Improved Multi-resolution Analysis of the Motion Patterns in Video for Human Action Classification

by

Hossein Shabani

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2011

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Hossein Shabani

Abstract

The automatic recognition of human actions in video is of great interest in many applications such as automated surveillance, content-based video summarization, video search, and indexing. The problem is challenging due to a wide range of variations among the motion pattern of a given action such as walking across different subjects and the low variations among similar motions such as running and jogging.

This thesis has three contributions in a discriminative bottom-up framework to improve the multi-resolution analysis and classification of the motion patterns in video for better recognition of human actions.

The first contribution is the introduction of a novel approach for a robust local motion feature detection in video. To this end, four different multi-resolution temporally causal and asymmetric filters of log Gaussian, scale-derivative Gaussian, Poisson, and asymmetric sinc are introduced. The performance of these filters is compared with the widely used multi-resolution Gabor filter in a common framework for detection of local salient motions. The features obtained from the asymmetric filtering are more precise and more robust under geometric deformations such as view change or affine transformations. Moreover, they provide higher classification accuracy when they are used with a standard bag-of-words representation of actions and a single discriminative classifier. The experimental results show that the asymmetric sinc performs the best. The Poisson and the scale-derivative Gaussian perform better than log Gaussian and that better than the symmetric temporal Gabor filter.

The second contribution is the introduction of an efficient action representation. As the salient features at different spatial and temporal scales characterize different motion information, a multi-resolution action signature is developed for a more discriminative video representation.

The third contribution is on the classification of different human actions. To this end, an ensemble of classifiers in a multiple classifier systems (MCS) framework with a parallel topology is utilized. This framework can fully benefit from the multi-resolution characteristics of the motion patterns in the human actions. The classification combination concept of the MCS has been then extended to address two problems in the configuration setting of a recognition framework, namely the choice of distance metric for comparing the action representations and the size of the codebook by which an action is represented. This implication of MCS at multiple stages of the recognition pipeline provides a multi-stage MCS framework which outperforms the existing methods which use a single classifier.

Based on the experimental results of the local feature detection and the action classification, the multi-stage MCS framework, which uses the multi-scale features obtained from the temporal asymmetric sinc filtering, is recommended for the task of human action recognition in video.

Acknowledgements

Let me start by thanking God for all the givings and his mercy. I am grateful of his guidance in my life to think and act positive. I wish to thank my great family for their unconditional love, continuous support, and encouragement. I am thankful of my parents follow up and support through my long, and maybe unfinished journey of science and education.

I would like to express my sincere acknowledgement in the support and help of my supervisors Professor David Clausi and Professor John Zelek. They always keep their doors open in order to discuss the research with passion and great enthusiasm. I really appreciate their kind consideration and help when I needed to be with family overseas. They have contributed tremendously to my growth as a dedicated researcher by their advice and great support throughout my PhD program.

My sincere thanks to my committee members for their guidance and valuable inputs on this research, Professor Paul Fieguth (Systems Design Eng.), Professor Arsen Hajian (Systems Design Eng.), Professor Oleg Michailovich (Electrical and Computer Eng.), and Professor Greg Mori (Computing Science, Simon Fraser Univ.).

I wish to thank my great friends and the Vision and Image Processing (VIP) lab members who made my stay at Waterloo an important chapter of my life with their inspiration and social support. Throughout my graduate program, I have collaborated and subsequently became friends with many people. Special thanks to these friends and the faculty members of Engineering, Computer Science, Mathematics, Psychology, and Vision Science at the University of Waterloo who were open to discuss about the research with passion. Special thanks to Ms. Vicky Lawrence and Ms. Carrie Gilmour who were always nice and helpful.

Finally, I wish to thank all of the funding agencies and corporations that supported my research through financial and knowledge support. In particular, I wish to thank Geomatics for Informed Decision (GEOIDE), a Network for Centers of Excellence supported by the Canadian funding agency Natural Sciences and Engineering Research Council (NSERC), the Ontario Graduate Scholarship (OGS) program, and the Faculty of Engineering at the University of Waterloo.

Dedication

I dedicate this thesis to my parents whom I love unconditionally.

Table of Contents

List of Tables

List of Figures

List of Symbols

Symbol	Definition
$\bigcap$	intersection
$\bigcup$	union
Σ	covariance matrix
s	scale
σ	spatial scale
τ	temporal scale
div	divergence
∇	gradient operator
Δ	Laplacian operator
G	Gaussian function
S	step function
h	Hilbert transform
sin	sine function
cos	cosine function
$arctan$	arctangent
ln	natural logarithm
log	logarithmic function
$hist$	histogram
q^{even}	even component of a complex filter q
q^{odd}	odd component of a complex filter q
R	energy of a complex filter
$\tilde{U}$	Fourier transform of function u
w_x	spatial horizontal frequency
w_y	spatial vertical frequency
w_t	temporal frequency
$(x; y)$	spatial coordinate
$(x; y; t)$	spatial and temporal coordinate
$u(x; y; t)$	intensity of a pixel
$i(x; y; t)$	current between two adjacent pixels

List of Abbreviations

Abbreviation	Complete form
BOW	Bag-Of-Words
HCRF	Hidden Conditional Random Field
HOG	Histogram of Oriented Gradients
LDS	Linear Dynamic System
MAD	Median Absolute Deviation
MAP	Maximum A Posterior
MCS	Multiple Classifier Systems
MKL	Multiple Kernel Learning
ML	Maximum Likelihood
NN	Nearest Neighbor
NLDS	Non Linear Dynamic System
PCA	Principle Component Analysis
PDE	Partial Differential Equations
pLSA	probabilistic Latent Semantic Analysis
RBF	Radial Basis Function
SIFT	Scale-Invariant Feature Transform
SURF	Speed-Up Robust Features
SVD	Singular Value Decomposition
SVM	Support Vector Machine
2D	Two Dimensional spatial space

Chapter 1

Introduction

Humans can easily detect and recognize the type of actions performed by individuals in a video. The automated recognition of human actions such as ”walking”, ”running”, and ”hand waving” is however a challenge in computer vision. An action is a sequence of movements of the limbs performed by a single person. Human actions usually have some predictable nature such as periodicity in walking or running.

Automated human action recognition is important in many video analytic applications such as activity analysis for automated surveillance, elderly home monitoring, video retrieval, video summarization, and human-computer interaction. For security applications, for example, automatic recognition of suspicious activities such as loitering, fighting, abandoning a bag, and entering into a restricted zone is helpful for security personnel. For a home monitoring application, an automatic action recognition system can be utilized in an assistive technology system purpose to provide feedback on the health condition of the elder by looking at the changes in the pattern of daily activities or by sending a help signal to the emergency station if the elder falls on the ground. In the web technology, the popularity of multimedia share environments such as Youtube and the social networks on the internet demand the search by visual query, indexing, and video summarization in which the human action recognition and retrieval is a fundamental task towards semantic description of the video contents.

The difficulty in automating the action recognition task is due to a number of challenges. There exist high variations in motion patterns across different individuals. That is, different people walk differently and even an individual walks differently with different moods. Moreover, the motion patterns of actions such as running and jogging are similar and hence, their discrimination is difficult. In addition, variations due to camera zooming or shaking, perspective view change, and illumination changes can cause more challenge to automatically interpret the scene.

This thesis will describe novel tools for improving computer-aided human action recognition. The new techniques are proposed for a bottom-up human action recognition system. The contributions are in three stages of the process: (1) low-level feature extraction, (2) mid-level action representation, and (3) high-level action classification. First, a new approach for robust feature extraction in a real-time video is proposed (Chapter 3). This approach is inspired by the motion perception models from biological vision research. The next contribution is on the introduction of a multi-resolution action encoding which provides a more discriminative action representation for better discrimination of similar motions (Chapter 4). The third contribution is on using a multiple classifier systems to retain the motion characteristic of low-level features all the way to the classification stage and hence, obtain a higher classification rate (Chapter 5).

Chapter 2

Background

2.1 Overview of existing methods

Different criteria can be adopted for the categorization of the general motion patterns in video. According to Polana and Nelson [102], a motion pattern in a video can be from one of three major categories: (1) Structured motions which include actions or activities such as human walking or animals flying. These types of motions are temporally periodic and have compact structure, (2) Isolated events such as opening a door, starting a car, or throwing a ball which are isolated simple motions with no temporal repetition, and (3) the statistical motions which are temporal (or dynamic) textures such as flowing water, smoke, or fluttering leaves. There is a statistical regularity in these motion patterns but the spatial and temporal extent of the motion is indeterminate.

According to Guerra-Fillho and Aloimonos [42, 43], there exist three different spaces and correspondingly three different languages to represent an action. (a) The visual space, (b) The motor space, and (c) The natural language (i.e., linguistic) space. Humans use a visual language to see and understand actions, use a motor language to produce an action, and use a natural language to talk about actions. Motivated from the developments in the natural language processing and speech processing literature, the authors developed the phonology, morphology and syntax of these different languages as well as their relationships. In their human activity language (HAL), an action is represented by the structure of a sentence using four lexical categories of noun, adjective, verb, and adverb. One however should note that the range of actions in the visual space, for example, might differ from those in the motor or the linguistic space. For example, there is not a direct visual evidence for a lot of verbs such as ”thinking”, ”feeling”, and ”affording”, and hence, there is not a one-to-one mapping of the verbs with the visual actions. Moreover, the definitions might not necessarily be equivalent across these three spaces. This inequivalence

is much more for the actions compared to the objects for which there exist only two spaces of visual and linguistic. To give an example, an "apple" and a "heart" are close in the visual space of objects, but they are quite different in the linguistic/semantic space.

In the human action/activity recognition literature [2, 19, 22, 25, 53, 68, 112, 131], however, there is not a well-defined hierarchy/taxonomy of human activities or a clear definition of human actions. Different publications used action, activities, or events interchangeably. This thesis uses a four sub-categories of human motions in video to make them distinguishable from each other (Fig. 2.1):

1. A local motion event such as moving arms which has a determinate space and time extension.

2. Atomic actions such as walking or running are performed by a single person and do not necessarily require a context to be specified.

3. An activity has a context and it describes (a) the interaction of an individual with the objects in the environment such as swinging a golf club or (b) the interaction of multiple people with each other such as payment at the cashier desk.

4. A global event such as a marriage or people playing soccer is a large-scale phenomenon.

In a chronological order, a local motion event might span just in a fragment of a second while a global event such as the soccer game might take an hour or more. This hierarchical of human motions is just a sample categorization which satisfies the purpose of this thesis. From now on, the focus will be on atomic actions.

Human action recognition in video can be generalized into two different approaches [3, 104]: top-down and bottom-up. A top-down approach requires foreground segmentation and explicit motion modeling for human detection and tracking [4, 16, 17, 53, 92, 94]. In a constrained environment, a top-down approach works fine if the body parts are quite visible and there is not significant geometric/photometric variations, background or camera motion, clutter, and occlusion [104]. For processing of videos captured under real-world situations such as Youtube videos or low quality (e.g., low frame-rate and low-resolution) surveillance videos, a bottom-up approach [25, 68, 122, 129] is more promising as it does not require explicit video segmentation and body limb tracking. Bottom-up approaches are thus more universal and attractive for a wider range of applications. For action recognition, they can be used in a wider scales from near-field to far-field situations. The focus of this thesis is on the bottom-up approach. In this chapter, we mainly introduce the components of this approach. Each component is then discussed in detail in subsequent chapters in which the related literature review is provided.

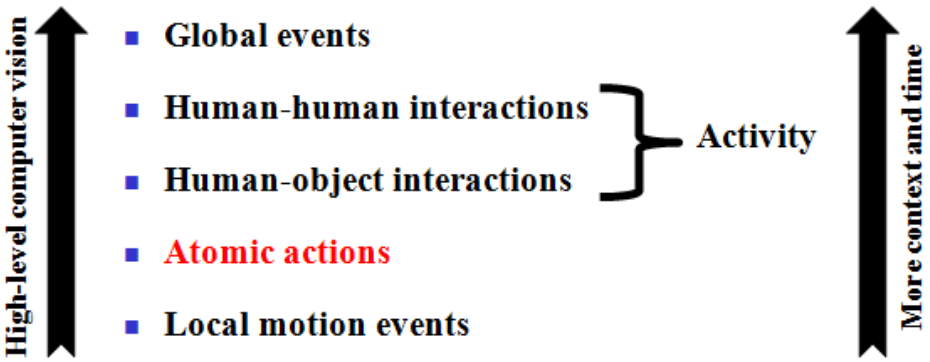


Figure 2.1: Hierarchy of human motions. Note that high-level motion patterns such as global events take longer and context is an important component of their description. The focus of this thesis is on atomic actions.

2.2 Bottom-up methods

A typical bottom-up approach performs the video analysis task using motion features. Motion features characterize the variations in the sequence of images. The global motion features such as motion history image/volumes [23, 128] or histogram of optical flow [17] have not been demonstrated to be reliable in the presence of occlusion, clutter, and noise [68, 104]. Moreover, decomposed motion templates require background subtraction to extract the optical flow which is not a reliable measure in presence of fast foreground motions, camera/background motions, or occlusion and clutter. In contrast, local motion features which correspond to local events are more robust under these situations and hence, more attractive for motion analysis. In the video of an action, for example, the local salient motion features occur when there is a change in the motion of a body limb [25, 67]. Intuitively, the time when and the location where a body limb changes its motion direction is a salient local motion event.

Drawing inspiration from the promising performance of local salient spatial features in the object recognition literature [84], the extraction of spatio-temporal salient features has been of interest for action recognition [25, 67, 97, 129]. These features are the low-level visual cues from the sequences of the frames (Fig. 2.2). An intermediate representation of human actions can thus be obtained by grouping these low-level features into visual words. A statistical measure such as the frequency of appearance of the features in an action is then the action signature. This representation of an action in a video is referred to as a global bag-of-words (BOW) representation [25, 68]. Using a supervised classification approach, a discriminative classifier such

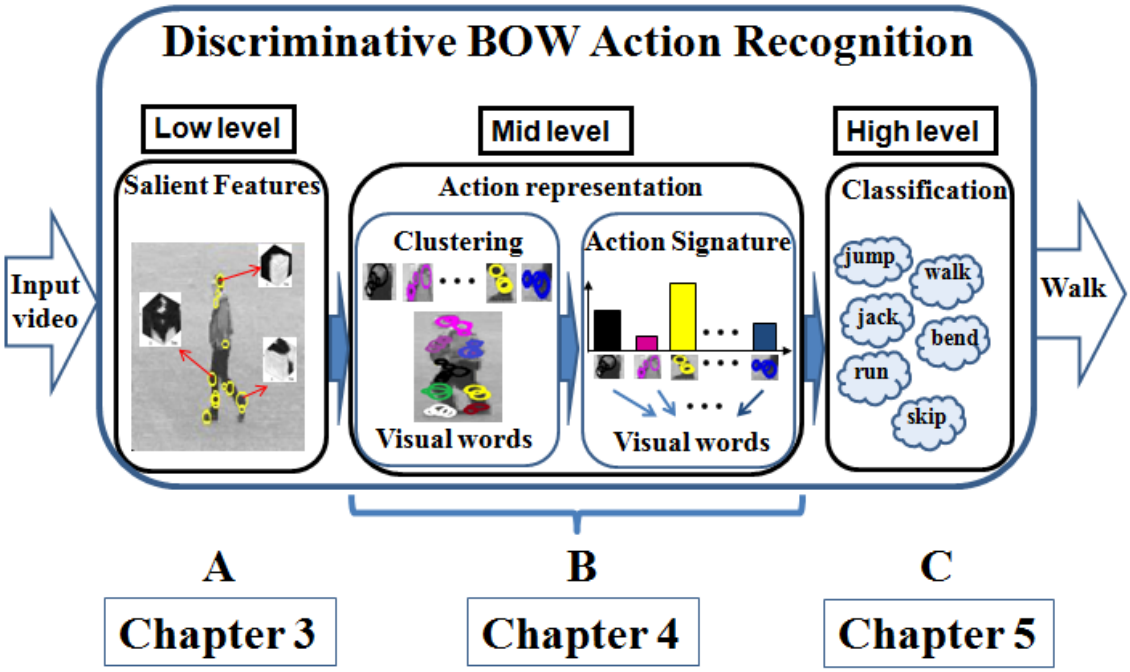


Figure 2.2: Different components of a discriminative bag-of-words (BOW) framework for human action recognition. Chapter 3 explains the salient motion feature extraction (module A). Chapter 4 explains the action representation using visual words and the action signature (module B). Chapter 5 describes different classification approaches (module C).

as support vector machine (SVM) can then use the signatures of different actions to learn the decision boundaries for the classification of different motion patterns. An unsupervised classification approach, such as a probabilistic latent semantic analysis (pLSA) [95] can use the BOW representation of different actions to learn a model for each actions. One might use a structural pLSA [135] to model the temporal relationship among the features or use a hidden conditional random fields (HCRF) [124] to learn the hidden body parts. The structural pLSA is a generative model and differs from a discriminative SVM classifier. In essence, the discriminative classifier performs directly a maximum likelihood (ML) estimation, while a generative model performs a maximum a posteriori (MAP) estimation [24]. Fig. 2.2 shows the components of a typical discriminative BOW framework which is considered as the baseline method for this thesis. One should note that the design of each component is not independent from the other components and an overall understanding of the whole recognition framework is necessary before choosing any specific approaches for each component. More specifically, the choice of action representation

or classification might require specific features. For example, a dynamic system modeling might require (spatial) feature detection at each frame, while a global BOW representation of video is obtained from the spatio-temporal features.

