Create Counter Matrix
Viterbi Path counting Re-estimation
Normalize Counter Matrix as Current Estimated Model
Viterbi Algorithm to find most likely path for current observation sequence
Form New Counter Matrix according to the Viterbi path
Apply Model Surgery
Output Final Estimated HMM Parameters

Figure 16: Viterbi Count Learning procedure

4.3.3. Sequence Scoring and Action Recognition

Pattern recognition with profile HMM is determining whether a new sequence contains the motif. Standard profile HMM forward algorithm (3.38-3.42) is applied to calculate probability of the observed sequence generated model $P(O|\lambda^r)$. Although this gives a score the states that generate the score must also be analyzed. Profile HMM structure has *insertion* and *deletion* states as explained before. However underlying motif of a human action is captured by *matching* states. If no symbols were emitted by *matching* states, then the motif is not present in the observation sequence.

The most likely path is found by using the Viterbi algorithm; to trace the location of the motif corresponds to the symbols emitted by the match states.

Each human action class is modeled by a distinct HMM (λ^r). For each unknown action sample to be recognized, measurement of observation sequence $O=\{O_1,O_2,\ldots,O_T\}$ is carried out via feature selection. To recognize an action sample its final scores are calculated for all possible models, and the action whose final score is highest is selected as label.

$$N_M(\lambda^r)=(1/N)\Sigma\,\delta(M_i\in \text{Viterbi path}) \quad i=1,\ldots,N \tag{4.49}$$

$$\text{Final score}(\lambda^r)=N_M(\lambda^r)\,P(O|\lambda^r) \tag{4.50}$$

$$r^*=\text{argmax}[\text{Final score}(\lambda^r)]\ r=1,\ldots,R \tag{4.51}$$

where R is the number of learned motions.

Proposed profile HMM structure allows subsequence recognition as well. Moreover our feature selection algorithm finally clusters spatio-temporal features in time domain to ensure temporal locality. Once the cluster boundary is reached, our algorithm generates an observation symbol, which forms a subsequence of observation symbols conjointly with the previously generated observation symbols, if any. Although observing the complete sequence will generate better results, it is still useful to match subsequences in order to achieve an early decision.

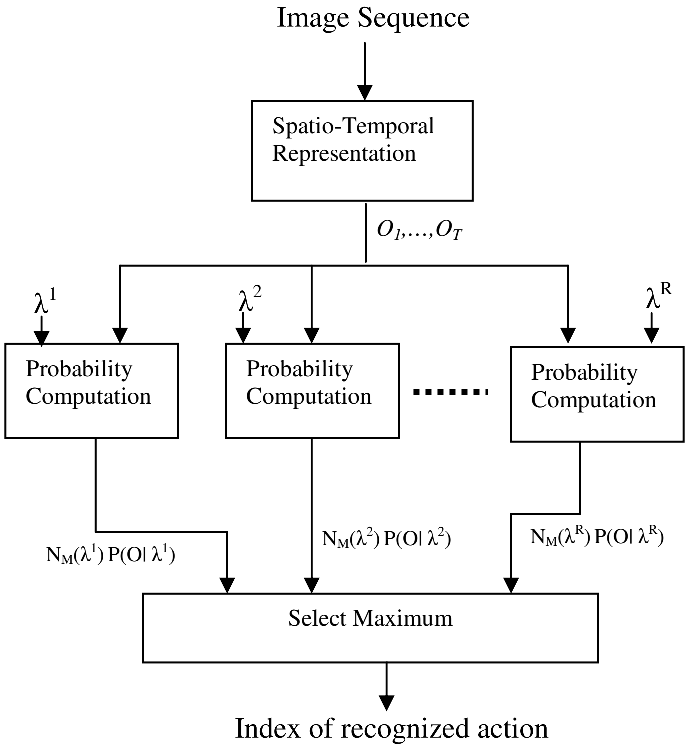


Image Sequence
Spatio-Temporal Representation
$O_1,\ldots,O_T$
λ^1
λ^2
λ^R
Probability Computation
Probability Computation
Probability Computation
$N_M(\lambda^1)\,P(O|\lambda^1)$
$N_M(\lambda^2)\,P(O|\lambda^2)$
$N_M(\lambda^R)\,P(O|\lambda^R)$
Select Maximum
Index of recognized action

Figure 17: Proposed profile HMM recognition algorithm

CHAPTER 5

RESULTS

We evaluate the proposed method with two publicly available human action databases, namely Weizmann human action database [158] and KTH human action database [157].

5.1. KHT Action Dataset

The KTH actions dataset is the most commonly used dataset in evaluating human action recognition. The KTH actions dataset consists of six human action classes: walking, jogging, running, boxing, waving, and clapping (figure 18). Each action class is performed several times by 25 subjects. The sequences were recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. Each sequence is further divided into shorter "clips" for a total of 2391 sequences. We use the original training and test split so our results are directly comparable to the recent survey by Wang *et al.* [97]; divide the samples into test set (9 subjects: 2, 3, 5, 6, 7, 8, 9, 10, and 22) and training set (the remaining 16 subjects).

5.1.1. Comparative Results

We compare our method to the state of the art methods. The comparison of results between ours with others is listed in Table 3. Our confusion matrix on KTH database is given in figure 19. In figure 20-23 confusion matrixes of other methods can be seen. Proposed algorithm is achieved %96.98 average accuracy on KTH database. This is, to the best of our knowledge, the best performance on KTH amongst methods that uses spatio-temporal features. It must be also noted that for certain actions in KTH such as running vs. jogging, even humans have difficulties in distinguishing them.

Table 2: Number of sequences in KTH dataset

Activity	**Number of training sequences**	**Number of test sequences**
Boxing	254	143
Handclapping	252	144
Hand waving	254	144
Jogging	256	144
Running	256	144
Walking	256	144
Total	**1528**	**863**

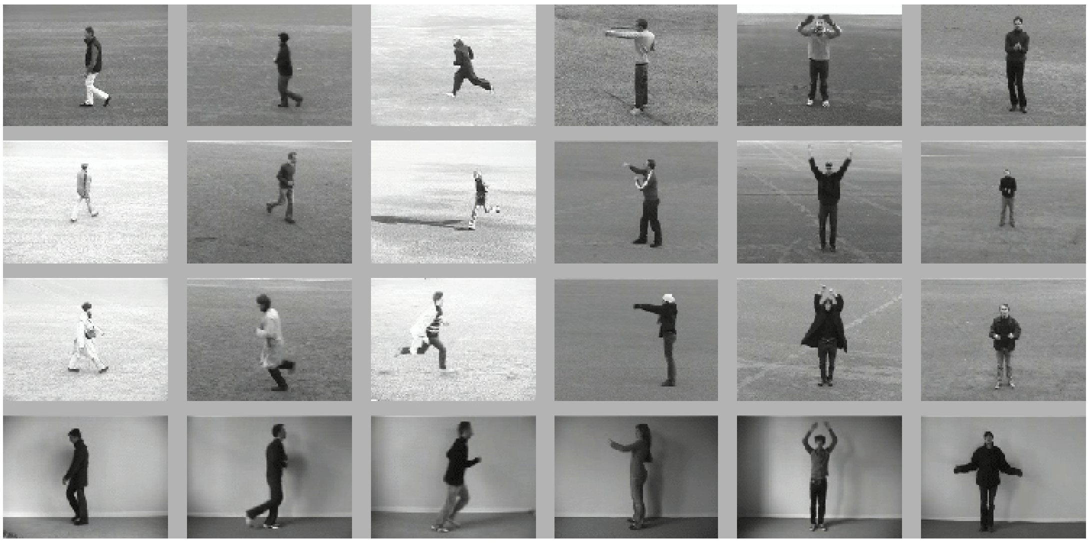
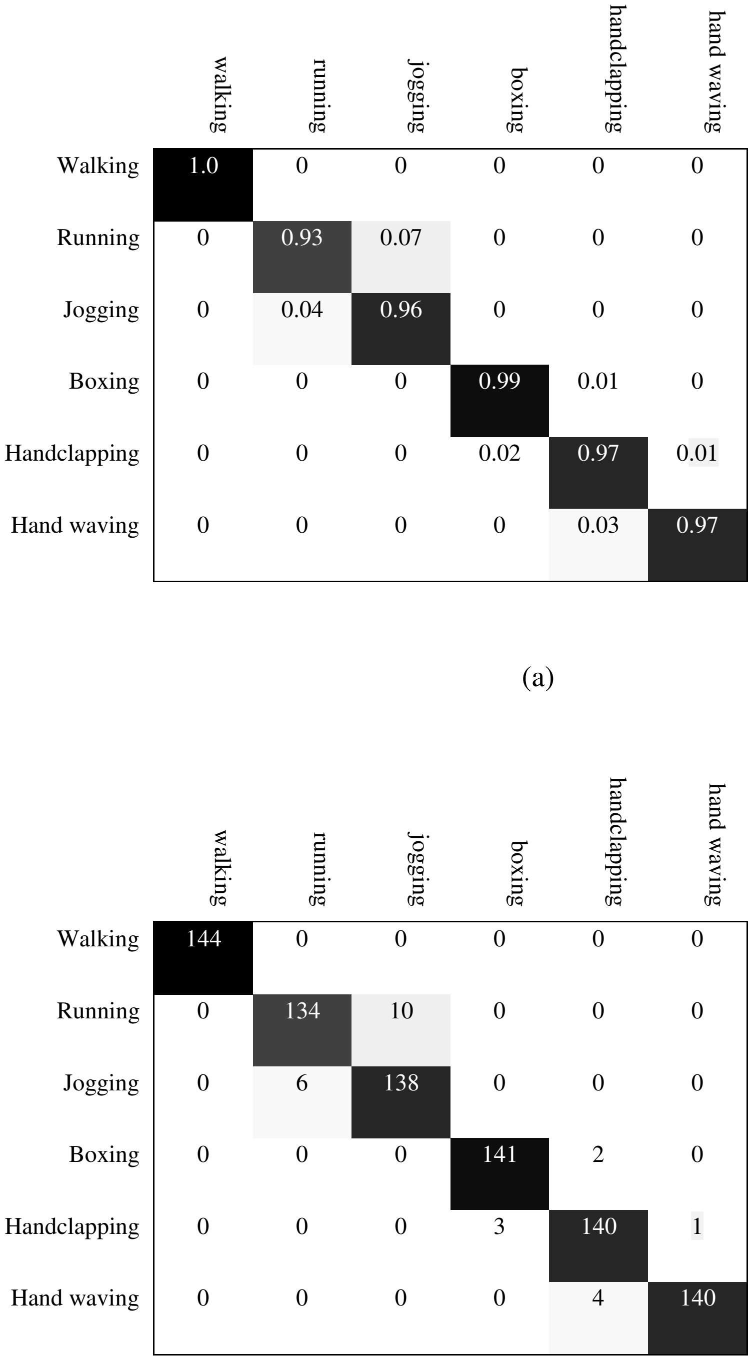