2.2.1 Spatio-temporal salient features

A local motion event occurs in a space-time volume centered at a spatio-temporal key/interest point. The event is then described using the characteristic appearance and motions of the object in the local volume. The detection and description of the local space-time volumes are usually referred to as salient feature extraction [25, 68, 129]. In a recognition task such as action classification, to address the variability in the spatial size and temporal interval of an action performed by different people or even an individual, the video events should be detected at multiple spatial and temporal scales [68, 129].

Most of the existing spatio-temporal salient feature detectors and descriptors are the extension of their counterparts from a still image. For feature detection, for example, Laptev et al. [67, 80] extended the 2D Harris corner detector [12] to video to detect where and when the sub-actions start and stop (Fig. 2.3). Willems et al. [129] extended 2D Hessian blob detection to 3D to detect 3D ellipsoid-like shapes in a video. The information theoretic salient feature of Kadir et al. [51] has been extended to time domain in [97] for a sparse representation of the video content. Dollar et al. [25] proposed "cuboids" which occur at the salient motion points in the video. For feature description, for example, Scovanner et al. [111] extended the descriptor of a 2D scale-invariant feature transform (SIFT [84]) to video. This 3D SIFT descriptor is the histogram of oriented gradients (HOG) in space and time. To compute this histogram, the local volume of the salient feature is divided into 4×4 non-overlapped sub-windows. The spatio-temporal oriented gradients of all sub-windows are then sorted in order to provide the final 3D SIFT histogram of the feature. The overlapped sub-windows version of this histogram is referred to as a 3D HOG [57] which is the extension of its counterpart from a still image, a 2D HOG [22]. A combination of HOG and HOF (histogram of optical flow) [68] is also used to describe the characteristic appearance and motion information in the feature's volumetric extension. In [68], the extension of local jets (i.e., steerable n^{th}-order gradients) to the video domain is used to describe the characteristics of a salient feature. The Haar wavelet coefficients are the descriptor used for 2D SURF (speed-up robust features) [9] and has been extended to the video in [129].

A comparative study of some of these spatio-temporal feature detectors and descriptors and a method for dense sampling of features has been published by Wang et al. [122] for human action classification. Without any motion analysis, dense sampling considers the salient features to be at the nodes of a regular grid with a given spatio-temporal resolution. The dense sampling

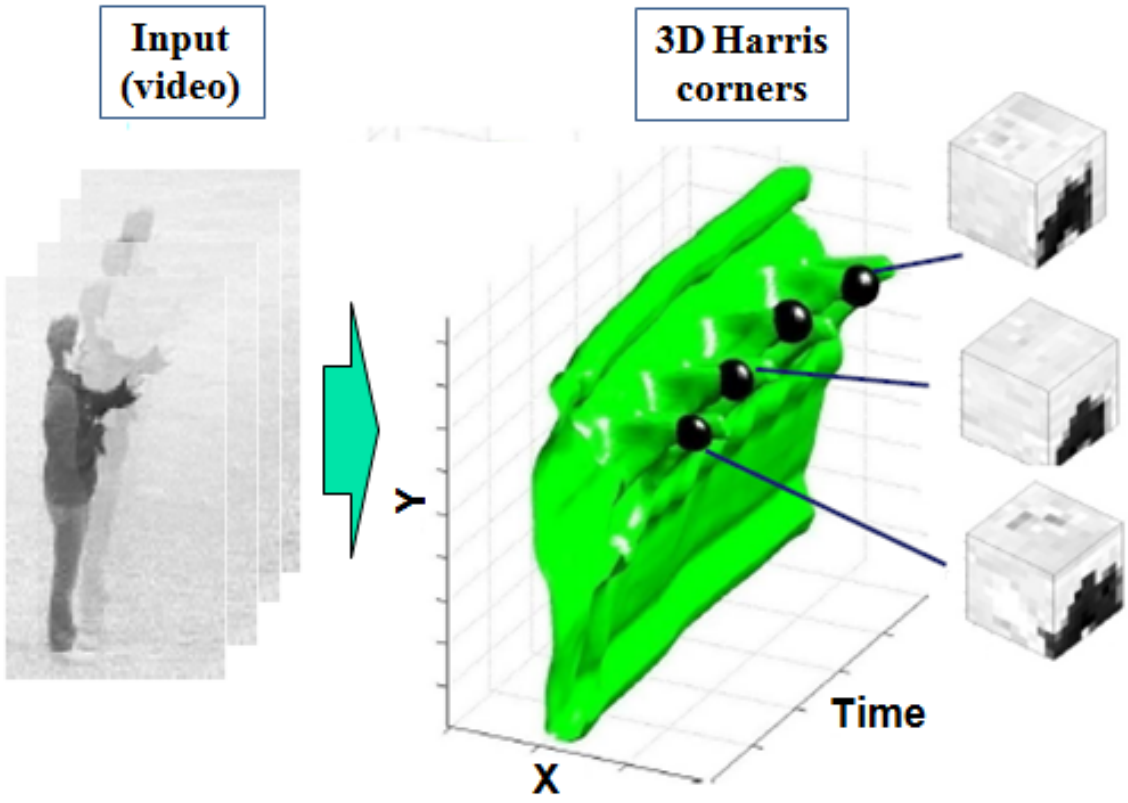


Figure 2.3: The input is a video of a person performing a ”boxing” action. As depicted in [68], a 3D Harris method detects the local space-time volumes (filled dark ellipsoids) in the video in which the sub-actions start or stop.

method did not perform better than the salient features on the recognition of atomic actions. However, in the dense sampling approach, some of the features are from the moving background by which some contextual information on where the actions are happening is encoded. This is believed to be a possible explanation why dense sampling performed better for the recognition of actions such as sports actions in which the context is a significant clue. The authors then concluded that the dense sampling method performs better on the real world videos, but not on simple videos. There are two critiques on the experimental setting of this paper which make the final conclusion of this study not valid. The most serious problem is that the number of scales at which the salient features have been detected varies from one feature to another. More specifically, the cuboids have been detected at just one spatio-temporal scale, the 3D Harris corners at twelve spatio-temporal scales (six spatial and two temporal scales), and the 3D Hessian

features and the dense sampling have also been detected at multiple scales. The performance of the cuboids is very much close to that of other multi-scale features or dense sampling. With this consideration in mind, it is expected that the multi-scale cuboids should perform better than the other feature detectors. The second critique is that the action classification accuracy is just one measure on the performance evaluation of the feature detectors. More fundamental measures such as the precision and the reproducibility under different geometric variations is a requirement for robustness evaluation of feature detectors. In addition to these experimental critiques, as we will show in this thesis (Section 3.7.4), using a proper temporal asymmetric filtering, one can detect salient features which perform better than dense sampling not only on simple datasets such as KTH, but also on real-world dataset such as UCF sports action dataset [107]. These results contradict the generalization that has been made in [122]. In fact, our finding shows that one can obtain high accuracy using a set of sparse salient features detected using a proper temporal filter such as asymmetric sinc (Section 3.3.2).

2.2.2 Action representation

Video salient features of a unique action performed by different people or the same person might vary due to appearance and motion difference. To reduce some of these intra-class variations and provide a more robust matching at the later stage of action recognition, similar salient features are grouped in a similar cluster. The collection of these clusters for all action types provides a dictionary (or a code book) of action prototypes, referred to as visual words [68]. During training, the system learns the dictionary of the visual words by which an action can then be represented. A typical and simple representation of an action is the global BOW representation. Instead of a global BOW representation, one might directly feed the set of features of a video to a classifier for matching [68]. The BOW representation originated from natural language processing for document classification [60] and has been widely used for visual recognition systems as well [83, 68]. The global BOW is an orderless representation as it does not consider any spatial or temporal relationship among the salient features in the action modeling. To incorporate some spatial and temporal localization of the features, Marszalek et al. [88] proposed to consider several channels of BOW representation localized at different spatio-temporal extents of the video. The set of channels that jointly provides a better discrimination of actions in the training is then considered to be used as the final multi-channel classification at the testing stage.

2.2.3 Action classification

The support vector machine (SVM) classifier has been widely used for the classification of human action using a discriminative bottom-up approach [25, 68, 122]. Most of these recognition

frameworks use a single SVM classifier for discrimination of motion patterns. The standard approach for performing a multi-class classification problem is considered to be the one-against-all approach [21, 119]. With the promising performance of multi-channel action representation [88] versus the single global BOW representation, Weinland et al. [127] developed a more dense sampling of these channels to better handle the partial occlusions and the view changes. In this extension, the collection of bounding boxes around the person of interest (i.e., the foreground volume) is first obtained. The foreground volume is then divided into densely distributed overlapping sub-volumes to each of which a separate SVM classifier is assigned. In an adaptive weighting approach, the sub-volume which is occluded contributes less in the final decision making.

2.2.4 Limitations of the existing bottom-up methods

A typical bottom-up discriminative action recognition is a serial multi-stage process in which the performance of each stage is affected by the previous stage(s). More specifically, the low level features that feed into a bottom-up framework play a significant role in an appropriate representation of different actions and finally in the recognition performance. An appropriate discriminative representation of different actions should definitely help the classifier to provide a better estimation of the class label.

Generally speaking, video analysis methods are an extension of their counterparts in image analysis. The solutions to many problems in the video action recognition task have been significantly inspired by similar approaches in the image object recognition task. Establishment of a new video analysis approach on a well-studied method of object recognition is an advantage. However, one should note that the time domain is different from the spatial domain [77]. The limitations of the existing discriminative bottom-up approaches for the action recognition are explained according to the major components of such a framework (Fig. 2.2) as follows.

1. **Feature extraction**. The extraction of features which are discriminative and robust is a crucial step for the task of action classification. To this end, the detection of foreground salient features which are reproducible under geometric/photometric variations is the first step. The features should also capture the characteristics of the motions of interest, such as local periodicity in walking or running. The literature of spatio-temporal salient feature extraction suffers from a similar comprehensive study as in the case of spatial salient features [90, 89] to evaluate the robustness of the features. Moreover, the existing methods consider the 2D+t video signal as a 3D object and hence, they apply similar spatial filters to the temporal domain. Considering the nature of time and motion which are different from the space and the shape structures, this treatment might not be valid.

2. **Action representation**. Despite the extraction of salient motion features at different spatial and temporal scales, existing methods do not explicitly benefit from the motion information across different scales. For example, in the BOW representation, either the motion redundancy across different spatio-temporal scales is removed when the scale-invariant features [67, 68, 70, 129] are used or it is implicit in the representation when the features of all scales are blindly combined [122]. A possible source of confusion of actions such as running and jogging with relatively similar motion patterns might be due to the fact that the motion characteristic of the feature is buried in the action representation. In fact, considering the high intra-class variations and the low inter-class variations of human motions, a multi-resolution analysis of actions should be a requirement. A multi-resolution approach can represent the salient motion features at different scales. The hypothesis is that the multi-resolution representation explicitly benefits from the full range of motion characteristics which exist across scales. It should thus provide an over-complete representation with a potentially better discrimination power.

3. **Action classification**. Most of the existing discriminative approaches use a single SVM classifier for the action discrimination [68, 129, 122, 112]. In these methods, the multi-scale salient features are combined at the early stage of the process. This treatment dose not retain the multi-resolution characteristics of the motion patterns and hence, the classifier dose not have access to this explanatory information. Note that the average motion pattern of ”jogging” and ”running” are quite similar and hence, their coarse-scale motion features are probably similar. However, the fine-scale motion features of these actions should be different. One possible explanation on the low performance of most of existing methods is that they feed mixture of all of these features into a classifier, directly. The classifier can not use the discrimination power of the features at different scales.

2.3 Problem statement

The objective of this thesis is to develop a robust algorithm to recognize human actions in video. For conciseness, the focus is on a scenario in which single person performs an atomic action. The type of actions of interest varies from moving a body limb such as a ”hand waving” to whole body movements such as ”walking”, ”running”, and ”diving”. As a bottom-up discriminative approach does not perform the difficult task of video segmentation, the human action recognition is formulated as a classification problem. In this context, action recognition refers to the labeling of an unknown action with a set of predefined classes. There are three main requirements that should be considered in the design of such a system. The low-level salient feature extraction method should provide robust and good quality features for motion encoding. The action

representation and classification method should address the high intra-class variations within a unique action type and the low inter-class variations among different actions with similar motion patterns. Moreover, the system should not require explicit parameter settings at different stages of the recognition framework such as the choice of similarity metric by the classifier for the comparison of the representation of two videos.

2.4 Thesis contributions

This thesis introduces a biologically-inspired approach for robust local motion event detection and a multi-resolution analysis approach to improve action recognition accuracy. These contributions are briefly listed here to justify new approaches to address the problems of a typical existing discriminative bottom-up action recognition framework.

2.4.1 Robust salient feature extraction

Existing methods of salient spatio-temporal feature extraction treat the time domain in the same manner as the spatial component and use a 3D filtering on the 2D+t video signal. Time causality is a significant characteristic which makes the time domain different from the non-causal spatial domain. Despite the theoretical discussion of time-causal multi-resolution video filtering in the literature of scale-space theory [59, 81, 30, 115], these developments have not been used by the action recognition community. Causal and asymmetric filtering is consistent with the functionality of the motion receptive fields in the human visual systems [1, 35, 115, 125]. This thesis promotes the use of asymmetric temporal filtering for robust motion feature detection. The hypothesis is that the consistency of asymmetric filtering with biological vision should provide a better motion representation and hence, a better local motion event detection.

This thesis adopts three existing time-causal multi-resolution filters for the salient feature detection task. In addition, a brand new asymmetric video filtering is developed. To this end, the classical model of Perona-Malik [100] for anisotropic diffusion filtering in a still image is extended to the time domain with the causality of time as the boundary constraint. The resulting kernel is an asymmetric sinc for the time-causal domain.

A comprehensive set of experiments to test the precision of the detected salient features and their reproducibility under different geometric variations such as affine transformation and view/scale changes have been performed. The results show that the salient features detected using an asymmetric temporal filtering are more robust than those detected using the widely used symmetric Gabor filtering. Moreover, these features provide higher action classification accuracy

in a standard BOW recognition framework. Based on these results, the use of an asymmetric sinc filtering is recommended for local salient motion feature detection in video.

2.4.2 Action modeling using multi-resolution BOW representation

Most of the existing bottom-up approaches do not characterize the multi-resolution nature of human motions in the action representation. This thesis promotes the use of multi-resolution action representation in which the motion information across multi-scale salient features are retained for a better action representation. The hypothesis is that since this representation fully benefits from the motion redundancy across scales of different features, it should provide more discriminative action encoding. In this setting, the action signatures of the features at different spatio-temporal scales are concatenated in order to obtain a multi-resolution action signature. This way the fusion of multi-resolution motions from the low-level of features is moved to the intermediate level of action representation. The experimental results on the discrimination of human actions shows that this approach provides a more discriminative representation of similar motions such as ”running” and ”jogging” and hence, it improves the classification accuracy.

2.4.3 Action classification using multiple classifier systems

Existing discriminative human action recognition approaches typically use a single classifier on the multi-scale salient features. This thesis introduces the multiple classifier systems (MCS) as an alternative classification approach. The MCS shifts the fusion of multi-resolution motion information across multi-scale salient features to the last stage of classification. This retainment of the motion information at the higher level of classification should empower the MCS to more easily discriminate similar actions. In this setting, a separate classifier with the features of each scale is used. Consequently, the MCS benefits from the estimation of several classifiers to provide a more reliable decision. The experimental results show that an MCS framework provides higher classification accuracy. Moreover, this thesis introduces a multi-stage MCS framework to address the problem of the choice of distance metric and the choice of the length of the dictionary of the visual words over which different actions are defined. Consequently, this multi-stage MCS framework is recommended as a robust multi-resolution analysis for the human action classification.