Figure 18: Samples from KHT database



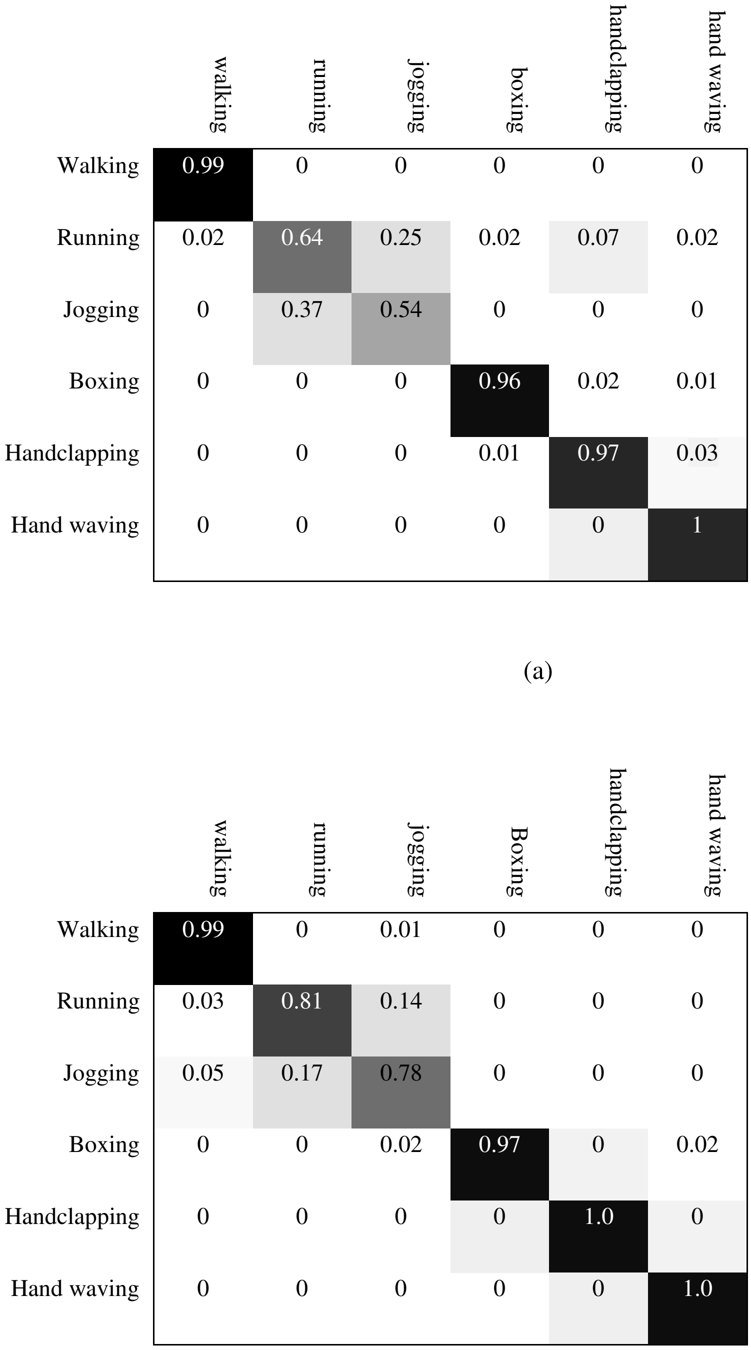
	walking	running	jogging	boxing	handclapping	hand waving
Walking	1.0	0	0	0	0	0
Running	0	0.93	0.07	0	0	0
Jogging	0	0.04	0.96	0	0	0
Boxing	0	0	0	0.99	0.01	0
Handclapping	0	0	0	0.02	0.97	0.01
Hand waving	0	0	0	0	0.03	0.97

(a)

	walking	running	jogging	boxing	handclapping	hand waving
Walking	144	0	0	0	0	0
Running	0	134	10	0	0	0
Jogging	0	6	138	0	0	0
Boxing	0	0	0	141	2	0
Handclapping	0	0	0	3	140	1
Hand waving	0	0	0	0	4	140

(b)

Figure 19: Confusion matrix on KHT database a) performance, b) number of frames



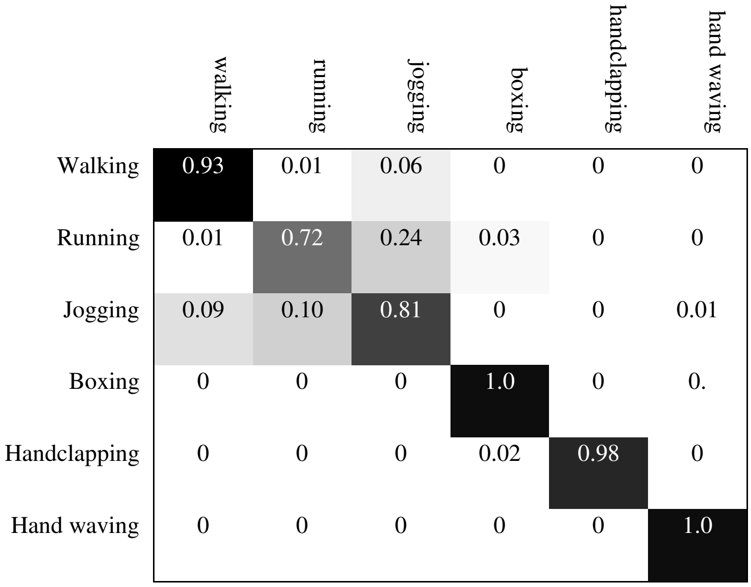
	walking	running	jogging	boxing	handclapping	hand waving
Walking	0.99	0	0	0	0	0
Running	0.02	0.64	0.25	0.02	0.07	0.02
Jogging	0	0.37	0.54	0	0	0
Boxing	0	0	0	0.96	0.02	0.01
Handclapping	0	0	0	0.01	0.97	0.03
Hand waving	0	0	0	0	0	1

(a)

	walking	running	jogging	Boxing	handclapping	hand waving
Walking	0.99	0	0.01	0	0	0
Running	0.03	0.81	0.14	0	0	0
Jogging	0.05	0.17	0.78	0	0	0
Boxing	0	0	0.02	0.97	0	0.02
Handclapping	0	0	0	0	1.0	0
Hand waving	0	0	0	0	0	1.0

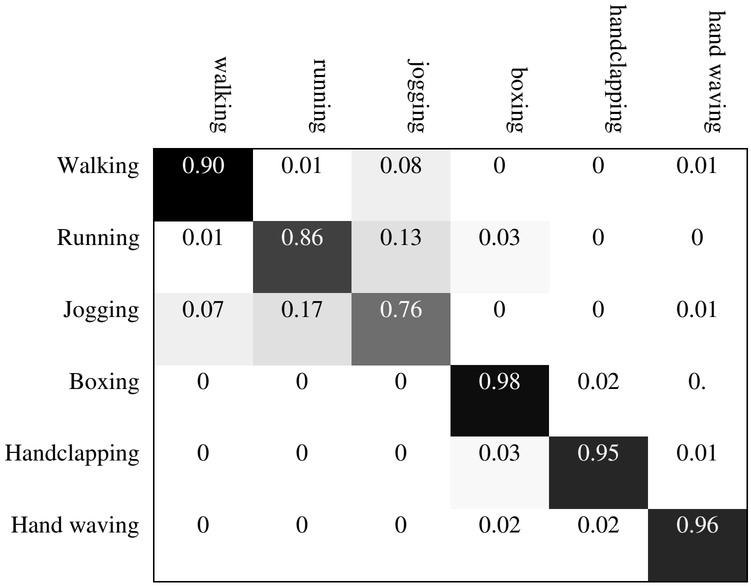
(b)

Figure 20: Confusion matrix on KHT database a) s-LDA [146], b) MMHCRF [134]



	walking	running	jogging	boxing	handclapping	hand waving
Walking	0.93	0.01	0.06	0	0	0
Running	0.01	0.72	0.24	0.03	0	0
Jogging	0.09	0.10	0.81	0	0	0.01
Boxing	0	0	0	1.0	0	0.
Handclapping	0	0	0	0.02	0.98	0
Hand waving	0	0	0	0	0	1.0

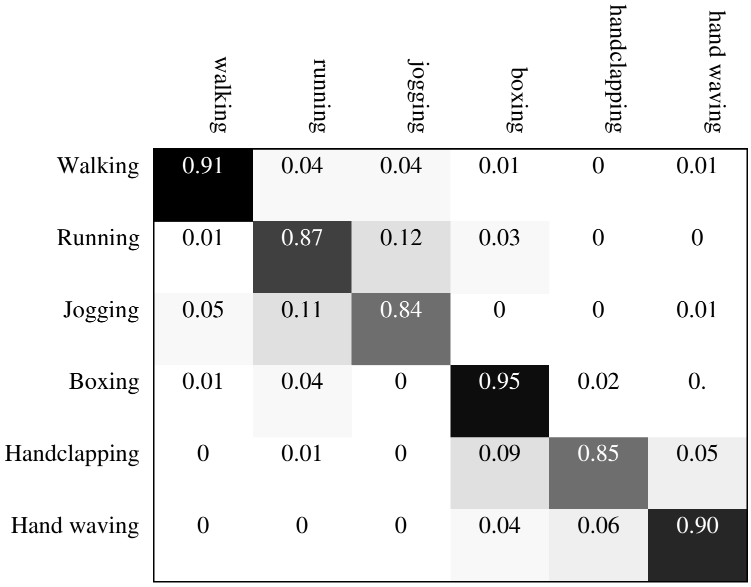
(a)



	walking	running	jogging	boxing	handclapping	hand waving
Walking	0.90	0.01	0.08	0	0	0.01
Running	0.01	0.86	0.13	0.03	0	0
Jogging	0.07	0.17	0.76	0	0	0.01
Boxing	0	0	0	0.98	0.02	0.
Handclapping	0	0	0	0.03	0.95	0.01
Hand waving	0	0	0	0.02	0.02	0.96

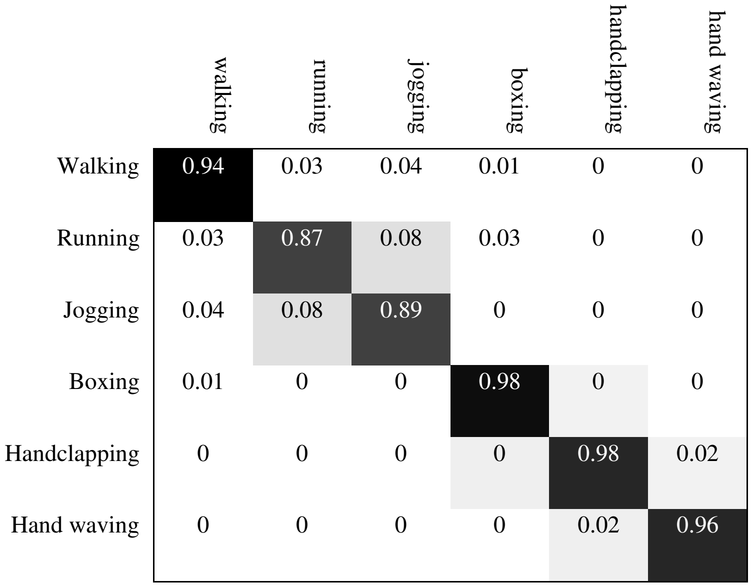
(b)

Figure 21: Confusion matrix on KHT database a) Mid level motion features [115], b) Local trinary patterns [126]



	walking	running	jogging	boxing	handclapping	hand waving
Walking	0.91	0.04	0.04	0.01	0	0.01
Running	0.01	0.87	0.12	0.03	0	0
Jogging	0.05	0.11	0.84	0	0	0.01
Boxing	0.01	0.04	0	0.95	0.02	0.
Handclapping	0	0.01	0	0.09	0.85	0.05
Hand waving	0	0	0	0.04	0.06	0.90

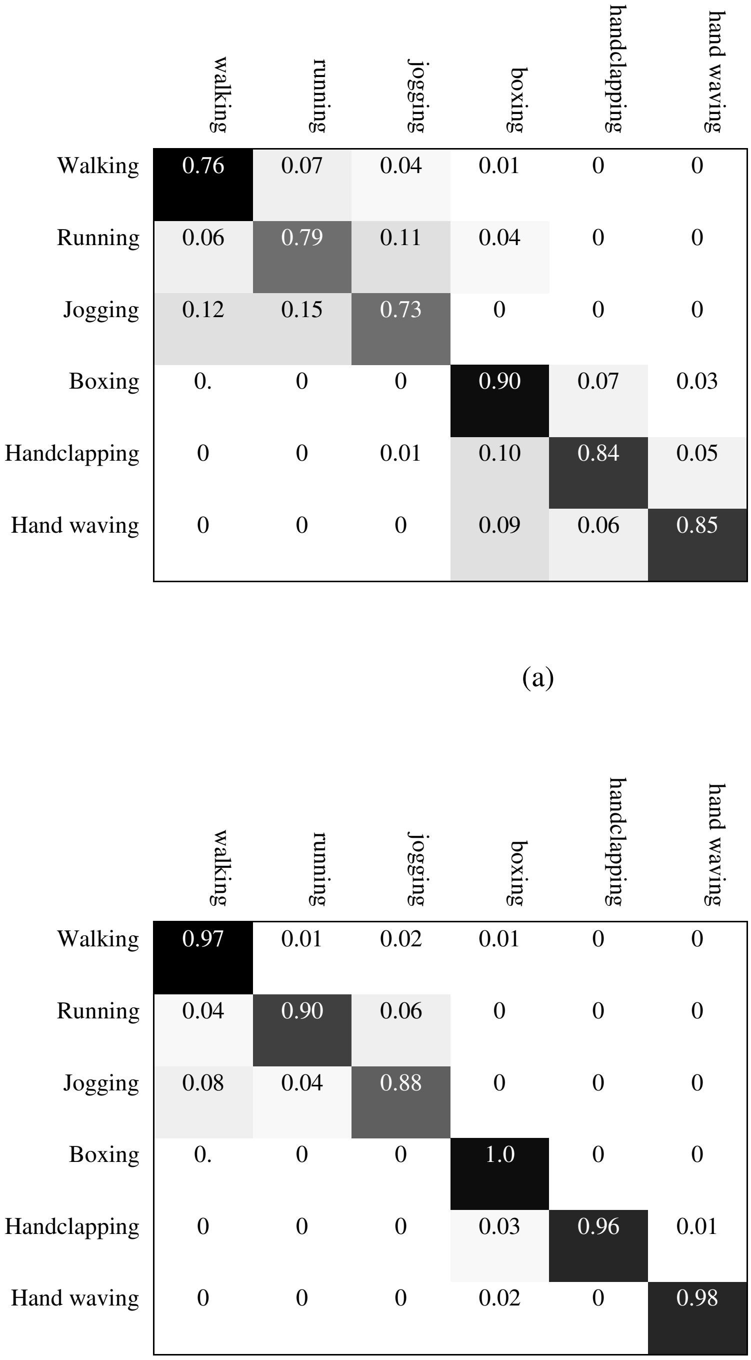
(a)



	walking	running	jogging	boxing	handclapping	hand waving
Walking	0.94	0.03	0.04	0.01	0	0
Running	0.03	0.87	0.08	0.03	0	0
Jogging	0.04	0.08	0.89	0	0	0
Boxing	0.01	0	0	0.98	0	0
Handclapping	0	0	0	0	0.98	0.02
Hand waving	0	0	0	0	0.02	0.96