2.5 Structure of the thesis

Based on the contributions of this thesis on the three main components of a discriminative bottom-up action recognition, the next three chapters will focus on these components. Chapter 3 explains in more details the limitations of existing approaches in salient feature extraction in video. Later in the same chapter, the use of asymmetric temporal filtering versus symmetric temporal filtering is promoted. A comprehensive set of experiments are presented on the validity of this hypothesis that the asymmetric filtering provides more robust and informative features. Chapter 4 presents the importance of an explicit multi-resolution analysis for the action representation. It will be shown how a simple independent treatment of the motion features can provide a more discriminative action representation. Chapter 5 introduces the use of MCS for better estimation of action types using a combined classification approach. The MCS is then utilized at multiple stages of the recognition framework to address the choice of dictionary length and the choice of action similarity matching. This simple idea is a significant jump towards a systems that is less sensitive to the parameter tuning. In fact, a unified framework in Chapter 6 is presented which benefits from all the improvements that have been made in the previous sections.

Chapter 3

Spatio-temporal salient feature extraction

The focus of this chapter is on salient feature extraction. In general, spatio-temporal salient features capture the important changes in the scene [67, 98]. Salient feature extraction is the first low-level process (module A in Fig. 2.2) in a standard BOW representation for discriminative action recognition. The performance of the subsequent operations is therefore related to the quality of these features. The detection of features which are robust under geometric or photometric variations is therefore crucial. Moreover, the features should be informative in terms of encoding relevant information about the motions of interest.

In this chapter, we break down the salient feature extraction into feature detection (Section 3.1.1) and feature description (Section 3.1.2). The main contribution of this chapter is the introduction of an improved temporal filtering approach for local motion detection from which more robust to geometric deformations and more informative (i.e., leading to better classification results) local salient features are detected in a video. Moreover, a data-driven measure for removing weak features is introduced which replaces the hard thresholding measures.

3.1 Literature review

A salient feature should be distinct in both space and time. For example, a corner is distinct in space as it has high derivative value in both horizontal and vertical direction. The change in motion direction (i.e., high temporal derivative) makes a point distinct in time. Moreover, a robust salient feature should be re-detected under geometric deformations such as view/scale changes or affine transformation [68, 90]. Salient features such as the start/stop of an action in a video [67] (Fig. 2.3) are examples of such low level visual cues which are widely used

for a compact representation of human motions in a video. A salient feature extraction method consists of feature detection and feature description. The detection component localizes where the salient feature occurs in the video. The description characterizes the appearance and motion information of the feature. The descriptor is chiefly used for the correspondence (e.g., in object recognition [84, 89]) and it is useful for action recognition correspondence as well [68]. To account for the variations in temporal duration and frequency of a unique action performed by different subjects, salient features should be detected at different spatial and temporal scales.

3.1.1 Multi-scale salient feature detection

A multi-scale salient feature detector performs three main steps: (1) spatio-temporal multi-resolution representation of the video signal, (2) saliency map construction, and (3) non-maxima suppression. At a given scale, the local maxima in a spatio-temporal local volume of the saliency determines where a salient feature occurs in space and time. Depending on the type of the feature of interest, existing detectors incorporate different saliency structural or motion maps (Section 3.4). Most of the spatio-temporal feature detectors are the extension of their spatial salient feature detector and hence, they are looking for salient structures in a video. A 3D Harris corner detector [67], 3D Hessian ellipsoid detector [129], 3D SIFT [6, 34] are examples of the type of features which are directly obtained by extending their 2D counterparts. The saliency map of these feature detectors can thus be referred as "structured-based saliency map". As the main characteristic of the video signal is the motion, Dollar et al. [25] proposed temporal Gabor filtering to detect "cuboids" which occur at locations with salient motion energy (Fig. 3.1). From this perspective, the cuboids detector is an example that uses a "motion-based saliency map" to detect the local spatio-temporal salient features in a video. In the rest of this section, we briefly review some of these feature detectors.

Laptev et al. [68] extended the Harris corner detection from a still image to a video signal to localize moving structures at their moments of non-constant motion. To this end, the original video signal $u(x, y, t)$ is scaled using a spatial Gaussian G_σ and a temporal Gaussian kernel G_τ using the convolution operation $L = G_\sigma * G_\tau * u$. The autocorrelation matrix $A = dL^T \times dL$ is then computed from the spatio-temporal derivative vector $dL = [L_x \; L_y \; L_t]^T$. To compute the second-order approximation for the local distribution of gradients within a spatio-temporal neighborhood, matrix A is integrated within a spatio-temporal Gaussian window.

$$M = G_{2\sigma} * G_{2\tau} * A = G_{2\sigma} * G_{2\tau} * \begin{bmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_yL_x & L_y^2 & L_yL_t \\ L_tL_x & L_tL_y & L_t^2 \end{bmatrix} \tag{3.1}$$

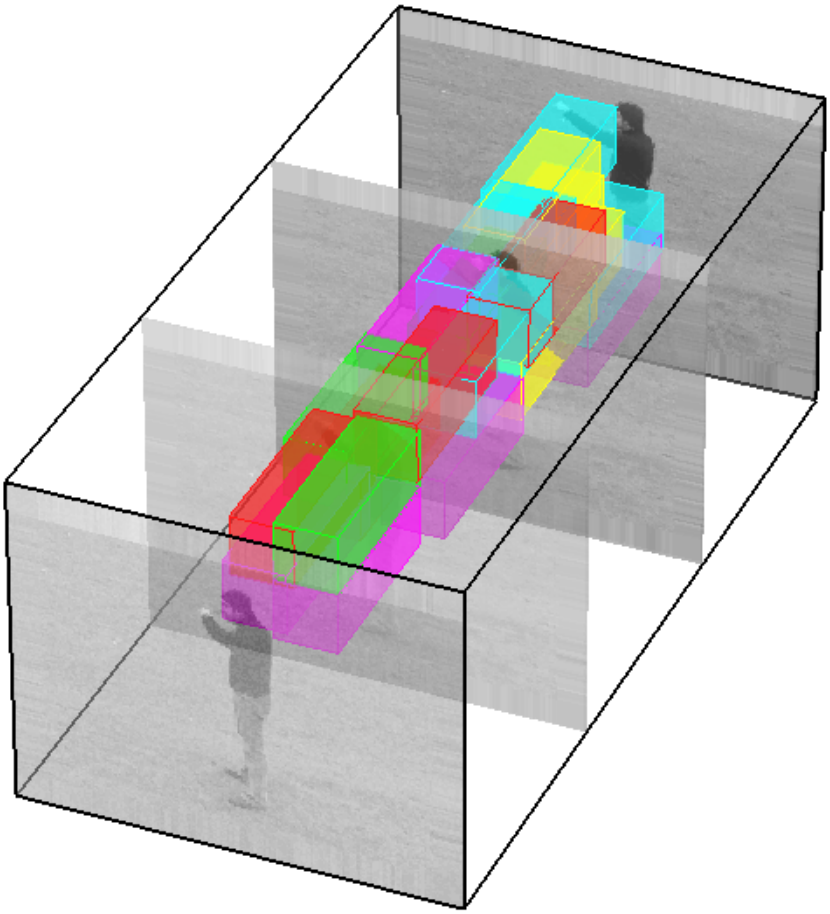


Figure 3.1: Typical "cuboids" features [25] occur when there is a significant motion energy at a point in space and time. The cuboids are then overlapping local space-time volumes in the video.

Using the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of the Harris matrix M, one can compute the spatio-temporal corner map C in which pixels with high level of cornerness are magnified and the rest are suppressed ($\alpha = 0.005$).

$$C = \det(M) - \alpha(trace^3(M)) = \lambda_1\lambda_2\lambda_3 - \alpha(\lambda_1 + \lambda_2 + \lambda_3)^3 \tag{3.2}$$

Another example of extending a 2D salient feature detector to video is the 3D ellipsoid detector. Willems et al. [129] used the trace of a 3D Hessian matrix H of a video signal (3.3) as a saliency map to find where salient features occur. Points with the highest level of saliency in their spatio-

temporal neighborhood are selected as the center of the salient ellipsoids.

$$H = \begin{bmatrix} L_{xx} & L_{xy} & L_{xt} \\ L_{yx} & L_{yy} & L_{yt} \\ L_{tx} & L_{ty} & L_{tt} \end{bmatrix} \tag{3.3}$$

A widely used example of motion-based salient features is the cuboids [25] which occur at points with salient motion energy. To this end, the video signal is first smoothed with a Gaussian filter with standard deviation of σ. The smoothed video is then filtered with a Gabor filter with quadrature pair of even K_τ^{even} and odd K_τ^{odd} at the temporal scale of τ. The energy of the filtered video $R(x, y, t; \sigma, \tau) = (G_\sigma \star q_\tau^{even} \star u)^2 + (G_\sigma \star q_\tau^{odd} \star u)^2$ is the motion saliency map. Searching in a local spatio-temporal window (x, y, t) of the motion map provides the local maxima which identifies the location of the salient features in the video.

3.1.2 Feature description

A feature descriptor characterizes a salient feature using the neighbor pixels in its scale extension. As the descriptor is chiefly used for the correspondence, it should be informative, distinctive, and robust to illumination changes, noise, and rotation [89]. There are several spatio-temporal feature descriptors for encoding the shape and motion information. Descriptors carry the shape and motion changes of the salient features in their spatio-temporal extensions. For example, the Histogram of Optical Flow (HOF) [73] gives the motion change, the Histogram of Oriented Gradient (HOG) [22] provides the spatial appearance changes, local jets (n^{th}-order derivative) encode different spatial and temporal gradients around the feature points [50], and the trajectory of features provides the temporal pattern of movement [98]. Laptev and Lindeberg [68, 69] compared different feature descriptors and their combinations. They concluded that an appropriate feature descriptor should carry both shape and temporal information. Scovanner [111] extended the spatial SIFT descriptor to video and showed that the 3D SIFT descriptor as a spatio-temporal histogram of oriented gradients outperforms the stack of just the spatial histograms or just the temporal histograms. In this section, we explain the 3D SIFT descriptor as it has been shown to be more effective relative to the other descriptors [111, 122].

The 3D SIFT descriptor encodes the shape and motion using the spatial and temporal orientated gradients. A 3D volume of the salient feature is partitioned into $4 \times 4 \times 4$ sub-windows. The orientation of the spatio-temporal gradients is utilized to compute the histogram of each sub-window. The histogram of sub-windows are then concatenated. To reduce the effect of illumination change, the contribution of each orientation in the histogram is weighted by its magnitude. Moreover, the orientations are aligned by the dominant spatial and temporal orientations.

A similar approach has been used in the 3D Histogram of Oriented Gradient (3D HOG [57]) with overlapping sub-windows.

For simplicity in presentation, Fig. 3.2 shows the division of spatio-temporal window into $2 \times 2 \times 2$ instead of $4 \times 4 \times 4$ sub-windows in the actual implementation by Scovanner et al. [111]. Each pixel has a magnitude m_{3D} and two spatial and temporal orientations (θ, ϕ). Any pixel (x', y', t') in spatio-temporal window around the center point (x, y, t) weighs the histogram by its 3D magnitude of gradient (m_{3D}) and a Gaussian kernel centered in the feature position (3.7). The peak of the histogram gives the dominant orientations (θ_{max}, ϕ_{max}). For rotation invariancy, the orientation of each pixel is deducted from the dominant orientation before applying its contribution in the histogram. A more precise version of this descriptor which more accurately computes the 3D orientations has been introduced in [6, 34].

$$m_{3D}(x, y, t) = \sqrt{u_x^2 + u_y^2 + u_t^2}, \tag{3.4}$$

$$\theta(x, y, t) = \arctan\left(u_y / u_x\right), \tag{3.5}$$

$$\phi(x, y, t) = \arctan\left(u_t / \sqrt{u_x^2 + u_y^2}\right) \tag{3.6}$$

$$hist(i_\theta, i_\phi) = m_{3D}(x', y', t') e^{\frac{-((x-x')^2+(y-y')^2+(t-t')^2)}{2\sigma^2}} \tag{3.7}$$

In the next section, we explain different processes required for feature detection in more detail.

3.2 Salient motion feature detection

Salient motion features such as cuboids have shown better performance compared to structured-based features such as 3D corners for action representation [25, 82, 95, 122]. From now on, the focus in this thesis will be on the components of this motion feature detector that can be used as a baseline. Fig. 3.3 shows the operations required for local salient motion feature detection. In this framework, more specifically, we focus on the multi-resolution temporal filtering from which the motion saliency map is constructed. The existing methods are limited to performing a symmetric non-causal 3D video filtering, while we promote the use of asymmetric and causal temporal filtering for better feature detection. Asymmetric temporal filtering is consistent with the physiology and functionality of the visual receptive fields [35, 115]. In the following sections, we explain different modules of a local motion feature detector, namely the scale-space video filtering in Section 3.3, the saliency map in Section 3.4, and the non-maxima suppression in Section 3.5.

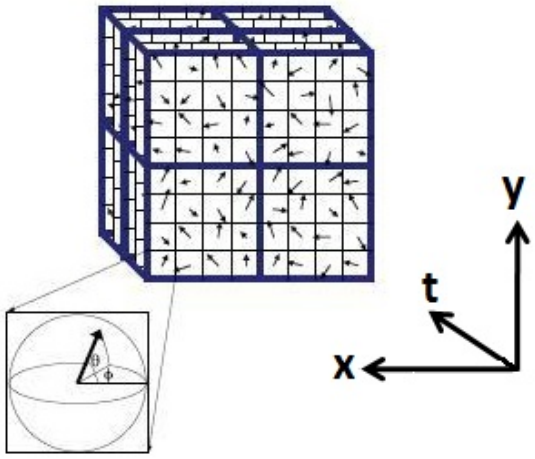


Figure 3.2: 3D SIFT descriptor computation [111]. The spatio-temporal volumetric extension of the features is divided into $4 \times 4 \times 4$ sub-windows. For simplicity, this figure shows the $2 \times 2 \times 2$ sub-windows. A histogram of the spatio-temporal oriented gradients for each sub-window is computed. The histograms are then concatenated in order to construct the final 3D SIFT descriptor.

3.3 Scale-space video filtering

Scale-space theory provides a multi-resolution representation of a signal to detect features at different sizes, from the pixel level to the semantic level [126]. The Gaussian kernel is the standard choice for linear scale-space construction for a still image [8]. In this construction, the structures of the image are more smoothed with the increase in the scale size. In this context, scale refers to the window through which the signal is observed (e.g., the standard deviation of the Gaussian kernel). Gaussian smoothing is consistent with the functionality of the human's early vision system for a spatial multi-resolution representation [75, 76, 77, 130].

A scale-space representation of a video signal can be obtained using linearly separable spatial and temporal scale-space filtering. In the literature of spatio-temporal salient feature detection, the norm is to use 2-D Gaussian filters [84, 90] with different standard deviations in the spatial directions. Temporally, a variety of approaches have been used. For example, for 1-D temporal filtering, Laptev et al. [68] used a Gaussian, Willems et al. [129] used a Gaussian approximation, and Dollar et al. [25] used a Gabor filter. These temporal filters are all non-causal and symmetric. However, biological vision has demonstrated that a motion perception model with causal and asymmetric temporal filtering [1, 35, 109, 115, 125] is more consistent with the electrophys-

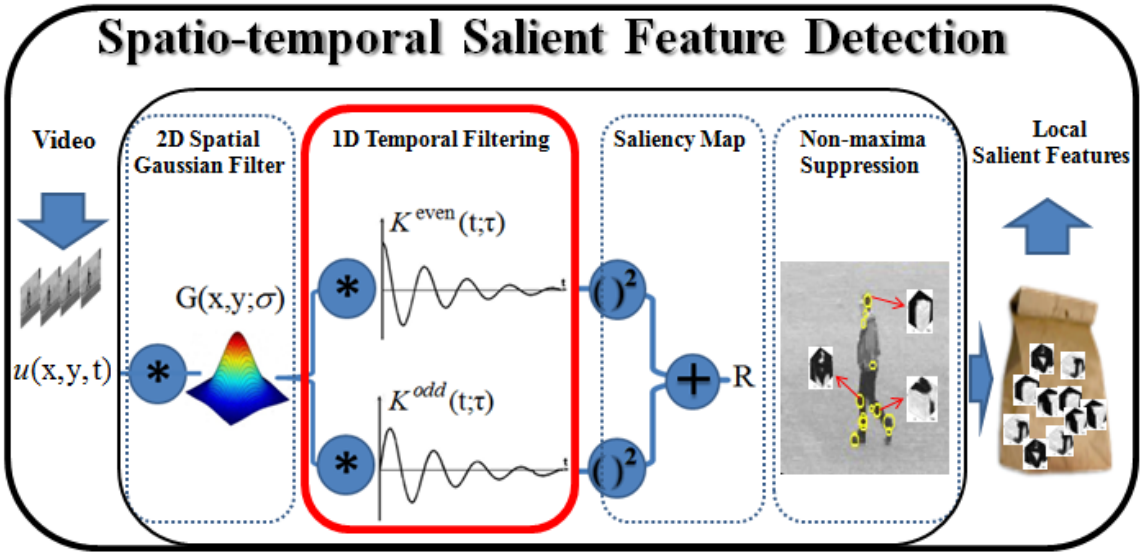


Figure 3.3: Salient motion feature detection procedure. The input to the system is a video and the output is a set of local salient motion features. A typical detector filters the video signal $u(x, y, t)$ by a spatial Gaussian filter $G(x, y; \sigma)$ and quadrature pair temporal filters. The local maxima in the spatio-temporal energy map R represent local salient features. Section 3.3 focuses on different temporal scale-space filters (solid red box).

iological data of the visual sensitivity to motion and the physiology of the LGN cell and the cortical V1 simple cell receptive fields [115]. Asymmetric temporal kernels has been proven to be more efficient than a symmetric kernel for video segmentation as well [123]. In scale-space theory of computer vision, also, several time-causal scale-space kernels have been theoretically developed [30, 59, 81, 115], but have not yet been utilized by the action recognition community.