(b)

Figure 22: Confusion matrix on KHT database a) Action MACH [133], b) Feature tracking [121]



	walking	running	jogging	boxing	handclapping	hand waving
Walking	0.76	0.07	0.04	0.01	0	0
Running	0.06	0.79	0.11	0.04	0	0
Jogging	0.12	0.15	0.73	0	0	0
Boxing	0.	0	0	0.90	0.07	0.03
Handclapping	0	0	0.01	0.10	0.84	0.05
Hand waving	0	0	0	0.09	0.06	0.85

(a)

	walking	running	jogging	boxing	handclapping	hand waving
Walking	0.97	0.01	0.02	0.01	0	0
Running	0.04	0.90	0.06	0	0	0
Jogging	0.08	0.04	0.88	0	0	0
Boxing	0.	0	0	1.0	0	0
Handclapping	0	0	0	0.03	0.96	0.01
Hand waving	0	0	0	0.02	0	0.98

(b)

Figure 23: Confusion matrix on KHT database a) Boosted EigenActions [117], b) Extreme learning machine [58]

Table 3: Comparison of our results with others in the literature

Method	**Accuracy**
Our method	96.98%
Chen *et al.* [100]	95.8%
Kim *et al.* [56]	95%
Liu *et al.* [117]	94.16%
Shindler [60]	92.7%
Wang *et al.* [134]	92.51%
Wang *et al.* [156]	92.1%
Laptev *et al.* [128]	91.8%
Jhuang *et al.* [98]	91.7%
Klaser *et al.* [104]	91.4 %
Niebles *et al.*[125]	91.3%
Ji *et al.* [144]	90.2%
Yeffet *et al.* [126]	90.1%
Wong *et al.* [101]	86.6%
Nieble *et al.* [137]	83.3%
Dollár *et al.* [139]	81.5%
Schuldt *et al.* [116]	71.7%
Ke *et al.* [109]	62.9%

5.2. Weizmann Action Dataset

Weizmann contains two databases, namely classification database and robustness database. The Weizmann classification database is for testing action classification performance. It contains 93 low-resolution 180 x 144 video sequences from nine people, each performing 10 natural actions: ‘‘run”, ‘‘walk”, ‘‘skip”, ‘‘jumping jack”, ‘‘jump forward on two legs”, ‘‘jump in place on two legs”, ‘‘galloping sideways”, ‘‘wave one hand”, ‘‘wave two hands” and ‘‘bend” (figure 24). All the videos are captured from a fixed viewpoint. The Weizmann robustness database is for testing the robustness of an activity recognition method. The robustness dataset has 10 videos, which were captured in front of non-uniform backgrounds, with partial occlusions and non-rigid deformations. In particular, each of the video contains a walking person with different conditions: ‘‘normal walking”, ‘‘walking with a skirt”, ‘‘walking while carrying a briefcase”, ‘‘limping”, ‘‘walking while Legs occluded by boxes”, ‘‘walking with knees Up”, ‘walking with a dog”, ‘‘sleep walking”, ‘‘walking while swinging a bag” and ‘‘walking when occluded by a pole” (figure 25).

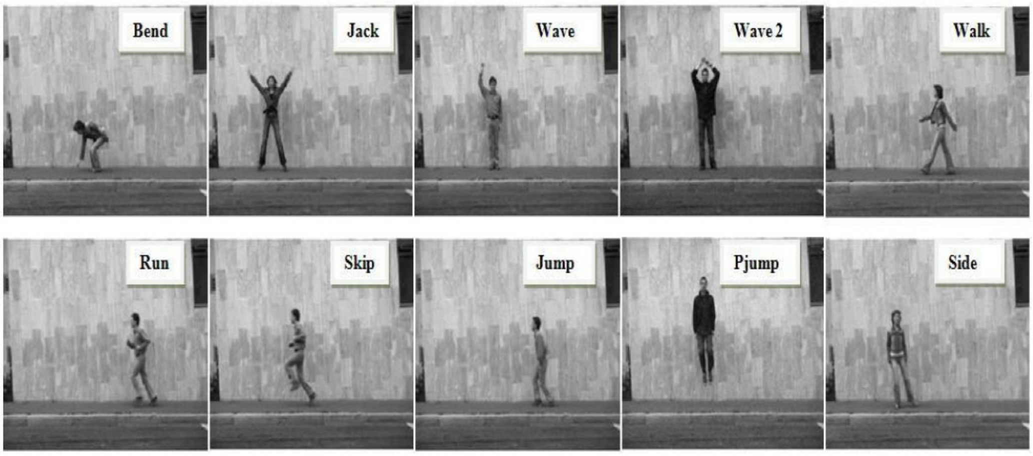


Figure 24: Weizmann dataset samples

5.2.1. Comparative Results

Recently, several approaches reported close to perfect recognition rates on this relatively easy dataset. We employ the leave-one-out scheme for evaluation, by randomly taking one person (totally 10 actions) as the testing data, and the other eight persons for training, which is a common evaluation method used by other researchers as well.

Note that existing approaches use slightly different evaluation methodologies on the data. Some evaluate on the whole sequences, others split sequences into multiple subparts, which is possible because most actions are periodic. We report here results for the full sequences, where our method yields perfect recognition rates that is 100%. In Table 4, we summarize our recognition results and compare them against other approaches.

Table 4: Comparison of our results with others in the literature on Weizmann dataset

Method	**Accuracy**
Our method	100%
Yeffet *et al.* [126]	100%
Fathi *et al.* [115]	100%
Blank *et al.* [131]	100%
Duygulu *et al.* [69]	100%
Jhuan *et al.* [98]	98.8%
Liu *et al.* [117]	98.3%
Wang *et al.* [61]	97.8%
Thurau *et al.* [118]	94.40%
Ali *et al.* [59]	92.6%
Dollar *et al.* [139]	86.7%
Niebles *et al.* [129]	72.8%

5.2.2. Robustness test

In this experiment, we demonstrate the robustness of our proposed method using Weizmann robustness testing dataset. This data set includes 10 different scenarios of “walk” activity: normal walk, walking in a skirt, carrying a briefcase, limping man, occluded legs, knees up, walking with a dog, sleep walking, swinging a bag, occluded by a pole (figure 25). Our method achieves perfect recognition rate (%100) on this database as well as other reported methods in the literature: [57] %100, [131] 100% [117] 100%. Although this result shows that our algorithm applicable on the occluded cases, number of test samples is not enough for a proper performance evaluation.

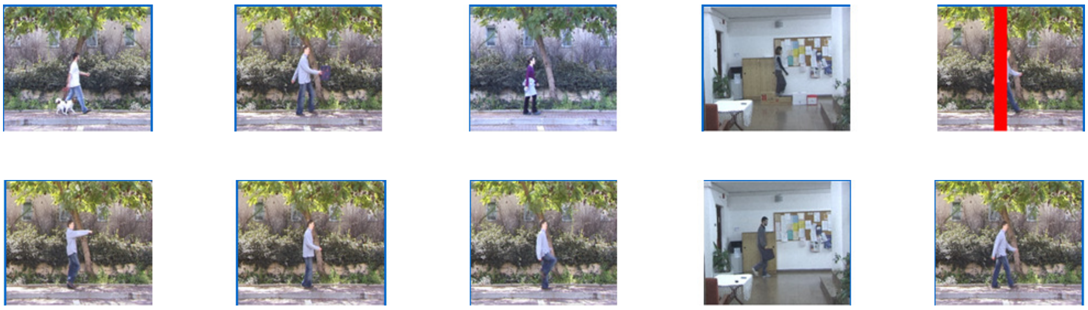


Figure 25: Samples from the robustness test set of the Weizmann dataset.

5.3. Discussion

Activity recognition accuracy of the proposed method is compared to the other methods on both KTH (table 3) and Weizmann datasets (table 4).

To further analyze evaluation results it is required to look closer to the previous approaches. Instead of giving a detailed explanation of the methods, here only the key points of the algorithms that affect their performances are reviewed.

It is seen that recent methods mostly focused on spatio-temporal representation of the activities. Although “bag of features” based approaches reports better results compared to the early activity recognition algorithms, such as silhouette based approaches or direct temporal quantization based approaches, their discriminative power over different activity classes still appear to have a limit of %80-%90 accuracy. Histogram of gradients and histogram of flow are the two commonly used feature descriptors of successful spatio temporal representations. Methods mostly differ according to their interest point detectors. Locations with high variation such as spatial and/or temporal corners, corners of motion are found as successful interest point locations [139, 101, 128, 98, 125, 156].