To clarify the terminology, from now on, filtering is referred to as the operation in which the signal is convolved with a kernel. In this context, the kernel is a weighting function with different compact supports. In the following sections, we first explain the spatial Gaussian scale-space theory and then, we introduce both non-causal symmetric and asymmetric causal temporal kernels. We argue that even if the delay of a symmetric filtering is not an issue, the question of proper temporal filtering and the choice between asymmetric versus symmetric filtering should be investigated. We hypothesize that due to consistency with the human visual system, the asymmetric causal filters should capture more relevant features for the representation of human actions in video and, and hence, provide better video filtering for salient feature detection. Our experiments in Section 3.7 support this hypothesis.

3.3.1 Spatial scale-space filtering

A physical process that can explain scale-space filtering is a diffusion equation (3.8) with a (time-like) scale variable s and a diffusivity g (in general, g can be a tensor [126]). The diffusion equation is obtained from the continuity equation stating that evolution of the image throughout the scales is due to the divergence of the gradient's concentration (∇u). With a constant diffusivity, the laplacian of the image (Δu) provides a linear diffusion equation. In fact, the linear smoothing (3.9) of an image $u_0(x, y)$ at different scales s can be obtained by its convolution ($*$) with a Gaussian kernel (3.10) with varying standard deviations $\sigma_s = \sqrt{2s}$ (the pixel location is denoted by (x, y)).

$$\partial u/\partial s = div(g.\nabla u) = g\Delta u \tag{3.8}$$

$$u(x, y; s) = (G_{\sigma_s} * u_0)(x, y) \tag{3.9}$$

$$G_{\sigma_s}(x, y) = \frac{1}{2\pi\sigma_s^2}\, e^{-\frac{x^2+y^2}{2\sigma_s^2}} \tag{3.10}$$

In addition to physical diffusion-based construction of spatial scale-space, Gaussian scale-space has also been extracted using an axiomatic approach [77]. These axioms require the continuity of the scale-space representation, its positiveness, translation invariancy, and scaling covariance. Moreover, the local maxima (zero-crossing) should not be created as the scale increases [126]. This constraint of Gaussian scale-space representation assures a unique correspondence between the structures in the higher scales to the finer scales.

Duits et al. [27, 28] followed the axiomatic approach of the spatial scale-space theory and showed that the Poisson scale-space is an alternative to the Gaussian scale-space theory. In fact, Poisson and Gaussian scale-spaces are related by an intermediate representation of the fractional power of (minus) the spatial Laplacian operator (3.11). The free parameter α in this representation examines the nature of the distribution that a particle undergoes to redistribute its energy with its neighbors. In the case of $\alpha = 1$, the Gaussian scale-space and in the case of $\alpha = 1/2$, the Poisson scale-space can be derived (note that for the simplicity, here the conduction g is normalized to one).

$$\partial u/\partial s = -(-\Delta)^{\alpha} u \tag{3.11}$$

Felsberg et al. [31, 32] developed the (complex) monogenic scale-space in which the real part contains the Poisson scale-space and the imaginary part is the flow of the image. Gilboa et al. [39] was inspired from the fusion of (imaginary-valued) Schrodinger's equation of a free particle with a (real-valued) diffusion equation to provide another type of complex scale space. In this case, a complex-valued conductivity resulted in the presence of Gaussian scale-space in

the real part and a (scaled) Laplacian of Gaussian in the imaginary part. In a more general case, the Laplacian Beltrami operator can better capture the variations on the manifold and hence it is useful for structure preserving multi-resolution filtering and color image processing [55, 54, 26].

In the literature of salient feature detection in a still image or in a video, the Gaussian scale-space filtering has been widely used [68, 77, 84, 129]. The other above-mentioned scale-space filters found their application for image enhancement or denoising in which a non-linear and an anisotropic diffusion filtering is performed [126]. As we are interested in the detection of local motion features for which proper temporal filtering is crucial, we restrict ourself to use the Gaussian filtering for spatial scale-scale filtering and focus on the choice of proper temporal scale-space filtering. One research direction is to evaluate the performance of other spatial scale-space filtering for the feature detection which is beyond the scope of this thesis.

3.3.2 Temporal scale-space filtering

For the temporal filtering of a video signal, one can use a non-causal symmetric or a causal asymmetric kernel. In this section, we review existing non-causal kernels and three asymmetric causal kernels from scale-space literature for salient feature detection. We then develop a brand new asymmetric time-causal filtering using circuit theory. The objective is then to examine the role of causality and asymmetry in temporal filtering of a video for salient feature detection. We hypothesize that due to consistency with the human visual system, the asymmetric causal filters should capture more relevant features for the representation of human motions in video and hence, provide better video filtering for salient feature detection.

We highlight three research questions regarding the role of causality and symmetry of the temporal filtering in key point detection and action classification in video.

1. How does the symmetry/asymmetry of the temporal filter affect the performance of salient feature detection? What is the role of the causality of the temporal filter in salient feature detection? Precision tests are performed in Section 3.7.2 to answer these questions.

2. How is the quality of salient features using different temporal filters affected by geometric changes? This question is addressed using reproducibility tests in Section 3.7.3.

3. How successful are salient features generated using different temporal filters for action classification? This research question is examined in Section 3.7.4.

Symmetric non-causal kernels

The isotropic diffusion for temporal smoothing of a video requires access to both past and future frames. Similar to the spatial domain, the temporal scale-space is obtained by convolving the signal with a temporal Gaussian kernel (3.12) with a scale of τ.

$$G_\tau(t) = \frac{1}{\sqrt{2\pi}\tau}\, e^{-\frac{t^2}{2\tau^2}} \tag{3.12}$$

Laptev et al. [70] and Willems et al. [129] used this Gaussian filtering to extract spatio-temporal Harris and Hessian features at multiple scales. Dollar et al. [25] used a temporal Gabor filter to capture motions at a given frequency. In the frequency domain, the Gabor filter is the shifted version of a Gaussian to a central frequency w_0 and hence, it is a band-pass filter. The shift in the frequency domain is obtained by multiplying the temporal Gaussian kernel with a complex function e^{-jw_0t} resulting in a symmetric even and odd kernels.

$$G_\tau^{even}(t) = cos(\omega_0 t)e^{-\frac{t^2}{2\tau^2}} \tag{3.13}$$

$$G_\tau^{odd}(t) = sin(\omega_0 t)e^{-\frac{t^2}{2\tau^2}} \tag{3.14}$$

Logarithmic Gaussian kernel

To address time causality, Koenderink [59] and then Romeny et al. [115] mapped the time domain to the natural logarithmic domain. The causal interpretation of the time axis is thus obtained by sampling the time in a logarithmic fashion. In this case, the half-line from the past time t to the current time t_0 is mapped to a line by logarithmic transformation. Prior to the logarithmic mapping, the time difference $(t_0 - t)$ is normalized by the time scale τ to make a dimensionless variable μ (3.15). The Gaussian kernel is then applied in the new logarithmic domain (3.16). Note that due to inversion of the time direction in the convolution, a time-causal filter is actually applied to the past frames, and hence, there is no need to introduce a delay.

$$\mu = \ln\left(\frac{t_0 - t}{\tau}\right) \tag{3.15}$$

$$G_\tau^L(t, t_0) = \frac{1}{\sqrt{2\pi}\tau}\, e^{\frac{-\mu^2}{2\tau^2}} \tag{3.16}$$

Poisson kernel

Lindeberg and Fagerstrom [81] developed a 2D+t scale-space theory with causal time. The spatial kernel is a Gaussian, but for the time domain, the Poisson kernel (3.17) is suggested as the canonical temporal scale-space of a causal discrete signal. The Poisson kernel is obtained from the cascade of "n" truncated exponentials. The mean and variance of the Poisson kernel are equal to $\lambda = \tau$ and n denotes discrete time. Note that the shape of the Poisson kernel varies by changing the temporal scale and hence, the Poisson is not scale covariant. This can be interpreted as non-stationarity of the kernel in the scale direction.

$$P_\lambda(n) = \frac{\lambda^n e^{-\lambda}}{n!} \tag{3.17}$$

In biological vision, Fourtes et al. [35] modeled the functionality of an axon in the Limulus visual system using a Poisson kernel. The derivative of Poisson kernel has been utilized to fit with the data representing the temporal contrast sensitivity of the human visual system [125] and to compute the well-received opponent-based motion perception model [1].

Scale-derivative Gaussian kernel

Lindeberg [78] and Fagerstrom [29] extended the axioms of spatial scale-space theory to the spatio-temporal domain. They removed the symmetry requirement for the time domain, added a time causality constraint, and kept the other axioms of spatial scale-space theory (i.e., continuity, positivity, non-enhancement of local maxima, translation invariance, and scaling covariance). Fagerstrom [30] then developed a closed-form solution by considering a convenient value for a free parameter α which determines the kernel shape. He then argued that in temporal smoothing, one should use the temporal memory (scale-space) rather than the actual past frames. For the choice of $\alpha = 1/2$, the signaling equation (3.18) as a dual evolution equation for the time-causal diffusion is obtained. The local generator for time-causal scale-space on this new space results in a kernel (3.19) which is proportional to the scale-derivative of a Gaussian kernel.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial \tau^2} \ , \quad \lim_{\tau \longrightarrow 0} u(t,\tau) = u(t) \ , \quad u(0,\tau) = 0 \tag{3.18}$$

$$G^D_\tau(t) = \frac{\tau \ e^{\frac{-\tau^2}{4t}}}{\sqrt{4\pi} \ t^{3/2}} = -2\partial_\tau G_t(\tau) \quad , \ t > 0 \tag{3.19}$$

In contrast to the Poison kernel which is not scale covariant and the log Gaussian kernel which uses the logarithmic time domain (i.e., it is inhomogeneous), the scale-derivative Gaussian uses linear time domain and it is scale covariant.

Asymmetric sinc

We model the scale-space filtering of a 2D+t video signal using circuit theory [35, 41, 38], but considering the time causality. To this end, the circuit representation of an image is extended to a time-causal video signal from which we derive the corresponding spatio-temporal diffusion equation. This formulation gives a brand new asymmetric multi-resolution kernel for time-causal video filtering. The circuit representation has been already used for modeling biological systems and the filter design in image processing. In biological vision literature, Fourtes [35] modeled the functionality of an axon in the Limulus's visual system by its segmentation into several stages. Each stage is then modeled as an RC circuit connected to the next stage by an amplifying or isolating unit. In terms of the electrical analogue, they conclude that there are about ten stages of low-pass filters. In image processing, Grady [41] used the circuit analogy to derive the equation of random walk for image segmentation. Perona and Malik [100] provided a circuit representation for implementation of an anisotropic diffusion on a still image.

In our filter design, a digital image is viewed as a graph network in which each pixel x is a node connected to the neighbor pixel by a resistance R (Fig. 3.4). The brightness u at each pixel is the charge on a capacitor C. The flux of the brightness between each two adjacent pixels depends on their relative brightness. To extend this model to a 2D+t video signal, we consider the effect of the potential of the corresponding pixel from the immediate past frame on the potential evolution at the current frame. This consideration is modeled as an external conductance R_{ext}^{-1}. From Kirchhoff's Laws (3.20-3.23), we can therefore derive the diffusion equation as the change of the brightness at each pixel at a given (time-like) diffusion scale s. The potential at this pixel is denoted by $u(x,t;s)$ and the current by $i(x,t;s)$.

$$
\begin{aligned}
i(x,t;s) &= (u(x,t;s) - u(x-dx,t;s))/R_1 && (3.20)\\
i(x+dx,t;s) &= (u(x,t;s) - u(x+dx,t;s))/R_2 && (3.21)\\
i_{ext} &= (u(x,t;s) - u(x,t-dt;s))/R_{ext} && (3.22)
\end{aligned}
$$

$$
i(x,t;s) + i(x+dx,t;s) + I_{ext} + C\frac{\partial u(x,t;s)}{\partial s} = 0 \qquad (3.23)
$$

Consider $R_1 = R_2 = R$ and the per unit quantities, $R = rdx$, $R_{ext} = r_{ext}dt$, $C = cdx$, and $I_{ext}(x,t;s) = i_{ext}(x,t;s)dx$. Substitute equations 3.20- 3.22 in the KCL equation (3.23). At the limit $dt \longrightarrow 0$ and $dx \longrightarrow 0$, the KCL equation results in the spatio-temporal diffusion equation (3.24).

$$\frac{\partial u(x,t;s)}{\partial s} = \frac{1}{rc}\frac{\partial^2 u(x,t;s)}{\partial x^2} - \frac{1}{r_{ext}c}\frac{\partial u(x,t;s)}{\partial t} \tag{3.24}$$

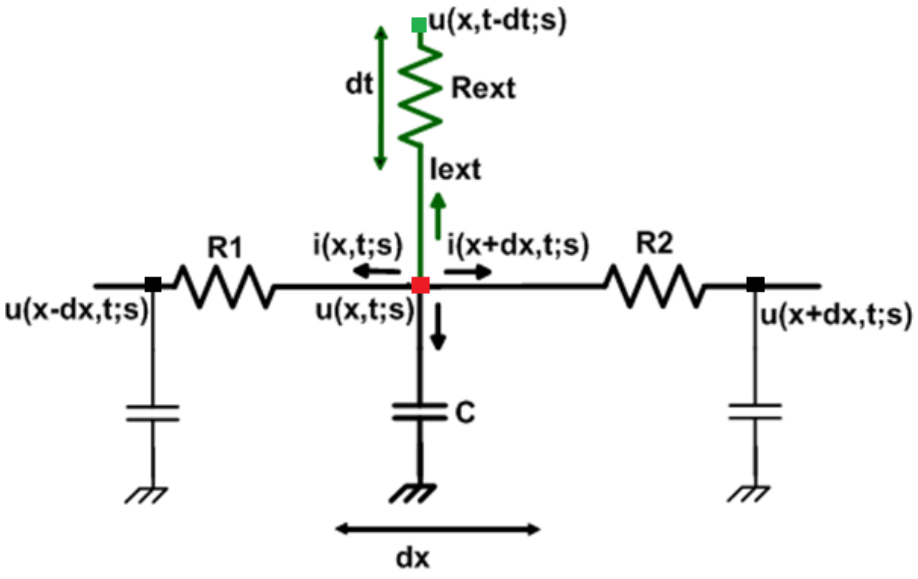


Figure 3.4: Modeling spatio-temporal (1D+t) diffusion using a RC circuit. The brightness $u(x,t;s)$ denotes the potential charge on the capacitor C at node x at the current time t and current diffusion scale s. A pixel is connected to the neighbor spatial pixel by a resistor R and to temporal past neighbor pixel (green resistor) by R_{ext}.

This equation is consistent with the observation of Lindeberg [81] in that the variation of a signal with respect to scale ($\frac{\partial u}{\partial s}$) should be proportional to its second-order derivative in space ($\frac{\partial^2 u}{\partial x^2}$), but first-order derivative in time ($\frac{\partial u}{\partial t}$). For a 2D spatial network of pixels, (x,y), the second-order spatial derivative $\frac{\partial^2}{\partial x^2}$ is replaced with a 2D Laplacian $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$.

To obtain a closed form solution for the equation (3.24), we consider constant value conductances and capacitance. This consideration results in a linear diffusion. However, the spatial diffusion is isotropic while the temporal diffusion is anisotropic due to the prohibition to dif-

fuse to the future. Moreover, there is a linear interaction between internal forces (i.e., spatial concentration gradients) and external forces (i.e., temporal gradient).

The solution of the equation (3.24) is obtained using the Fourier transform. Consider $\widetilde{U}(w_x, w_y, w_t; s)$ as the (discrete-time) Fourier transform of $u(x, y, t; s)$ with respect to pixel coordinates $X = (x, y, t)$, in which (w_x, w_y, w_t) are the corresponding spatial frequencies and temporal frequency. Consider $\alpha = 1/rc$ and $\beta = 1/r_{ext}c$ which are related to the diffusivity in space and time.

1. Apply a Fourier transform to the equation (3.24).

$$\frac{\partial \widetilde{U}}{\partial s} = [-\alpha(w_x^2 + w_y^2) - \beta(jw_t)]\widetilde{U} \tag{3.25}$$

2. Solve the ordinary differential equation knowing that $\widetilde{U}(w_x, w_y, w_t; 0) = \widetilde{U}_0$ is the Fourier of the current original image at the finest scale $u(x, y, t; 0) = u_0$.

$$\begin{aligned} \frac{d\widetilde{U}}{\widetilde{U}} &= [-\alpha(w_x^2 + w_y^2) - j\beta w_t]ds & (3.26)\\ \widetilde{U} &= [e^{-\alpha[w_x^2+w_y^2]s} e^{-j\beta w_t s}]\widetilde{U}_0 & (3.27) \end{aligned}$$

3. Using the convolution theorem, the inverse Fourier transform of the equation (3.27) provides the convolution ($\star$) of the original image sequence u_0 with the (discrete analogue of) spatial Gaussian kernel G with standard deviation $\sigma = \sqrt{2\alpha s}$ and a causal asymmetric sinc as the temporal kernel $q(t; \tau)$ with temporal scale $\tau = \beta s$. The $S(t)$ denotes the Heaviside step function.

$$u(x, y, t; \sigma, \tau) = G \star q \star u_0 \ , \ \ G(x, y; \sigma) = \frac{e^{-\frac{x^2+y^2}{2\sigma^2}}}{2\pi\sigma^2} \ , \ \ q(t; \tau) = sinc(t - \tau)\ S(t) \tag{3.28}$$

The spatial Gaussian kernel is symmetric and non-causal and the temporal kernel is an asymmetric sinc function due to causality and shift of the peak to the temporal scale τ. Note that if we release the causality constraint and allow the diffusion to the future frames as well, we will have the second-order temporal derivative in equation (3.24) and hence, a temporal Gaussian kernel.

An insight into the temporal part of the Fourier transform in (3.27) suggests that if we consider the stability criteria of $|w_t\tau| < 1$, we can actually use the Taylor approximation of

$e^{-jw_t\tau} \sim 1/(1+jw_t\tau)$ and hence, approximate the temporal kernel with a truncated exponential function ($E(t,\tau) = e^{-t/\tau}S(t)/\tau$) for a continuous signal. The stability constraint shows that for small time scales, higher temporal frequencies are captured and vice-versa. As explained in [81], the truncated exponential is the base for the derivation of the Poisson kernel (3.17).

To detect salient features, after scale-space filtering, a saliency map should be computed for each spatio-temporal scale. Despite the linearity of scale-space video filtering, the saliency map construction usually requires non-linear filtering. The local maxima in the saliency map are the salient features (Fig. 3.3). In the next sections, we explain the motion saliency map and introduce a data driven approach to retain the stronger salient features after non-maxima suppression.

3.4 Saliency map

The saliency map is constructed based on the characteristics of the feature of interest. From a biological point-of-view, a saliency map is the response of the video signal to a local filter applied to surrounding volume of a pixel in order to mimic the functionality of the receptive field of the early visual cortex [37]. From biological vision [1, 125, 112], a motion map should be phase insensitive to provide a response independent of the phase of the motion stimulus and be constant for a periodic motion. To obtain a phase insensitive motion map, we perform quadrature pair filtering for motion computation. The quadrature can be seen as the pairs of adjacent simple cells in the visual cortex [103]. A complex receptive field can therefore have these quadratures in the real and imaginary parts. Moreover, the motion response should be the same for two stimuli with opposite contrast but similar motion [1]. The latter requirement ensures a similar response for a white foreground against dark background and vice versa. To obtain a contrast-polarity insensitive motion map, we consider the magnitude of the motion response.

3.4.1 Phase insensitivity

To obtain a phase-insensitive motion response, we have two options: complex quadrature pair filtering and derivative filters, which are considered to be quasi-quadrature. There are two reasons that the complex quadrature filters are preferred over the quasi-quadrature derivative filters: (1) the complex quadrature pair filters provide better localization than the derivative filters [58]; and (2) The complex quadrature pair filtering is straightforward by multiplying the original kernel with a complex exponential term of e^{-jw_0t}. This is the way Gabor filter is obtained from the Gaussian. In contrast, for the computation of derivatives of an asymmetric filter, one should take into account the shape of the filter [79]. According to Lindeberg [79], it is more natural to

compute the derivatives of a time-causal kernel with respect to a transformed axis (3.29). This is due to skewness of a time-causal kernel.

$$t' = \phi(t) \tag{3.29}$$

The transformation function ϕ should be a monotonically increasing function such as a self-similar logarithmic transformation (3.30) or a self-similar power law (3.31).

$$t' = log(t/t_0) \tag{3.30}$$

$$t' = t^{\alpha} \tag{3.31}$$

Figure 3.5 shows the temporal smoothing kernels with their first- and second-order derivatives as reported by Lindeberg [79] (Fig.15). Note that the first-order derivatives (second column) have two peaks and the change of the time variable does not change the position of the peak (third and fourth rows in the second column). However, the second-order derivative changes qualitatively. In fact, the second-order derivative with respect to the original time axis (last column of the second row) has two peaks and one interior zero crossing, while its counterpart with the time transformed variable makes three peaks with two interior zero-crossings (the last column of the third and fourth rows) which shows similar behavior as a second-order derivative of a scale-space kernel (last column of the first row).

Due to the simplicity and better localization of the quadrature pair filtering, we consider this filtering scheme to address the phase insensitivity requirement for the motion computation. The quadrature filters for each of the temporal kernels (q^{even} and q^{odd}) provide a band-pass filter centered at $w_0 = 1/2\tau$. For the asymmetric sinc, we however use the Hilbert transform $h(t) = \frac{1}{\pi t}$ [125] to compute its quadrature pair $q_h(t;\tau) = q(t;\tau) \star h(t)$ due to the derivative nature of this filter.

Fig. 3.6 shows the quadrature pair of the Gabor and the complex scale-derivative Gaussian, log-Gaussian, Poisson, and asymmetric sinc. These complex filters are used for multi-resolution representation of the video and hence feature detection (See Fig. 3.3).The even and odd Gabor filters are symmetric and non-causal, while the other filters are causal and asymmetric. Note that the Poisson tends to decay faster and the scale-derivative Gaussian has the longest tail. The magnitude of the frequency response of these filters are shown in Fig. 3.7. Note the different shape of these filters; the asymmetric filters have wider bandpass than the symmetric Gabor. The scale-derivative Gaussian G^D has the sharpest decay from the central frequency. The asymmetric sinc is the closest filter to an ideal band-pass filter.

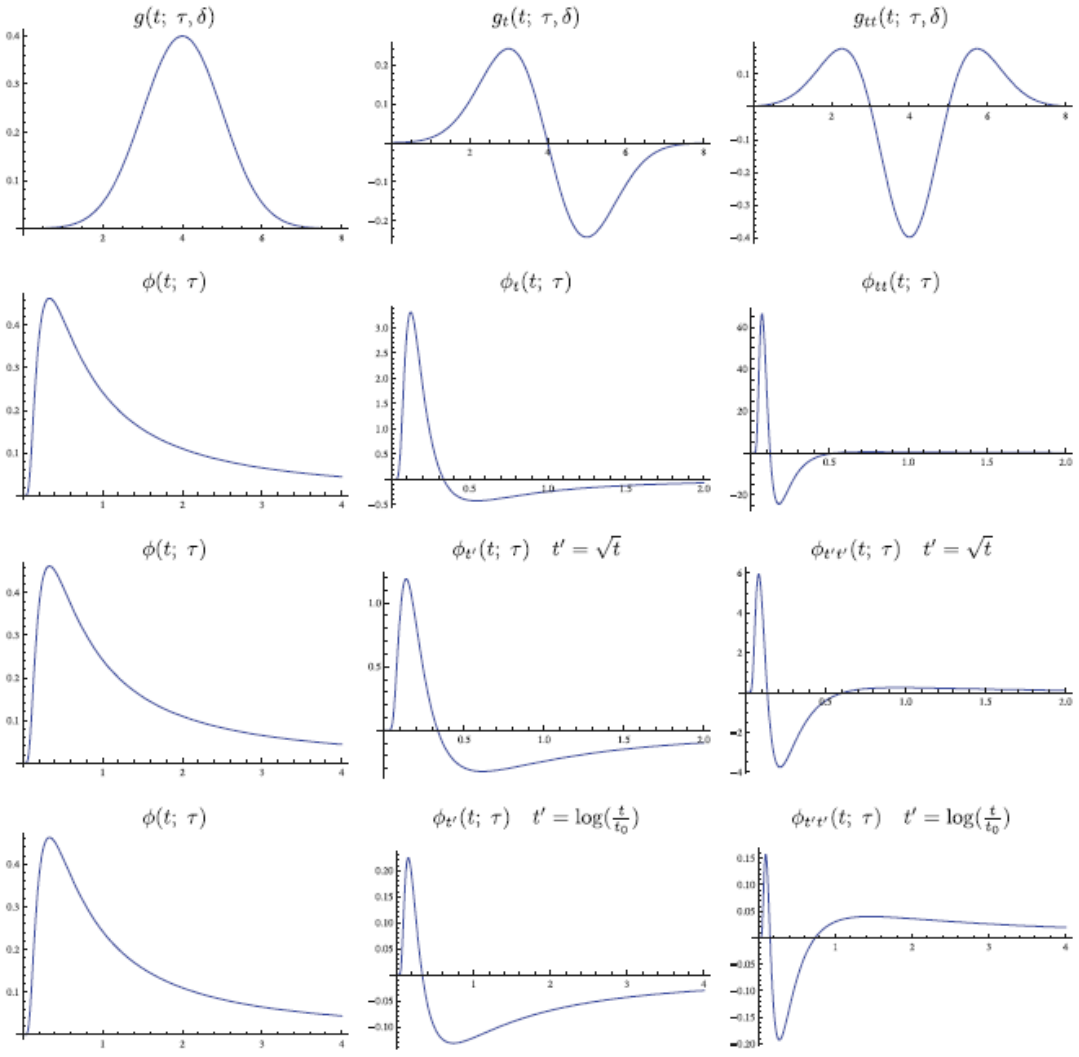


Figure 3.5: The temporal smoothing kernels with their first- and second-order derivatives as reported by Lindeberg [79] (Fig.15). In figure, same function $\phi(t;\tau)$ is used to describe each of the scale-derivative Gaussian functions. The first row is the time-shifted Gaussian and its derivatives. The scale-derivative Gaussian kernel and its derivatives are plotted in the second row for the original time variable and its derivative with respect to the self-similar transformed time are plotted in the third and fourth rows.

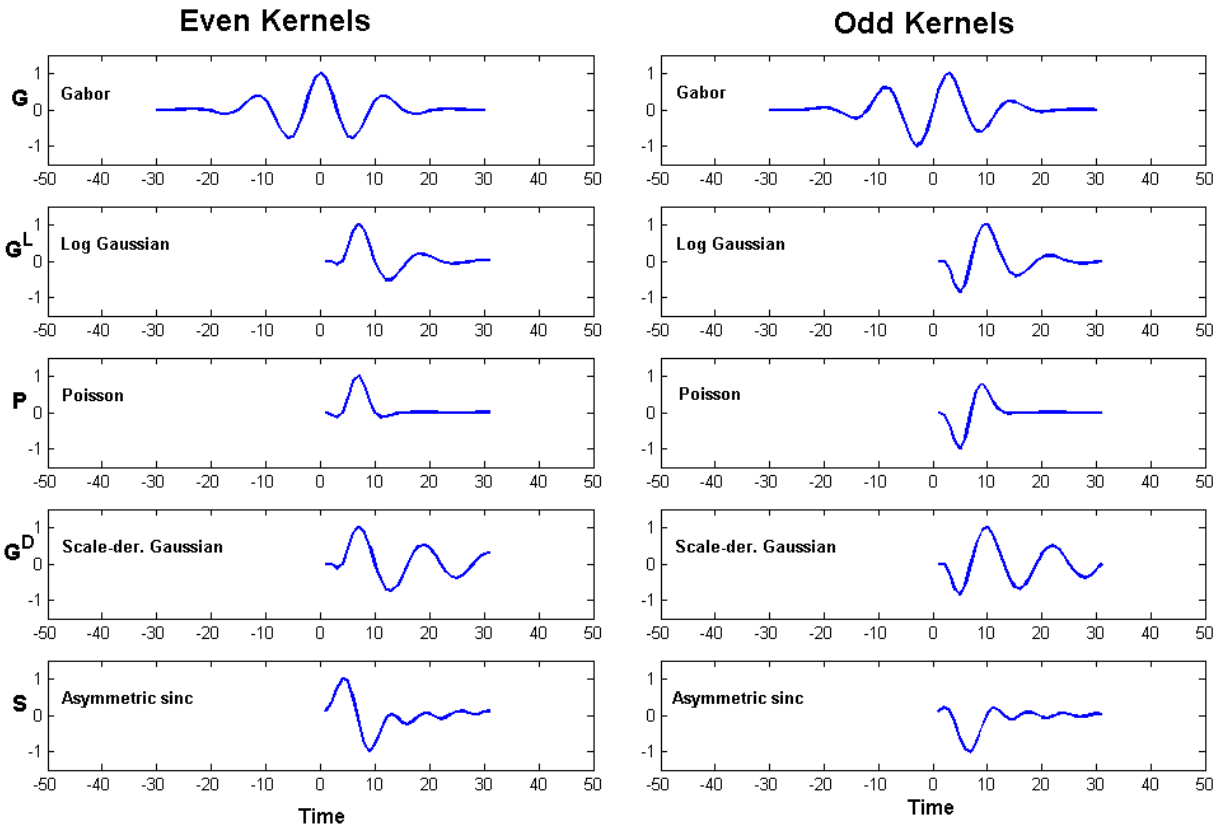


Figure 3.6: Quadrature pair of Gabor G and complex form of log-Gaussian G^L, Poisson P, scale-derivative Gaussian G^D, and asymmetric sinc S filters at temporal scale of $\tau = 6$. The left column shows the even filters and the right column shows the odd filters. The horizontal axis shows the time. Only the Gabor filters are symmetric and non-causal. The causal filters are asymmetric.

Note that there is no guarantee that a subject performs with the same motion pattern in different periods of a given action such as walking. The imperfection of the motion pattern requires a reasonable band-pass filter with high weight at the central frequency, but at the same time a longer tail so that it can capture more neighbor frequencies. This requirement is in support of causal asymmetric kernels which have a longer tail in the frequency domain and show this behavior better than the non-causal symmetric Gabor. A longer tail band-pass filter does not attenuate the signal as much at higher frequencies or passes more energy and hence, it results in

less smoothing of fast motions.

3.4.2 Contrast-polarity insensitivity

To obtain a contrast-polarity insensitive motion map, we use the sum of the square filter responses of the quadrature filters which corresponds to the local energy of the video signal in the spatio-temporal extent of the filters.

The salient motion map from which feature are detected is defined as $R = R(x, y, t; \sigma, \tau) = (G_\sigma \star q_\tau^{even} \star I)^2 + (G_\sigma \star q_\tau^{odd} \star I)^2$ for the spatio-temporal scale of (σ, τ). This energy-based saliency map highlights the points with local periodic motion and hence, it is more suitable for action-related salient feature detection [25, 122].

3.5 Non-maxima suppression

As the local salient features are centered at point with high level saliency, we need to perform non-maxima suppression on the saliency map to localize the salient features (Fig. 3.3). To this end, the neighborhood of each pixel in all spatio-temporal directions is searched to find points with the highest saliency value. These points localize the salient fetaures. Having detected the local maxima, existing methods [25, 68] usually apply a hard thresholding on the local maxima to remove the weak noisy local maxima and, consequently, retain stronger salient features.

Instead of hard thresholding to remove weak features, a data-driven measure for removing the noisy local maxima is introduced here. In the robust estimation and anisotropic diffusion filtering literature [10], this threshold is estimated using the median absolute deviation (MAD) for statistical noise removal. This operation is performed in two steps. First, those maxima that have an energy level higher than the MAD of the energy of the whole video R (i.e., MAD_R in equation 3.32) are kept as initial salient features. Second, another level of thresholding is performed based on the MAD of the saliency values of the initial salient features to remove weaker salient features. Once the salient features are detected, due to the peak shift in the causal kernels, the central time at which salient features occur is compensated accordingly. That is, if the peak of the kernel is at τ_0, the time coordinate of the salient features is deducted by this amount. This compensation addresses the illusion in the motion map [33] which is the result of having a peak not at the origin.