Main drawback of the “bag of features” representation is their holistic approach; lack information about the relations between features. Especially on relatively simple Weizmann dataset inefficiency of such holistic approaches became clear. Since “bag of features” representation does not consider temporal ordering different activities result in similar patterns.

Methods based on temporal segmentation of the motion, “bag of words” representation, report intermediate performances [156, 128]. On the other hand the best results previously reported on the KTH dataset are based on the recent approaches which are also focused on capturing global structure of activities. Although those methods still depend on gradient and/or optical flow based descriptors, their activity representation utilizes relations between spatio-temporal features. In [100] Chen *et al.* constructs video codebook by k-means algorithm and use bigram to capture sequence information based on cooccurances of adjacent words. Kim *et al.* [56] inspect joint space-time linear relationships of two video volumes.

Liu *et al.* [117] segment human action to primitive periodic motion cycles, and represent each segment as action units by applying PCA. In their work corners of the motion is detected and the video segment between those corners are modeled, in contrast in our work motion corners are detected and used to model the activity.

Wang *et al.* [156] proposed a global motion feature for the entire frame and a set of salient patches to represent local motion patterns. The hidden part of their Maximum Margin Hidden Conditional Random Field (MMHCRF) model assigns a “part label” to a local patch. Relations between features are through an undirected graph.

Although performance gain provided by temporal ordering of the spatio temporal features, have captured researchers attention, instead of modeling the temporally ordered data, studies are focused on extracting additional features by evaluating temporal relations, then recognition is done using SVM based classification, nearest neighbor classification, voting or even direct similarity calculation. Sequence alignment is considered in [100], however results were reported manually aligned sequences.

Number of interest points/feature vectors also appears to have an important impact over classification accuracy. For a successful spatio-temporal representation authors conclude that "sufficient" number of interest points must be selected, both more and less interest points reduces performance. While [56] and [117] employs AdaBoost algorithm for feature selection, in [100] randomly sampled %1 of interest points are used. Heuristic limitation of the number of interest points in a range is also a common method.

Proposed method achieved the best accuracy on both datasets due to the following attributes:

- Proposed unified local spatio-temporal description provides better representation, instead of treating spatial and temporal domain separately.
- Proposed feature selection algorithm successfully captures distinctive features of each activity class and maintains a compact representation.
- Proposed structural modeling of temporally ordered data yields effective and flexible classification of activities.
- Proposed method automatically generates action specific representation without domain specific constraints.

CHAPTER 6

CONCLUSIONS

In this work a novel spatio-temporal activity representation based on 3D Gabor filters is proposed. 3D Gabor filters are inspired from the human visual system; they give high responses at spatio-temporal locations with high variation at different scales and orientation. Spatio-temporal features which includes both local shape and dynamics information is assigned as the primitives of the global motion pattern of the activity. Feature selection applied to locate the spatio-temporal texture primitives in both spatial and temporal domains.

Proposed feature selection algorithm not only eliminates redundancy but also segments sequence by clustering features on time domain. Filtered features of successive frames are combined by a genetic algorithm, Genetic Chromodynamics, to obtain an elite set of representative local features. Genetic Chromodynamics algorithm was proposed as a multimodal optimization tool [91]. In this work Genetic Chromodynamics is used as a framework for feature clustering and adopted to spatio-temporal domain. While cross over ensures to combine features from different frames in a local temporal region without redundancy, mutation is used as a feature elimination tool. State-space representation of each sequence is achieved automatically, without any constraints on number of states or predetermined features. Number of states and number of feature vectors of each state are not fixed.

While modeling the activity with these spatio-temporal features, their temporal ordering is also considered as additional discriminative information between different classes of activities. Common pattern of each activity class is learned with a Profile HMM, which is originally proposed by Durbin *et al.* [96] to align protein sequences in molecular biology. Profile HMM provides flexibility to extract the underlying common pattern of each activity class by dynamically enhancing the HMM structure with new sample sequences of different length. One of the important contributions of this work is the incremental learning of the codebook as well as the underlying structure at the training stage of the Profile HMM. Profile HMM model is automatically specialized to each activity class. Proposed Profile HMM structure is also capable of sequence aligning.

In general activity recognition problem, beginning and ending points of the activity in an observed sequence are not labeled. This is one of the main challenges of the activity recognition schemes. Proposed structure enables to recognize sub-sequences and does not require end point constrains. This also provides robustness to the missing data and occlusions.

Proposed method is not domain specific and can be applied to wide range of activities. Moreover our method can also be applied to concatenated activities.

Experimental results show that the proposed algorithm achieved the best result on KHT dataset, among the methods in the literature. The main reason is better representation of the local discriminative primitives of the activity; proposed spatio-temporal descriptor has an important effect on accuracy. Moreover in this work locality of the spatio-temporal features is not only considered in spatial domain but in the temporal domain as well. Instead of considering the activity as a whole, temporal occurrence pattern of its spatio-temporal features is analyzed by Profile HMM. This leads to better modeling and recognition performance.

Weizmann dataset is a relatively easy dataset. Our method achieved 100% correct recognition rate on Weizmann dataset. There are also some altered samples of walking action included in the Weizmann dataset as robustness dataset. Our method again achieved 100% correct recognition rate on robustness dataset, however number of test samples is very low and this result shows the applicability of the proposed method to such altered cases, rather than a performance result.

As a future work proposed algorithm should be evaluated on different domains of actions. Sports videos are the popular application area of activity recognition algorithms; the main purpose is the video annotation. Activity recognition on sports video can also be used for athletic training. Tennis videos include well defined, highly structured player activities. Proposed algorithm will be evaluated to recognize tennis player activities. Our purpose will be both video annotation and statistical analysis of the game for player training. Since in a tennis video there are several successive activities, performance results to recognize concatenated activities can also be evaluated.

Today GPUs, which became a common PC component, provide a great low cost hardware tool to faster implementation of time consuming computer vision algorithms. Another future work could be the GPU implementation of the algorithm at least 3D Gabor filtering phase to achieve real time performance.

REFERENCES

[1] L. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, pp. 257-286, 1989.

[2] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine Recognition of Human Activities: A Survey, CirSysVideo(18), No. 11, pp. 1473-1488, November 2008.

[3] G. Johansson, Visual perception of biological motion and a model for its analysis, Perception and Psychophysics, vol. 14:2, pp. 201–211, 1973.

[4] G Johansson, Visual motion perception, Scientific American, vol. 232, pp. 76–88, 1975.

[5] W.H. Dittrich, Action categories and the perception of biological motion, Perception, vol. 22, pp. 15–22, 1993.

[6] J.F. Norman, S.M. Payton, J.R. Long, L.M. Hawkes, Aging and perception of biological motion, Psychology Aging, vol. 19, pp. 219–25, 2004.

[7] J.B. Mass, G. Johansson, G. Janson, S. Runeson, Motion perception I and II [film], Boston: Houghton Mifflin.

[8] T.J. Clarke, M.F. Bradshaw, D.T. Field, S.E. Hampson, D. Rose, The perception of emotion from body movement in point-light displays of interpersonal dialogue, Perception, vol. 34, pp. 1171–80, 2005.

[9] W.H. Dittrich, T. Troscianko, S.E.G. Lea, D. Morgan, Perception of emotion from dynamic point-light displays represented in dance, Perception, vol. 25, pp. 727–38, 1996.

[10] R.D. Walk, C.P. Homan, Emotion and dance in dynamic light displays. Bulletin of the Psychonomic Society, vol. 22, pp. 437–40, 1984.

[11] F.E. Pollick, H.M. Paterson, A. Bruderlin, A.J. Sanford, Perceiving affect from arm movement, Cognition, vol. 82B, pp. 51–61, 2001.

[12] G. Mather, K. Radford, S. West, Low level visual processing of biological motion, Proceedings of the Royal Society of London B: Biological Sciences, vol. 249, pp. 149–55, 1992.

[13] H. Lu, B. Tjan, Z. Liu, Shape recognition alters sensitivity in stereoscopic depth discrimination, Journal of Vision, vol. 6, pp. 75–86, 2006.

[14] H. Ikeda, R. Blake, K. Watanabe, Eccentric perception of biological motion is unscalably poor, Vision Research, vol. 45, pp. 1935–43, 2005.

[15] J. Pinto, M. Shiffrar, Subconfigurations of the human form in the perception of biological motion displays, Acta Psychologica, vol. 102, pp. 293–318, 1999.

[16] B.I. Bertenthal, J. Pinto, Global processing of biological motions, Psychological Science, vol. 5, pp. 221–25, 1994.