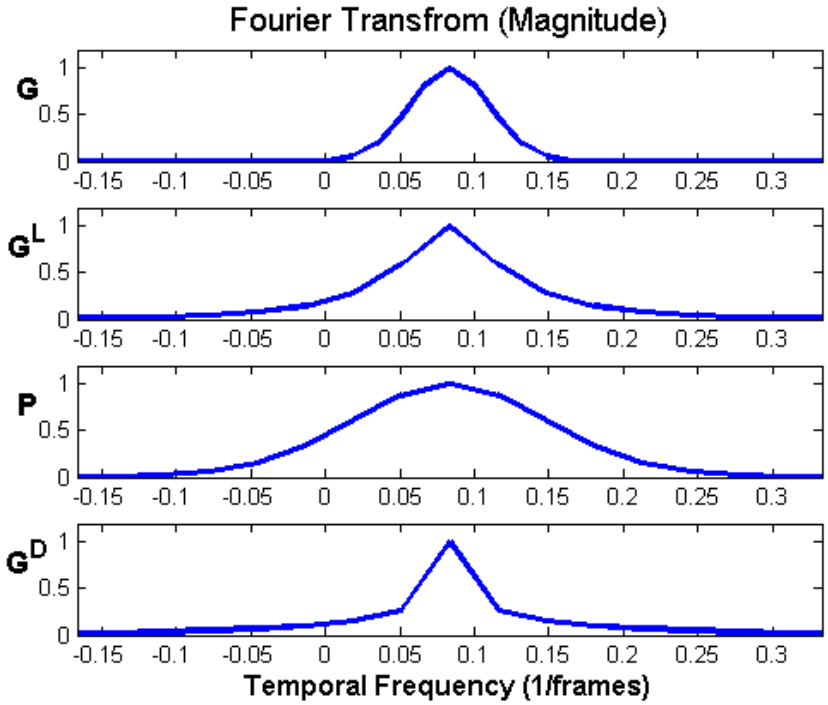


Figure 3.7: Magnitude of the Fourier transform of the complex form of the four temporal filters (Gabor G, log-Gaussian G^L, Poisson P, and scale-derivative Gaussian G^D) with $\tau = 6$ and $\omega_0 = 1/2\tau = 0.083$. Note that the three asymmetric filters (G^L, P, G^D) have wider bandpass than the symmetric Gabor filter G.

$$MAD_R = \alpha \;\; median(|R| - median(|R|)) \tag{3.32}$$

The constant value $\alpha = 1.4826$ normalizes the MAD to that of a normal distribution with zero-mean and unit variance for which $MAD = 1/\alpha$ [10].

3.6 Scale-invariant features

The scale-invariant spatio-temporal salient features [68, 129] were of interest due to the success of their counterparts for object encoding in a still image. In fact, in the absence of any prior knowledge about the size of the objects in an image, the scale-invariant features are the local visual cues that are utilized to match a query image with a test image under a possible scale

change. Scale-invariant salient features [84] are widely used in image analysis for applications such as object detection or object recognition. Detection of an scale-invariant feature requires an intrinsic scale estimation in which the feature is quite distinctive from its neighbors.

For video analysis, different approaches have been used to determine the intrinsic scale of the features. A straightforward way of detecting scale-invariant features is to extend the search domain of the feature detection to the scales direction as well [84, 129]. For spatio-temporal scale-invariant features, the search (i.e., non-local maxima suppression) should thus be performed in five directions (x, y, t, σ, τ). One might use other measures such as the Laplacian which is the trace of Hessian matrix [129] as a measure to compute the intrinsic scale of features. In this case, the intrinsic scale is the scale in which the salient feature has the maximum Laplacian value. The scale-invariant features provide a sparse representation of video contents, but they were not successful for discriminative representation of different motion patterns [88, 129]. In fact, as the scale selection approaches are heuristic [88], some of the motion information is thrown away. Moreover, the removal of motion redundancy from the set of features for action representation might not necessarily be useful for their discrimination [45, 129]. A unique representation might however require to retain all the salient features (Section 4.3).

3.7 Experiments

In this section, the methodology used to evaluate the performance of different temporal filters in local feature detection is described. The evaluation is based on three quantitative measures: precision score, reproducibility score, and action classification accuracy. The precision score and reproducibility score show the accuracy and robustness of the detected salient features and the action classification accuracy determines how informative the salient features are for action representation.

3.7.1 Methodology

To examine the performance of different temporal scale-space filters for salient feature detection, the set of spatio-temporal scales, the dataset, and the salient feature matching used in the quantitative measures must be defined.

Spatio-temporal scales

To experiment with the salient feature detection performances, a set of spatial ($\sigma_x = \sigma_y = \sigma_i$) and temporal scales ($\tau_i$) must be defined. We consider nine different spatio-temporal scales in which σ is in *pixels* and τ is in *frames*. The scale generation formula (3.33) is similar to the approach used by 2D SIFT [84]. Note that the minimum spatial scale $\sigma_0 = 2$ *pixels* determines the maximum spatial frequency of 0.5 *cycles/pixel*. With 25 *frames per second*, the maximum temporal frequency of 12.5 *cycles/sec* is obtained at $\tau = 2$ *frames*.

$$\sigma_i = (\sqrt{2})^i \, \sigma_0 \ , \quad \tau_i = (\sqrt{2})^i \, \tau_0 \ , \quad i \in \{0, 1, 2\} \tag{3.33}$$

Data sets

To provide quantitative results, video samples with ground truth are required. More specifically, segmented videos of the same action taken under different views, clutters, and occlusions for the precision and robustness dataset are required. For the evaluation of the feature detectors, we used the robustness and classification datasets of the Weizmann dataset [11] for which the ground truth has been provided. For the action classification test, we used the Weizmann classification dataset, the KTH [110] dataset, and the UCF sports dataset [107].

The **Weizmann robustness dataset** [11] consists of two sets: "deformations" and "viewpoint". The former set consists of ten videos of "walking" with high variations in the motion pattern in the presence of clutter and occlusion with a non-uniform background. The "viewpoint" set is the "walking" action performed at different horizontal views from the camera (0^o to 81^o with 9^o intervals). Figures 3.8 and 3.9 show sample frames from these sets. The tests on these videos can show the performance of different temporal filters when there exists partial clutter, occlusions, and non-rigid deformations. This dataset is a relatively easy dataset for the classification task and perfect classification on this dataset has been reported already. We use this dataset as part of our proof for the concept.

The **Weizmann classification dataset** [11] consists of ten different human actions: run, jumping-jack (jack), jump-forward-on-two-legs (jump), jump-in-place (pjump), wave-two-hands (wave2), wave-one-hand (wave), gallop sideways (side), bend, skip, and walk. Each action is performed by at least nine different people in front of a fixed camera. Each of the 93 video clips lasts about two seconds at $25Hz$ with an image frame size of 180×144. Fig. 3.10 shows a sample frame from each action. Having different action types by different people in this dataset, the performance of different temporal filters for the salient feature detection and action classification

can be evaluated. For the action classification, a leave-one-out approach is considered for the training and testing.

The **KTH data set** [110] consists of six actions (running, boxing, walking, jogging, hand waving and hand clapping). Twenty-five different subjects perform each action in four different scenarios: indoors, outdoors, outdoors with scale change (fast zooming in/out) and outdoors with different clothes. This data set contains 600 video samples in total. Each clip lasts between 10 to 15 seconds and is sampled at 25Hz with an image frame size of 160×120. Fig. 3.11 shows a sample frame from each action. This dataset is challenging due to fast camera zooming, strong shadows, and low resolution (blurry) video samples. According to the initial citation [110], the video samples are divided into a test set (9 subjects: 2,3,5,6,7,8,9,10, and 22) and a training set (the remaining 16 subjects).

The **UCF sports dataset** [107] includes actions such as diving, golf swing, kicking, lifting, riding horse, run, skate boarding, swing baseball, and walk with 150 video samples collected from Youtube website. Fig. 3.12 shows a sample frame from each action. This dataset is more challenging due to diverse ranges of views and scene changes with a moving camera. The sample videos are taken from a real-world scenarios with clutter and partial occlusion. Similar to the setting in [122], we increase the data samples by adding a horizontally flipped version of each video sample to the dataset. For each video sample, both its original version and its flipped version are removed from the training set when learning the classifier. We use a leave-one-out setup and test on the original video samples.

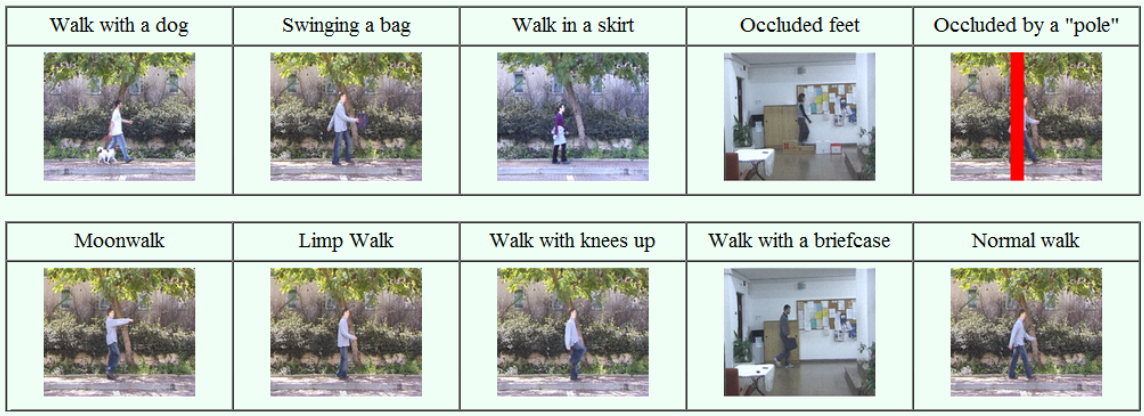


Figure 3.8: One sample frame from each of ten videos of the Weizmann robustness (deformations) dataset [11]. Different subjects ”walk” under different occlusions/clutters in front of a fixed camera.

Figure 3.9: One sample frame from each of ten videos of the Weizmann robustness (view-point) dataset [11]. The subject "walks" with different view angles in front of a fixed camera.

Salient feature matching

To obtain a quantitative measure to determine whether a salient feature from one filter corresponds to a salient feature from another filter, a volumetric similarity measure is used. Two salient features are assumed to capture the same information if they have sufficient volumetric intersection. Fig. 3.13 shows a sample scenario in which disjoint and overlapped spatio-temporal salient features are drawn. If the volumetric intersection of two salient features is sufficient, they are considered to correspond. If p_i is a salient feature in the original video, and p_j the sample salient feature in the tested (e.g., rotated) video, the regional similarity p_{ij} between i and j as the intersection of the two volumes over the union of the volumes ($p_{ij} = \frac{p_i \bigcap p_j}{p_i \bigcup p_j}$) is computed. The radii of the volume around a salient feature is considered three times its spatio-temporal scale [129] (i.e., $3\sigma \times 3\sigma \times 3\tau$). Note that we intentionally do not compare the salient features based on their descriptor, as our intention is to compare the salient feature detectors not the descriptors. The matching of the features under different views however requires the descriptor correspondence due to asynchrony across different views. This case is explained in more detail in Section 3.7.3.

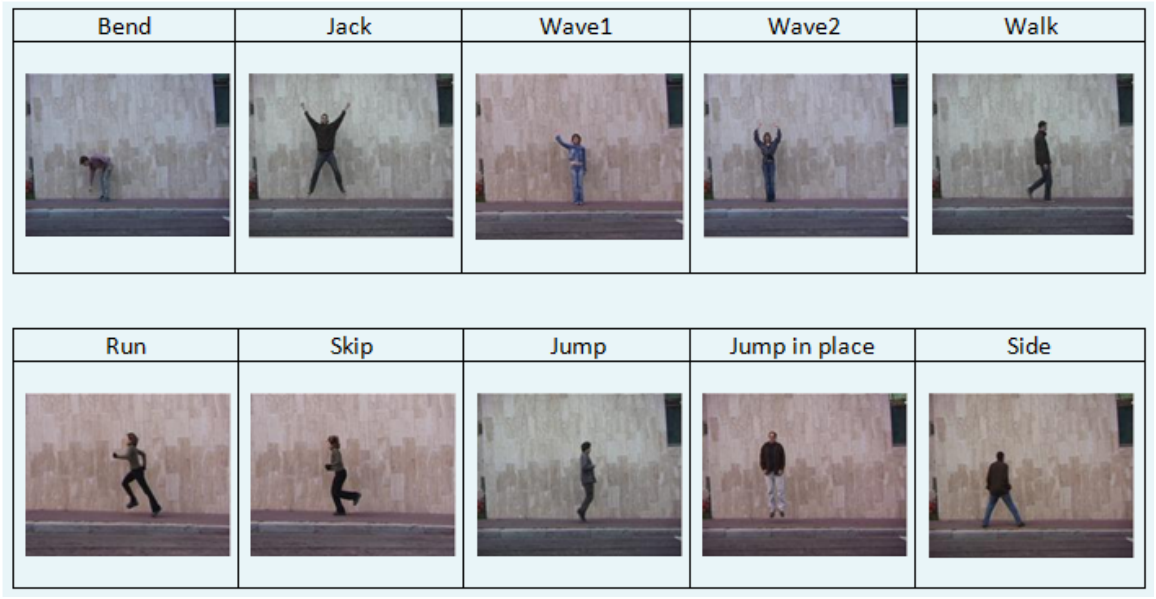


Figure 3.10: A sample frame from each action of the Weizmann classification dataset [11]. At least nine subjects perform ten different actions in front of a fixed camera.

Figure 3.11: A sample frame from each action of the KTH classification dataset [110]. Twenty-five subjects perform six different actions under four different scenarios. Note the strong shadow (in ”jogging”), zooming (in ”boxing”), and blur (in ”running”) in this dataset.

Precision score

The precision score is the ratio of the number of detected salient features from the foreground (true positive) to the number of all detected salient features. This measure determines how well

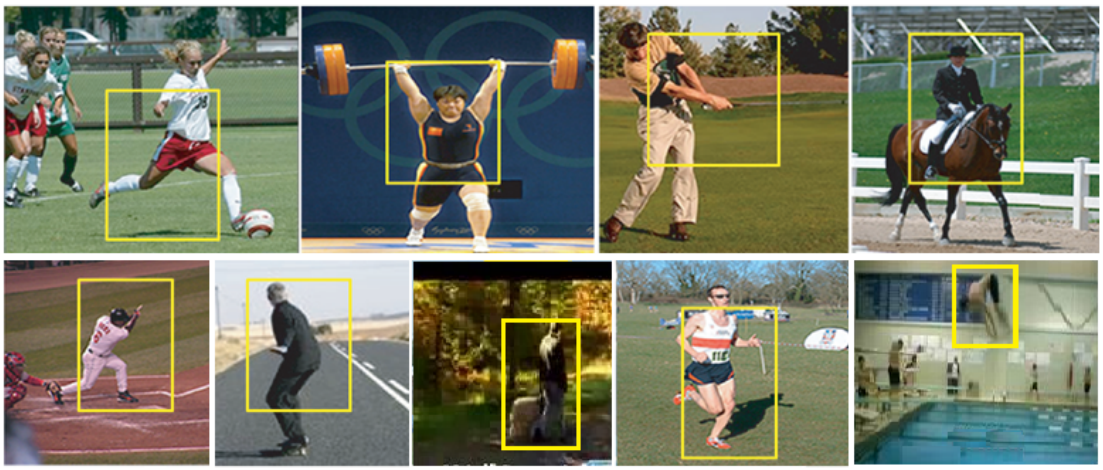


Figure 3.12: A sample frame from each of nine videos of the UCF sports dataset [107]. This dataset consists of several sports actions including diving, golf swing, kicking, lifting, riding horse, run, skate boarding, swing baseball, and walk.

a method can detect salient features which are from the foreground and not the background. A point is considered as a false positive detection (i.e., from the background), if the 3D volumetric intersection of the salient feature is less than a threshold with the available foreground volume. In determining the threshold, we should consider two main factors. (1) The spatial Gaussian filtering introduces a blurring and degradation of feature's localization (i.e., dislocation) [100, 126]. More specifically, at higher scales with more smoothing, distinct features might merge [63]. (2) the salient features are usually close to the silhouettes' boundary of the subject and hence, they have less intersection with the foreground volume. We thus consider 20% volumetric intersection threshold for reporting results, but the relative results are insensitive to this choice (Fig. 3.15). Note that the main motions in the video samples of the Weizmann classification dataset are from the moving person. As there is no video segmentation and there is not any specific parameter for removing non-relevant motions, the feature detector might find some false positives from the so-called background.

Reproducibility score

Reproducibility measures the ability to detect the same salient features under changes in view, spatial scale, and affine transformations. The reproducibility score is defined as the ratio of the number of re-detected salient features (summed between the original video and the video recorded under different settings) to the total number of salient features detected in both videos.

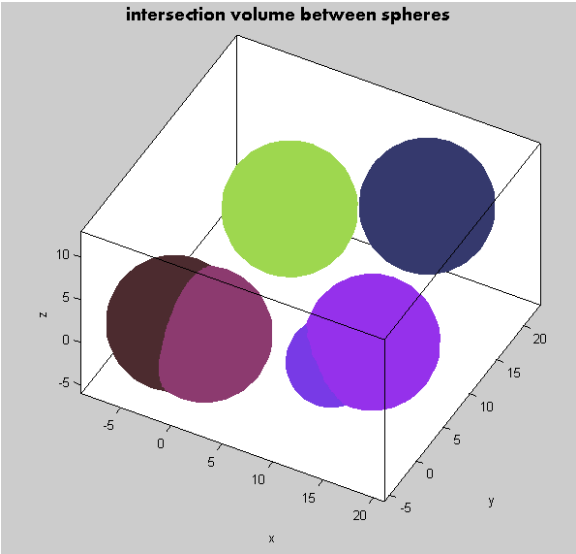


Figure 3.13: Salient features are plotted according to their spatio-temporal scales in a 3D plot showing (x, y, t) axes. Salient feature matching is based on whether two salient features are considered to correspond if they have sufficient volumetric intersection in their spatio-temporal extensions.

Action classification

To evaluate the quality of the detected salient features for action recognition, classification can be used. Classification provides a basis to relatively compare the five different temporal filtering techniques. A standard discriminative BOW approach is used for the classification [25, 68]. This classification approach satisfies our purpose which is performance comparison of the features. However, one should note that the BOW is order-less and does not take into account the spatial or temporal relationships among the features. A more promising approach would be to use, for example, a HCRF [124] or structured pLSA [135] to model these relationships. We explain this consideration in Section 7.2.1.

Here, three sets of experiments answer the research questions highlighted in Section 3.3.2.

3.7.2 Research question 1: Precision tests

Precision in classification dataset

Fig. 3.14 shows the precision score for each action type and for different spatio-temporal scales on the Weizmann classification dataset [11]. The results show that the asymmetric filters detect more true positives than the symmetric Gabor filter. Note that the filtering of the fast temporal changes in the "run" or "jump" actions results in more dislocation due to the increase of uncertainty in temporal correlation in these actions. This dislocation results in the detection of fewer true positives as Fig. 3.14(a) shows. The coarse scale results in a lower precision (Fig. 3.14(b)). This is due to the fact that wider kernels result in more smoothing of the structures and hence, more dislocation [126]. To ease the presentation of the scales in the plots, we show the rounded scales in the figures (e.g., scale $\sigma = 2.82$ is rounded to $\sigma = 3$).

As Figure 3.15 shows, varying the volumetric intersection threshold generates results that are relatively consistent. Here, the Gabor has the least precision and the asymmetric sinc performs the best. The Poisson and scale-derivative Gaussian perform the second best. The performance of the log Gaussian is higher than the symmetric Gabor filter, but less than the other asymmetric filters.

Fig. 3.19-3.22 show the salient features detected using different temporal filters as local 3D space-time volumes in the video. For clarity, 2D projection of these features on different frames have been shown in Fig. 3.23-3.24. The main observation is that the asymmetric filters detect more salient features and different features than the symmetric Gabor filtering.

Precision in robustness dataset

Fig. 3.16 shows the precision score for each cluttered video and for different spatio-temporal scales on the Weizmann robustness dataset [11]. As Fig. 3.16(a) shows, the asymmetric sinc has the best overall performance. The asymmetric filters performs better than the Gabor filter, except in the video when a person walks with a dog, the Gabor performs slightly better than the log Gaussian. This shows that the asymmetric filters detect fewer false positive salient features than symmetric Gabor filtering under irregular motion patterns, clutter, and occlusion. Fig. 3.16(b) shows the better overall precision of the asymmetric filters compared with the Gabor filter at different spatio-temporal scales.

The role of causality and symmetry

To better understand the role of causality and asymmetry on feature detection, we experimented with (a) shifting the peak of each causal kernel to the origin to create a non-causal (but asymmetric) version and (b) right-shifting the Gabor to create a causal filter. For conciseness, we do not report the results here as similar precision results were obtained between the original and shifted versions for each of the causal and non-causal (Gabor) filters. This observation makes sense as the convolution commutes with translation meaning that the time-shift in the convolution kernel enforces just a time-shift in the output which can be compensated accordingly.

$$f(t) \star g(t) = y(t) \Longrightarrow f(t) \star g(t - t_0) = y(t - t_0) \tag{3.34}$$

One should note that different temporal kernels have different tails, even though they use the same standard deviation. To better differentiate the role of asymmetry and the tail, we performed an experiment with three synthetic kernels derived from the Gaussian function which is the envelope of the symmetric Gabor filter. Fig. 3.17 shows a synthetic asymmetric kernel in which each side of the kernel is a truncated Gaussian with different standard deviations (3.35).

$$g(t) = G_{\sigma_1} S(-t) + G_{\sigma_2} S(t) \tag{3.35}$$

where the G_σ is a Gaussian function with standard deviation of σ and the $S(t)$ is the Heaviside step function. The ratio $\alpha = \frac{\sigma_1}{\sigma_2}$ controls the skewness of the kernel. We applied the precision test on the classification dataset with different values of the skewness, $\alpha = 1$ is the standard Gaussian function, $\alpha = 0$ (i.e., $\sigma_1 = 0$) makes a one-sided Gaussian kernel, and $\alpha < 1$ (i.e., $\sigma_2 > \sigma_1$)produces an asymmetric kernel with skewness towards the times before the peak. We then multiply the complex exponential term with each kernel (Section 3.4) and used it in our common framework (Fig. 3.3) for the local feature detection.

Table 3.1 presents the average precision score for different synthetic filters. Note that the asymmetric filter performs the best. The causal filter performs better than the symmetric Gabor filter as well. This results show that the asymmetry is the main reason why the four introduced asymmetric filters (i.e., log-Gaussian, scale-deriv. Gaussian, Poisson, and asymmetric sinc) perform better than the symmetric Gabor filter. However, the performance difference among the asymmetric filters is due to their shape and tails.

Research question 1: Conclusion

To answer the research question 1 regarding the effect of causality and symmetry of a temporal scale-space kernel for salient feature detection, the precision tests show that a non-causal sym-

Table 3.1: Precision score for synthetic filters. The score is averaged over all nine spatio-temporal scales and all video samples from the Weizmann classification dataset [11]. Note that the asymmetric filter performs the best. The causal filter performs better than the symmetric Gabor filter as well.

Temporal filter	α	Precision score
Symmetric Gabor	1	0.648
Causal Gabor	0	0.711
Asymmetric Gabor	2/3	0.757

metric temporal kernel just enforces a delay in the system and does not impact the performance of salient feature detection. The key distinction is that the asymmetric filters outperform the symmetric filter. More specifically, the left-skewed kernels such as the asymmetric sinc, Poisson, and scale-derivative Gaussian have a better performance for action-related salient feature detection compared with the less-skewed log Gaussian. All of these four asymmetric kernels perform better than the symmetric kernel of the Gabor. Better overall performance of asymmetric filters can be due to decaying slower than the Gabor filter (Fig. 3.7). In fact, the motion pattern in a real-world natural videos is not perfectly periodic from one cycle to another during performance of a given action by an individual. Slow decay of asymmetric filters can better capture this motion imperfection by passing more temporal frequencies than the ideal Gabor filter. Keeping the higher temporal frequencies in the band-pass asymmetric filters means less motion smoothing and better preservation of the the salient motions. This results in less dislocation of the salient motion features and hence, fewer false positive detection.

Number of detected salient features

To compare the computational cost and the encoding capability of different temporal filters is important to evaluate the number of detected salient features. As the number of frames is not the same for all the videos, we report the normalized number of the salient features in one second of the video. The normalization is obtained by the division of the number of detected salient features over the number of frames. The result is then multiplied by the frame rate (i.e, $25Hz$) to achieve the number of salient features per second. As Fig. 3.18 shows, the number of salient features detected using Gabor filtering is usually fewer than asymmetric temporal filters. Log-Gaussian and scale-derivative Gaussian detect the most salient features distributed mainly around

body limbs with significant motions in an action. Intuitively, detecting more salient features provides a more complete representation and hence, a better information encoding. Moreover, denser features have higher chance to be re-detected under different geometric deformations and hence, their matching rate will be higher. As can be seen in Fig. 3.18(a), fewer salient features are detected in actions such as "bend" and "wave" which are performed in place. More salient features are detected at actions such as "run" and "walk" in which the whole body is moving. As expected, fewer salient features are detected with a spatial and/or temporal scale increase, especially when the temporal scale increases (Fig. 3.18(b)). If we consider the number of salient features as an indicator of local extrema, the decrease in number of salient feature with spatial/temporal scale increase is consistent with the non-creation of local extrema axiom in the scale-space theory [77].

3.7.3 Research question 2: Reproducibility tests

View change

To test the reproducibility under different camera view angles, the Weizmann robustness (viewpoint) dataset [11] was used. Due to asynchrony of video frames under view change, the volumetric intersection for salient feature matching cannot be directly computed. We therefore describe a salient feature using a 3D SIFT descriptor [111] and use this feature vector for matching salient features similar to 2D SIFT matching [84]. As Fig. 3.25 shows, the Gabor filter has the lowest robustness compared to the asymmetric filters. The asymmetric filters can reproduce around 70% of the same salient features up to about 60^o in view change, while the Gabor can barely reproduce half of them.

Spatial scale change

Spatial scale change occurs when the camera's distance from the action or zooming might vary. To test, we generate video samples at three different spatial (sub-sampling) scales $\{1/2, 1/3, 1/4\}$ for each of the 93 video samples in the Weizmann classification dataset [11]. The original frame size of 180×144 is therefore shrunk into 45×36 for the maximum spatial sub-sampling of $1/4$. Results in Fig. 3.27 show that, on average, Gabor performs better than asymmetric filtering under spatial scale sub-sampling. The scale-derivative Gaussian and Poisson filters perform slightly better than the asymmetric sinc and log-Gaussian filter. In actions with limited moving parts such as "one/two hand waving" or "bend", Gabor performs better. In actions such as "walk", "pjump" and "jump", the Poisson performs better than Gabor. One possible explanation is that the symmetric temporal filtering is more responsive to symmetric structures which are better preserved

under spatial sub-sampling. Fig. 3.26 shows the shape change of the asymmetric Poisson and the symmetric Gaussian which is the cover of the Gabor filter. As can be seen, the shape of a Poisson kernel changes significantly with increase in the sub-sampling factor, while the shape of a symmetric Gaussian kernel is preserved under sub-sampling.

Affine transformation

To evaluate the effect of in-plane **rotation**, we generated synthetic rotated videos using a spatial rotation transformation A_{rot} (3.36) applied on all frames of a video. Eight videos were created from a single video using the 93 video samples from the classification dataset [11], rotating from -60^o to $+60^o$ with an interval of 15^o. In total, 744 rotated videos were generated. For salient feature matching, the detected salient features in the rotated video are rotated back in the original video domain to compute the relevant volumetric intersection with the salient features of the original video. As can be seen in Fig. 3.28, the asymmetric sinc filter performs better than the other kernel. The other symmetric filters are slightly better than the symmetric Gabor filter.

$$A_{rot} = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix} \tag{3.36}$$

To test the reproducibility under in-plane **shearing**, we used the vertical shearing factor of α set to $\{0, 0.2, 0.4, 0.6\}$ and the horizontal shearing factor of β set to $\{0, 0.2, 0.4, 0.6\}$ as part of shearing transformation matrix A_{shear} (3.37). This generates 15 synthetic sheared videos (excluding $(0, 0)$) from each of 93 videos of the Weizmann classification dataset [11]. In total, 1395 sheared videos are compared with the 93 original videos.

$$A_{shear} = \begin{bmatrix} 1 & \alpha \\ \beta & 1 \end{bmatrix} \tag{3.37}$$

The reproducibility score under shearing transformation decreases more in the actions with limited movements (Fig. 3.29(a)). We visually observed that the shape change due to shearing in the frames of actions with limited body limbs movement such as bending or one-hand waving is much more that this change in fast motions such as jump in place (pjump) or jump. As expected, the reproducibility decreases with increase in shearing (Fig. 3.29(b)). The asymmetric filters show better performance compared with the symmetric Gabor filter. The asymmetric sinc performs the best.

Research question 2: Conclusion

Based on the result of reproducibility tests, to answer the research question 2, we can say that the salient features which are detected using the asymmetric temporal filters are more robust under view change or affine transformation, compared to those salient features detected using temporal symmetric Gabor filtering. The Gabor filter provides better robustness under spatial scale change as the symmetric shape are better kept the same under sub-sampling compared to the asymmetric shapes. To summarize the experimental results of precision and reproducibility, Table 3.2 shows the performance averaged over all scales and all action types. As can be seen, except for scale change, the asymmetric filters perform better than symmetric Gabor filter. Among asymmetric filters, the asymmetric sinc performs the best. Poisson and scale-deriv. Gaussian show better overall performance than the log Gaussian filtering. As can be seen in Fig.3.18, asymmetric sinc detects fewer features while its performance is the best, overall. The Gabor also detects fewer features than the other asymmetric filters such as Poisson, while its performance is less compared to those asymmetric filters. This observation shows that detection of denser or sparser set of features might not be directly related to the performance of the features. The robustness of the features is related to the type of features they detect. In fact, there are more asymmetric-type motions than symmetric type in the human actions. The asymmetric temporal filters are thus more precise in modeling these type of motions than a symmetric filter.

3.7.4 Research question 3: Action classification

In this experiment, we use the Weizmann classification dataset [11] and the KTH dataset [110]. For a fair comparison, we fix all the components of the standard discriminative BOW framework [25, 68, 112] and change just the temporal filter.

We applied both causal and non-causal temporal filtering to detect local salient features at a given spatio-temporal scale. We then used a 3D SIFT descriptor [111] as the feature descriptor. This descriptor gives a normalized histogram of the spatio-temporal orientations of gradients and hence, both appearance and motion of the salient feature are encoded in the descriptor.

The salient features from all of the actions in the training set and all spatio-temporal scales are quantized into visual words, K-means was performed 10 times and the result with the lowest error was kept, consistent with the experimental setting used in [25, 122].

We experimentally set the number of clusters to 200 for the Weizmann dataset. For the KTH and UCF sports datasets, we set it to 1000 consistent with the setting in [122]. We used both nearest neighbor (NN) classifier and SVM [21, 119] as the classifier. The SVM does not show improvement over NN classifier on the Weizmann dataset due to small number of training

samples. SVM however slightly improves the NN classifier on the KTH and UCF sports datasets. For conciseness, we thus report the classification accuracy of NN for the Weizmann dataset and the SVM for the KTH and UCF sports datasets. To compare two action signatures S_i and S_j, we used χ^2 distance metric (3.38) for the NN classifier on the Weizmann classification dataset with leave-one-out setting and the radial basis function (5.8) (RBF) for the SVM classifier on the KTH dataset with training/testing setting mentioned by the initial paper [110]. For the UCf sports dataset, we used similar experimental setting as [122] in which horizontally flipped version of each original sample is added to the training set and a leave-one-out setting for testing. We used the LibSVM library [15] to learn the parameter γ of the RBF function using cross validation.

$$D_{\chi^2}(S_i, S_j) = \frac{1}{2}\frac{(S_i - S_j)^2}{S_i + S_j} \tag{3.38}$$

$$K_{RBF}(S_i, S_j) = e^{-\gamma|S_i - S_j|^2} \tag{3.39}$$

As Table 3.2 shows, the classification accuracy of using salient features detected using a temporal Gabor filter is the lowest. The asymmetric sinc provides the best results. For the Weizmann and the KTH datasets, the Poisson and scale-derivative Gaussian perform the second best. The performance of the log Gaussian is higher than the symmetric Gabor filter, but less than the other asymmetric filters. For the UCF sports dataset, the log-Gaussian performs the second best.

Fig. 3.30 shows the confusion matrices for action recognition on the Weizmann classification dataset [11] using the features detected by different temporal filters. Note that all methods have difficulty discriminating "skip" from "run". This is due to the similarity of motion patterns of these fast motions in this dataset. All asymmetric filters perform better than Gabor and the asymmetric sinc is the best.

Fig. 3.31 shows the confusion matrices for action recognition on the UCF sports dataset [107] using the features detected by different temporal filters. Note that all methods have difficulty discriminating some actions due to wide view change or scene changes in the video samples. The asymmetric sinc performs the best compared to other temporal filters.