[17] B.I. Bertenthal, J. Pinto, Complementary processes in the perception and production of human movement. In Dynamic Approaches to Development: Vol. 2 Approaches, ed. LB Smith, E Thelen, pp. 209–39. Cambridge, MA: MIT Press, 1993.

[18] J.A. Beintema, K. Georg, M. Lappe, Perception of biological motion from limited lifetime stimuli, Perception and Psychophysics, vol. 63, pp. 613-624, 2006.

[19] E.D. Grossman, R. Blake, Perception of coherent motion, biological motion and formfrom- motion under dim-light conditions, Vision Resesearch, vol. 39, pp. 3721–27, 1999.

[20] Valentine, Upside-down faces: a review of the effect of inversion upon face recognition, British Journal of Psychology, vol. 79, pp. 471–91, 1998.

[21] N.F. Troje, Reference frames for orientation anisotropies in face recognition and biological-motion perception, Perception, vol. 32, pp. 201–10, 2003

[22] M. Pavlova, A. Sokolov, Orientation specificity in biological motion perception, Perception and Psychophysics, vol. 62, pp. 889–98, 2000.

[23] E. Hiris, A. Krebeck, J. Edmonds, A. Stout, What learning to see arbitrary motion tells us about biological motion perception, Journal of Experimental Psychology: Human Perception Performance, vol. 31b, pp. 1096–106, 2005.

[24] G.P. Bingham, Scaling judgments of lifted weight: lifter size and the role of the standard, Ecological Psychology, vol. 5, pp. 31–64, 1993.

[25] T.A. Stoffregen, S.B. Flynn, Visual perception of support-surface deformability from humanbody kinematics. Ecol. Psychol. vol. 6, pp. 33–64, 1994.

[26] J. Shim, L.G. Carlton, J. Kim, Estimation of lifted weight and produced effort through perception of point-light display, Perception, vol. 33, pp. 277–91, 2004.

[27] G. Mather, S. West, Recognition of animal locomotion from dynamic point-light delays, Perception, vol. 22, pp. 759–66, 1993.

[28] M. Pavlova, I. Krageloh-Mann, A. Sokolov, N. Birbaumer, Recognition of point-light biological motion displays by young children, Perception, vol. 30, pp. 925–33, 2001.

[29] J.F. Norman, S.M. Payton, J.R. Long, L.M. Hawkes, Aging and perception of biological motion, Psychology and Aging, vol. 19, pp. 219–25, 2004.

[30] J.F. Norman, H.E. Ross, L.M. Hawkes, J.R. Long, Aging and the perception of speed, Perception, vol. 32, pp. 85–96, 2003.

[31] B. Kroustallis, Biological motion: an exercise in bottom-up vs top-down processing, Journal of Mind Behav, vol. 25, pp. 57–74, 2004.

[32] I.M. Thornton, R.A. Rensink, M. Shiffrar, Active versus passive processing of biological motion, Perception, vol 31, pp. 837–53, 2002.

[33] T.L. Watson, J. Pearson, C.W.G. Clifford, Perceptual grouping of biological motion promotes binocular rivalry, Current Biology, vol. 14, pp. 1670–74, 2004.

[34] K. Fujimoto, Motion induction from biological motion, Perception, vol. 32, pp. 1273–77, 2003.

[35] D. Tadi, J.S. Lappin, R. Blake, E.D. Grossman, What constitutes an efficient reference frame for vision?, Natural Neuroscience, vol. 5, pp. 1010–15, 2002.

[36] J.A. Beintema, M. Lappe, Perception of biological motion without local image motion, National Academy of Science, vol. 99, pp. 5661–63, 2002.

[37] M. Shiffrar, J. Pinto, The visual analysis of bodily motion, Common Mechanisms in Perception and Action: Attention and Performance, Vol. XIX, ed. W Prinz, B Hommel, pp. 381–99. Oxford: Oxford Univ. Press, 2002.

[38] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russel, Toward robust automatic traffic scene analysis in real-time, International Conference on Pattern Recognition, Israel, pp. 126–131, 1994.

[39] R. Cutler C. Benabdelkader and L.S. Davis, Motion based recognition of people in eigengait space, IEEE International Conference on Automatic Face and Gesture Recognition, pp. 267–272, 2002.

[40] J. Little and J. Boyd, Recognizing people by theirgait: the shape of motion, Videre, vol. 1(2), pp. 1–32, 1998

[41] G. Bailo, M. Bariani, P. Ijas, M. Raggio, Background estimation with Gaussian distribution for image segmentation, a fast approach, in Proc. Measurement Systems for Homeland Security, Contraband Detection and Personal Safety Workshop, pp.2-5, March 2005.

[42] S. Fukui, T. Ishikawa, Y. Iwahori, H. Itou, Extraction of Moving Objects by Estimating Background Brightness, The Journal of The Institute of Image Electronics Engineers of Japan, vol. 33(3), pp. 350-357, 2004

[43] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, Wallflower: principles and practice of background maintenance, Proceedings of the International Conference on Computer Vision, pp. 255–261, 1999.

[44] Haritaoglu, D. Harwood, and L. S. Davis, Real-time surveillance of people and their activities, IEEE Transactions on Pattern Analalysis and Machine Intelligence, vol. 22, pp. 809–830, Aug. 2000

[45] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, Tracking groups of people, Computer Vision and Image Understanding, vol. 80(1), pp. 42–56, 2000.

[46] H.-Y. Shum, M. Han, and R. Szeliski, Interactive construction of 3D models from panoramic mosaics, CVPR, Santa Barbara, CA, pp. 427–433, 1998.

[47] R. Koch, 3D-Scene Modeling From Image Sequences, ISPRS Archives, vol. XXIV, part 3/W8, Munich, 17.-19. Sept. 2003.

[48] D. S Tweed, Motion Segmentation Across Image Sequences, Department of Computer Science, University of Bristol, PhD Thesis, April 2001.

[49] J. Lipton, H. Fujiyoshi, and R. S. Patil, Moving target classification and tracking from real-time video, IEEE Workshop on Applications of Computer Vision, pp. 8–14, 1998.

[50] Triggs and A. Zisserman (Eds.), Variational Space-Time Motion Segmentation, IEEE International Conferece on Computer Vision (ICCV), Nice, Vol 2, pp. 886–892, October 2003.

[51] Meyer, J. Denzler, and H. Niemann, Model based extraction of articulated objects in image sequences for gait analysis, IEEE International Conference on Image Processing, pp. 78–81, 1998.

[52] Meyer, J. Psl, and H. Niemann, Gait classification with HMM's for trajectories of body parts extracted by mixture densities, British Machine Vision Conference, pp. 459–468, 1998.

[53] J. Barron, D. Fleet, and S. Beauchemin, Performance of optical flow techniques, International Journal of Computer Vision, vol. 12, no. 1, pp. 42–77, 1994.

[54] N. Friedman and S. Russell, Image segmentation in video sequences: a probabilistic approach, 13th Conference on Uncertainty in Artificial Intelligence, pp. 1–3, 1997.

[55] J. Gao, A. G. Hauptmann, H. D. Wactlar, Combining Motion Segmentation with Tracking for Activity Analysis, The Sixth International Conference on Automatic Face and Gesture Recognition (FGR'04), pp. 699-704, Seoul, Korea, May 17-19, 2004

[56] T.K. Kim, S.F. Wong, R. Cipolla, Tensor Canonical Correlation Analysis for Action Classification, CVPR, pp. 1-8, 2007.

[57] P. Natarajan, V.K. Singh, R. Nevatia, Learning 3D Action Models from a few 2D videos for View Invariant Action Recognition, CVPR, 2010.

[58] R. Minhas, A. Baradarani, S. Seifzadeh, Q.M. JonathanWu, Human action recognition using extreme learning machine based on visual vocabularies, Neurocomputing archive, vol. 73, pp. 10-12, 2010.

[59] S. Ali, A. Basharat, and M. Shah, Chaotic invariants for human action recognition, ICCV, pp. 1-8, 2007.

[60] K. Schindler, L Van Gool, Action snippets: How many frames does human action recognition, CVPR, pp. 1-8, 2008.

[61] L. Suter, D. Wang, Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model, CVPR, pp.1-8, 2007.

[62] S. A. Niyogi and E. H. Adelson, Analyzing and recognizing walking figures in XYT, CVPR, pp. 469–474, 1994.

[63] T. Zhao, T. S. Wang, and H. Y. Shum, Learning a highly structured motion model for 3D human tracking, Asian Conference on Computer Vision, Melbourne, Australia, pp. 144–149, 2002.

[64] J. C. Cheng and J. M. F. Moura, Capture and representation of human walking in live video sequence, IEEE Transactions on Multimedia, vol. 1, pp. 144–156, June 1999.

[65] C. Bregler, Learning and recognizing human dynamics in video sequences, CVPR, San Juan, Puerto Rico, pp. 568–574, 1997.

[66] H. Sidenbladh and M. Black, Stochastic tracking of 3D human figures using 2D image motion, European Conference on Computer Vision, Dublin, Ireland, pp. 702–718, 2000.

[67] E. Ong and S. Gong, A dynamic human model using hybrid 2D-3D representation in hierarchical PCA space, British Machine Vision Conference, U.K., pp. 33–42, 1999.

[68] H. Z. Ning, L.Wang,W. M. Hu, and T. N. Tan, Articulated model based people tracking using motion models, International Conference on Multi-Model Interfaces, pp. 115–120, 2002.

[69] N. Duygulu, P. Ikizler, Human Action Recognition Using Distribution of Oriented Rectangular Patches, ICCV, 2007.

[70] H. Ren, G. Xu, Human action recognition with primitive-based coupled-hmm, International Conference on Pattern Recognition, pp. 494-498, 2002.

[71] K. Jia, D.-Y. Yeung, Human action recognition using local spatio-temporal discriminant embedding, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008.

[72] L. Wang, D. Suter, Learning and matching of dynamic shape manifolds for human action recognition, IEEE Transactions on Image Processing, vol. 16, pp. 1646–1661, 2007.

[73] L. Wang, D. Suter, Visual learning and recognition of sequential data manifolds with applications to human movement analysis, Computer Vision and Image Understanding, vol. 110(2), pp. 153–172, 2008.

[74] L. Wang, T. Tan, H. Ning, W. Hu, Silhouette analysis-based gait recognition for human identification, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25(12), pp. 1505–1518, 2003.

[75] R. Souvenir, J. Babbs, Learning the viewpoint manifold for action recognition, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008.

[76] V. Parameswaran, R. Chellappa, View invariance for human action recognition, International Journal of Computer Vision, vol. 66(1), pp. 83–101, 2006.

[77] F. Bremond and G. Medioni, Scenario recognition in airborne video imagery, International Workshop on Interpretation of Visual Motion, pp. 57–64, 1998.

[78] M. Brand, N. Oliver, and A. Pentland, Coupled Hidden Markov Models for complex action recognition, IEEE Conference on Computer Vision and Pattern Recognition, pp. 994–999, 1997.

[79] T. Starner, J. Weaver, and A. Pentland, Real-time American sign language recognition using desk and wearable computer-based video, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 1371–1375, Dec. 1998.

[80] N. M. Oliver, B. Rosario, and A. P. Pentland, A Bayesian computer vision system for modeling human interactions, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 831–843, Aug. 2000.

[81] M. Brand and V. Kettnaker, Discovery and segmentation of activities in video, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 844–851, Aug. 2000.

[82] M. Yang and N. Ahuja, Extraction and classification of visual motion pattern recognition, IEEE Conference on Computer Vision and Pattern Recognition, 1998, pp. 892–897.

[83] U. Meier, R. Stiefelhagen, J. Yang, and A.Waibel, Toward unrestricted lip reading, International Journal of Pattern Recognition and Artificial Intelliggence, vol. 14, no. 5, pp. 571–585, Aug 2000.

[84] Y. A. Ivanov and A. F. Boblic, Recognition of visual activities and interactions by stochastic parsing, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 852–872, Aug. 2000.

[85] T.Wada and T. Matsuyama, Multi-object behavior recognition by event driven selective attention method, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 873–887, Aug. 2000.

[86] N. Johnson and D. Hogg, Learning the distribution of object trajectories for event recognition, Image and Vision Computing, vol. 14, no. 8, pp. 609–615, 1996.

[87] N. Sumpter and A. Bulpitt, Learning spatio-temporal patterns for predicting object behavior, Image and Vision Computing, vol. 18, no. 9, pp. 697–704, 2000.

[88] W. M. Hu, D. Xie, and T. N. Tan, A hierarchical self-organizing approach for learning the patterns of motion trajectories, IEEE Transactions on Neural Networks, vol. 15(1), pp. 135–144, Jan. 2004.

[89] J. Owens and A. Hunter, Application of the self-organizing map to trajectory classification, IEEE International Workshop on Visual Surveillance, pp. 77–83, 2000.

[90] D. H. Hubel and T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, Journal of Physiology, vol. 160(1), pp. 106–154, 1962.

[91] S. Marčelja, Mathematical description of the responses of simple cortical cells, JOSA, vol. 70(11), pp. 1297-1300, 1980.

[92] J. G. Daugman, Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 36(7), pp. 1169, 1988.

[93] Burcu Kepenekci and F. Boray Tek, Gozde Bozdagi Akar, Occluded Face Recognition Based on Gabor Wavelets, IEEE International Conference on Image Processing (ICIP), Rochester, New York, September 2002.

[94] D. Dumitrescu, Genetic chromodynamics, Studia Universitatis Babes-Bolyai Cluj-Napoca, Ser. Informatica, vol. 45(1), pp. 39–50, 2000.

[95] C. Stoean, M. Preuss, R. Gorunescu, D. Dumitrescu, Elitist Generational Genetic Chromodynamics - a New Radii-Based Evolutionary Algorithm for Multimodal Optimization, CEC 2005.

[96] R. Durbin, S. Eddy, A. Krogh and G. Mitchison. Biological Sequence Analysis. Cambridge University Press, 1998.

[97] H. Wang, M. Ullah, A. Kl¨aser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. BMVC, pp. 127–138, 2009.

[98] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, A biologically inspired system for action recognition, ICCV, 2007.

[99] L. C. Molina, L. Belanche, and A. Nebot. Feature Selection Algorithms: A Survey and Experimental Evaluation. Technical Report LSI-02-62-R, Universitat Politècnica de Catalunya, Barcelona, Spain, 2002

[100] Chen, Ming-Yu and Hauptmann, Alexander, MoSIFT: Recognizing Human Actions in Surveillance Videos, Computer Science Department, Paper 929, 2009.

[101] S.F. Wong and R. Cipolla, Extracting spatio-temporal interest points using global information, ICCV, 2007.

[102] Laptev and T. Lindeberg, Space-time interest points, ICCV, 2003.

[103] G. Willems, T. Tuytelaars, and L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, ECCV, 2008.

[104] Kläser, M. Marszałek, and C. Schmid, A spatio-temporal descriptor based on 3Dgradients, BMVC, 2008.

[105] T. Dean, G. Corrado, and R. Washington. Recursive sparse spatiotemporal coding, IEEE International Workshop on Multimedia Information Processing and Retrival, 2009.

[106] Galata, N. Johnson, and D. Hogg, Learning variable-length Markov models of behavior, Computer Vision and Image Understanding, vol. 81, no. 3, pp. 398–413, 2001.

[107] T. Kadir, M. Brady, Saliency, scale and image description, International Journal of Computer Vision, vol. 45, pp. 83–105, 2001.

[108] G. Willems, T. Tuytelaars, and L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, ECCV, 2008.

[109] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, ICCV, vol. 1, pp. 166–173, October 2005.

[110] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, vol. 20, pp. 91–110, 2003.

[111] A. Efros, A. Berg, G. Mori, and J. Malik, Recognizing action at a distance, ICCV, pp. 726–733, 2003.

[112] A. Kale, A. Sundaresan, A.N. Rajagopalan, N.P. Cuntoor, A.K. Roy-Chowdhury, V. Kruger, R. Chellappa. Identification of humans using gait, IEEE Transactions on Image Processing, vol. 13(9), pp. 1163–1173, 2004.

[113] A. Oikonomopoulos, I. Patras, M. Pantic. Spatiotemporal salient points for visual recognition of human actions, IEEE Transactions on Systems, Man, and Cybernetics, Part B 36 (3), pp. 710–719, 2005.

[114] A. Yilmaz, M. Shah, Actions sketch: a novel action representation, CVPR, pp. 984–989, 2005.

[115] A. Fathi, G. Mori, Action Recognition by Learning Mid-level Motion Features, CVPR, 2008.

[116] C. Schuldt, L. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach, CVPR, 2004.

[117] C. Liu, Pong C. Yuen, Human action recognition using boosted EigenActions, Image and Vision Computing 28 , pp. 825–835, 2010. .

[118] C. Thurau, V. Hlava,. Pose primitive based human action recognition in videos or still images, CVPR, pp. 1-8, 2008.

[119] A. Davis, J. Bobick, The recognition of human movement using temporal templates, PAMI, vol. 23(3), pp. 257–267, 2001.

[120] G. Th. Papadopoulos, V. Mezaris, I. Kompatsiaris2 and M. G. Strintzis, Accumulated Motion Energy Fields Estimation and Representation for Semantic Event Detection, CIVR, 2008.

[121] H. Uemura, S. Ishikawa, K. Mikolajczyk, Feature tracking and motion compensation for action recognition, BMVC, 2008.

[122] H. Yi, D. Rajan, L.-T. Chia, A new motion histogram to index motion content in video segments, Pattern Recognition, vol. 26 (9), pp. 1221–1231, 2005.

[123] E. Irani, M. Shechtman, Space-time behavior based correlation, CVPR, 2005.

[124] J.W. Davis, S.R. Taylor, Analysis and recognition of walking movements, ICPR, pp. 315–318, 2002.

[125] J.C. Niebles, C.W. Chen, and Li Fei-Fei, Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification, ECCV, 2010.

[126] L. Yeffet, L. Wolf, Local Trinary Patterns for Human Action Recognition, ICCV, 2009.

[127] I. Laptev, On space–time interest points, International Journal on Computer Vision vol. 64 (2–3), pp. 107–123, 2005.

[128] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, CVPR, 18, 2008.

[129] J. Li, F. Niebles, A hierarchical model of shape and appearance for human action classification, CVPR, 2007.

[130] M. Ahmad, S.-W. Lee, Human action recognition using shape and clg-motion flow from multi-view image sequences, Pattern Recognition, vol. 41(7), pp. 2237–2252, 2008.

[131] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, Actions as space-time shapes, ICCV, vol. 2, pp. 1395–1402, 2005.

[132] M. Brand, V. Kettnaker, Discovery and segmentation of activities in video, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22(8), pp. 844–851, 2000.

[133] D. Mikel Rodriguez, J. Ahmed, M. Shah, Action MACH A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition, CVPR, 2008.

[134] Y. Mori, W. Greg, Max-Margin Hidden Conditional Random Fields for Human Action Recognition, CVPR, 2009.

[135] N.M. Oliver, B. Rosario, A.P. Pentland, A Bayesian computer vision system for modeling human interactions, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22(8), pp. 831–843, 2000.

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

[137] J.C. Niebles, H. Wang, L. Fei-Fei, Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words, International Journal of Computer Vision, vol. 79, pp. 299-318, 2008.

[138] O. Masoud, N. Papanikolopoulos. A method for human action recognition, Image and Vision Computing, vol. 21(8), pp. 729–743, 2003.

[139] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, Behavior recognition via sparse spatio-temporal features, VS-PETS, 2005.

[140] Q. Shi, L. Wang, L. Cheng, A. Smola, Discriminative human action segmentation and recognition using semi-Markov model, IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[141] R.V. Babu, K.R. Ramakrishnanb, Recognition of human actions using motion history information extracted from the compressed video, Image and Vision Computing, vol. 22(8), pp. 97–607, 2004.

[142] S. Xiang, F. Nie, Y. Song, C. Zhang, Contour graph based human tracking and action sequence recognition, Pattern Recognition, vol. 41(12), pp. 3653–3664, 2008.

[143] S.M. Oh, J.M. Rehg, T. Balch, F. Dellaert, Learning and inferring motion patterns using parametric segmental switching linear dynamic systems, International Journal of Computer Vision, vol. 77(3), pp. 103–124, 2008.

[144] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, International Conference on Machine Learning, pp. 495-502, 2010.

[145] T. Xiang, S. Gong, Beyond tracking: modelling activity and understanding behaviour, International Journal of Computer Vision, vol. 67(1), pp. 21-51, 2006.

[146] Y. Wang, P. Sabzmeydani, and G. Mori, Semi-Latent Dirichlet Allocation: A Hierarchical Model for Human Action Recognition, 2nd Workshop on Human Motion Understanding, Modeling, Capture and Animation (at ICCV), 2007.

[147] O. Chomat, J. L. Crowley, Probabilistic Recognition of Activity using Local Appearance, CVPR'99, Fort Collins, Colorado, USA, June 23–25, 1999

[148] T. Jolliffe. Principal Component Analysis. Springer Verlag, New York. 1986.

[149] K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, New York, 2nd edtion. 1990.

[150] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification. JohnWiley, USA, 2nd edition, 2001.

[151] S.F. Wong, T.K. Kim, R. Cipolla, Learning Motion Categories using both Semantic and Structural Information, IEEE CVPR, pp. 1-6, 2007.

[152] D. Gabor, Theory of Communication, J. IEE, vol. 93, pp. 429-459, 1946.

[153] A. Forner-Cordero, H. Koopman, and F. Van der Helm, Describing Gait as a Sequence of States, Journal of Biomechanics, 2005.

[154] C. S. Myers, L. R. Rabiner, A comparative study of several dynamic time-warping algorithms for connected word recognition, The Bell System Technical Journal, vol. 60(7), pp. 1389-1409, 1981.

[155] M. Marszalek, I. Laptev, C. Schmid, Actions in context, IEEE CVPR, pp. 2929-2936, 2009.

[156] Y. Wang, G. Mori, Human action recognition by semilatent topic models, IEEE TPAMI, vol. 31, pp. 1762-74, 2009.

[157] KTH Dataset, http://www.nada.kth.se/cvap/actions/, last visited on March 2011

[158] Weizmann Dataset, http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html, last visited on March 2011.

[159] R. Blake, M. Shiffrar, Perception of Human Motion, Annual Review Psychology, vol.58, pp.47–73, 2007.

[160] T. Watson, J. Pearson, C. Clifford, Perceptual grouping of biological motion promotes binocular rivalry, Current Biology, vol. 14, pp. 1670–74, 2004.

[161] J. Liu, S. Ali and M. Shah, Recognizing human actions using multiple features, CVPR, 2008.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Kepenekci, Burcu
Nationality: Turkish (TC)
Date and Place of Birth: 11 April 1978 , Ankara
Marital Status: Single
Phone: +90 312 472 54 88
Fax: +90 312 210 11 78
email: burcu.kepenekci@paranavision.com.tr

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Electrical and Electronics Engineering	2001
BS	Başkent University Electrical and Electronics Engineering	2000
High School	High School Yükseliş Collage, Ankara	1996

WORK EXPERIENCE

Year	Place	Enrollment
2006-Present	ParanaVision Machine Vision Technologies	R&D Director
2000-2006	TUBİTAK BİLTEN	Research Engineer
1999-2000	Başkent University Computer Engineering	Student Assistant
1999 July	Başkent University Electrical and Electronics Engineering	Intern Engineering Student
1998 August	ASELSAN	Intern Engineering Student

FOREIGN LANGUAGES

Advanced English

PUBLICATIONS

1. B. Kepenekci, G. Akar, *Activity Modeling with Spatio-Temporal Texture Primitives*, Wiamis 2011, paper 75.

2. B. Kepenekci, G. Akar, *Motion analysis using 3D Gabor kernels*, Signal Processing, Communication and Applications Conference, 2008. SIU 2008. IEEE 16th, 20-22 April 2008, pp. 1 – 4.

3. Burcu Kepenekci, Gözde Bozdağı Akar, *Destekçi Vektör Makinası ile Yüz Sınıflandırma,* SİU 2004, Kuşadası, Nisan 2004.

4. Burcu Kepenekci and Gozde Bozdagi Akar, *GAYE: A Face Recognition System,* Electronic Imaging 2004, San Jose, California, 18-22 January 2004.

5. K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F. B. Tek, G. B. Akar, F. Deravi, and N. Mavity. Face verification competition on the XM2VTS

database. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, pages 964-974. Springer-Verlag, 2003.

6. Burcu Kepenekci and Gozde Bozdagi Akar, *Hareket Tabanlı Video Bölütleme (Motion Based Video Segmentation),* Siu2003 (National Conference on Signal Processing and Applications), İstanbul, June 2003.

7. Burcu Kepenekci and F. Boray Tek, Gozde Bozdagi Akar, *Occluded Face Recognition Based on Gabor Wavelets,* IEEE Int. Conf. on Image Processing (ICIP), Rochester, New York, September 2002.

8. Burcu Kepenekci, Gozde Bozdagi Akar, *Video Dizilerinde Tanıma Amaçlı Yüz İzleme (Face tracking in video squences for face recognition),* Siu2002 (National Conference on Signal Processing and Applications), Denizli, June 2002.

9. Burcu Kepenekci and F. Boray Tek, Gozde Bozdagi Akar, *Güvenlik Amaçlı Yüz Tanıma Sistemi (A Security Purposed Face Recognition System),* Siu2002 (National Conference on Signal Processing and Applications), Denizli, June 2002.

10. Burcu Kepenekci and F. Boray Tek, Gozde Bozdagi Akar, *Wavelet Based Face Recognition,* Siu2001 (National Conference on Signal Processing and Applications), obtained 2nd place in 'Alper Atalay Best Student Paper Competition', April 2001.

11. Burcu Kepenekci, Hasan S. Bilge and Mustafa Karaman, *A New Segmentation Method for Speckled Images*, Siu2000 (National Conference on Signal Processing and Applications), June 2000.

HOBBIES

Motorcycles, Computer Technologies, Swimming.