Research question 3: Conclusion

The action classification experiment shows that the asymmetric temporal filters provide salient features that better encode actions and generate higher classification results. As the asymmetric temporal filters provide more precise salient features with higher robustness under geometric

Table 3.2: Summary of the evaluation tests of different temporal filters for salient feature detection and action classification, averaged over all scales and all video samples. Bold numbers represent the highest performance. Note that overall the asymmetric sinc performs the best.

		Temporal filters				
Test	Type	Gabor [25]	**Log Gaussian**	**Scale-deri. Gaussian**	**Poisson**	**Asym. sinc**
Precision	classification (Sec. 3.7.2)	64.8 %	72.4 %	74.7 %	74.7 %	**79.6** %
	robustness (Sec. 3.7.2)	91.9 %	93.9 %	93.5 %	93.6 %	**96.1** %
Reproducibility	view change (Sec. 3.7.3)	55.5 %	70.7 %	69.3 %	73 %	**73.4** %
	scale change (Sec. 3.7.3)	**50.7** %	43.9 %	47.3 %	47.1 %	45.2 %
	rotation change (Sec. 3.7.3)	86.6 %	88.9 %	87.8 %	87.8 %	**93.3** %
	shearing change (Sec. 3.7.3)	84.1 %	86 %	86.7 %	86.8 %	**91.8** %
Recognition	Weizmann (Sec. 3.7.4)	91.1 %	91.5 %	93.5 %	93.5 %	**95.5** %
	KTH (Sec. 3.7.4)	89.7 %	91.1 %	90.2 %	91.7 %	**94.3** %
	UCF sports (Sec. 3.7.4)	73.3 %	87.3 %	74.7 %	82.7 %	**91.7** %

deformations (except for scale change), better performance for action representation and action classification was expected and did occur. The experimental results (Table 3.2) show improvement of asymmetric filtering over symmetric Gabor filter. The action classification test on the KTH dataset with camera zooming shows that the asymmetric filters perform better than Gabor due to better overall precision and robustness, even though the Gabor filter performs better under spatial scale change. The action classification in a more realistic scenario such as UCF sports dataset shows the better performance of an asymmetric temporal filtering over a symmetric Gabor filtering for good quality salient feature detection in video.

As Table 3.2 shows, action classification improves simply by changing the symmetric tem-

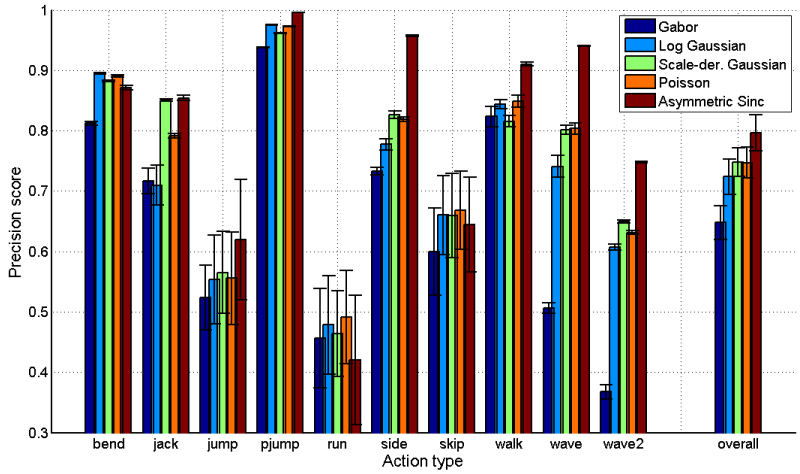
poral filter to asymmetric filters, under a standard discriminative BOW framework. There are many action recognition schemes that we could have used, some superior to the BOW that we used. However, the focus of this chapter is on the quality of the features and the choosing an appropriate action recognition method is left to future chapters. Without video segmentation and with a similar BOW setting, our 95.5% accuracy on the Weizmann dataset is comparable with the method in [40] with 83.7% and in [111] with 82.6%. Similarly, we obtained 94.3% accuracy on the KTH dataset which is comparable with 84.3% using 3D Hessian features [129], 88.3% accuracy using dense features [106], and 92.1% accuracy using 3D Harris features [122]. Moreover, our 91.5% accuracy on the UCF sports dataset is much better than the accuracy obtained using the dense sampling [122] with 85.6%, the unsupervised feature learning [72] with 86.5%, and the dense trajectory [44] with 88.2%. Considering the sparsity of asymmetric sinc filtering (Fig. 3.18), our results are in contrast to the previous observations about better performance of dense sampling/trajectories, showing the importance of proper temporal filtering for robust and informative key point detection.

3.8 Conclusion

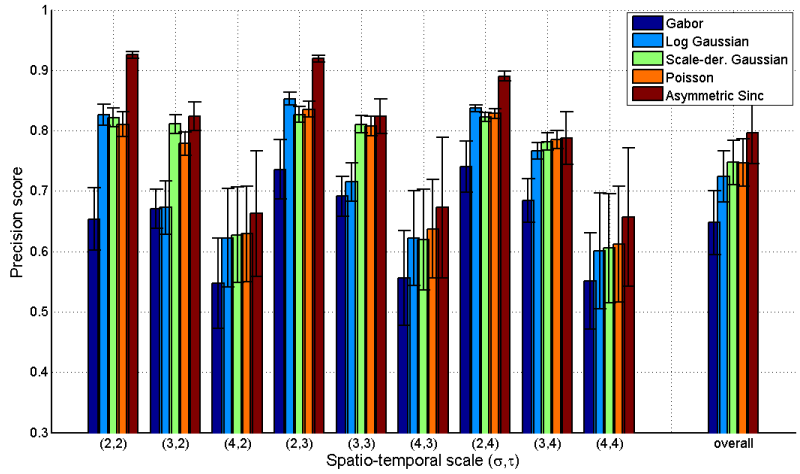
Existing salient feature detection methods utilize just the symmetric non-causal temporal filters. In this section, we introduced three asymmetric causal temporal scale-space kernels of Log Gaussian, Poisson, and Scale-derivative Gaussian. Moreover, we developed a brand new time-causal kernel of asymmetric sinc. We evaluated the role of asymmetry and causality of the temporal kernels on salient feature detection. A comprehensive study in terms of precision of the detected salient features, the reproducibility score under different geometric deformations, and action recognition task have been performed. We used both Weizmann classification and robustness data sets for the robustness tests and the Weizmann classification, the KTH, and the UCF sports dataset for the action classification purpose. The results show that asymmetric temporal filtering provides more precise and more robust (except under scale change) salient features than symmetric Gabor filtering. Moreover, the non-causality just enforces a temporal delay in the system and symmetry does not improve the performance. Among all the kernels, the asymmetric sinc has the best performance.

In a standard discriminative BOW framework, we obtained higher action classification accuracy using the salient features detected with asymmetric temporal filtering. However, the improved performance in feature detection permits us to explore unique representation and classification strategies. In the next Chapter (Section 4.3), we propose multi-resolution action representation which better encodes the motions compared to the global BOW representation that we used in this section. Moreover, in Chapter 5, we propose the multiple classifier systems which

improves the classification accuracy compared to the single classifier that we used in this section.



(a) Precision score for each action type using all scales



(b) Precision score for each scale for all action types

Figure 3.14: Precision test on the video samples of the Weizmann classification dataset [11]. (a) The result for each action type is averaged over all samples at nine spatio-temporal scales and the bar shows the variance. Note that the precision of all filters reduces at faster motions such as run. (b) The horizontal axis shows the nine spatio-temporal scales. The results are averaged over 93 video samples. Note that the precision is reduced by a spatial and/or temporal scale increase. The asymmetric sinc performs the best. The other asymmetric filters perform better than the symmetric Gabor filtering.

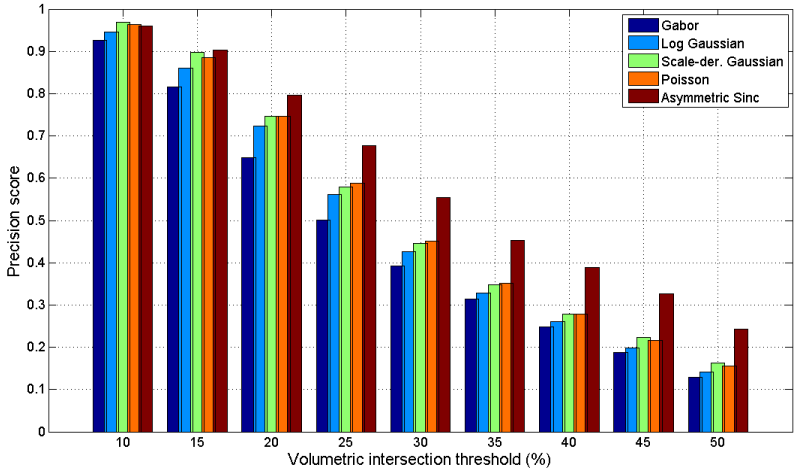
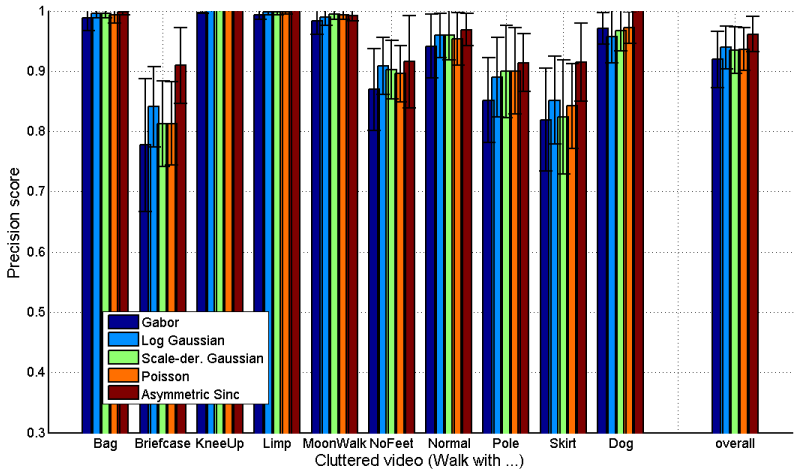
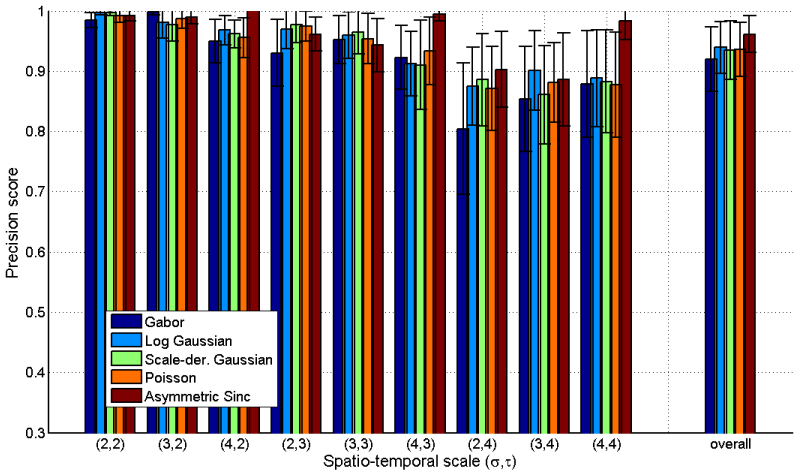


Figure 3.15: Precision test for different thresholds on the Weizmann classification dataset [11]. Note that the symmetric Gabor kernel has a lower performance compared with the asymmetric kernels. More specifically, the asymmetric sinc performs the best.



(a) Precision score for each action type using all scales



(b) Precision score for each scale for all action types

Figure 3.16: Precision test on the video samples from the Weizmann robustness (deformations) dataset [11]. (a) The horizontal axis represents the "walking" action cluttered in different situations. The result for each action type is averaged over nine spatio-temporal scales. (b) The horizontal axis shows nine spatio-temporal scales. Note that overall asymmetric filters perform better than symmetric Gabor filter. The asymmetric sinc performs the best.

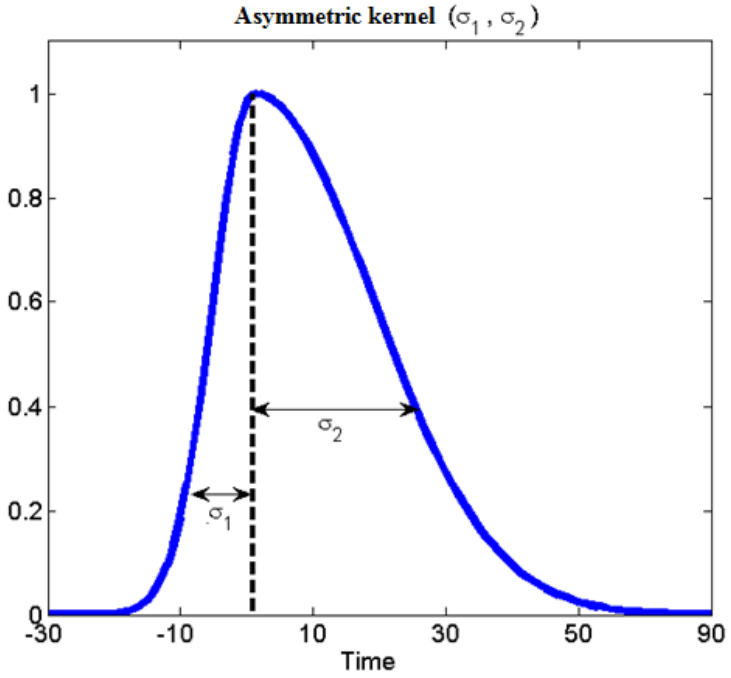
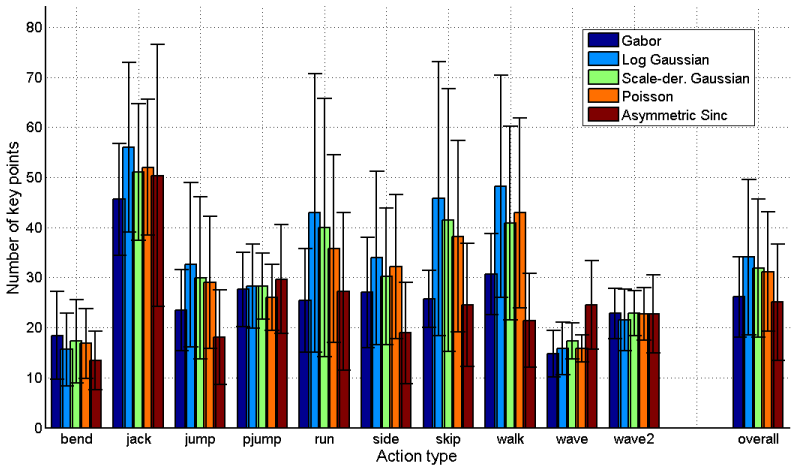
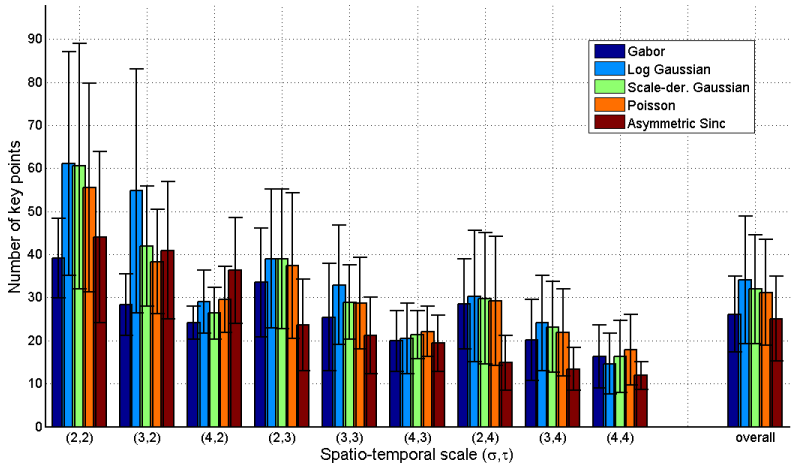


Figure 3.17: Synthetic asymmetric kernel. The ratio $\alpha = \frac{\sigma_1}{\sigma_2}$ controls the skewness of the kernel. The kernel with $\alpha = 1$ is the standard Gaussian function, $\alpha = 0$ (i.e., $\sigma_1 = 0$) makes a one-sided Gaussian kernel, and $\alpha < 1$ produces an asymmetric kernel with skewness towards the times before the peak.



(a) Average number of salient features for each action type using all scales



(b) Average number of salient features for each scale for all action types

Figure 3.18: Average number of salient features per second on the Weizmann classification dataset [11]. (a) The result for each action type is averaged over all samples at nine spatio-temporal scales and the bar shows the variance. Note that fewer salient features are detected for actions performed in place (e.g., bend, pjump). (b) The result for each spatio-temporal scale is averaged over 93 video samples. With a spatial and/or temporal scale increase, the number of salient features is reduced.

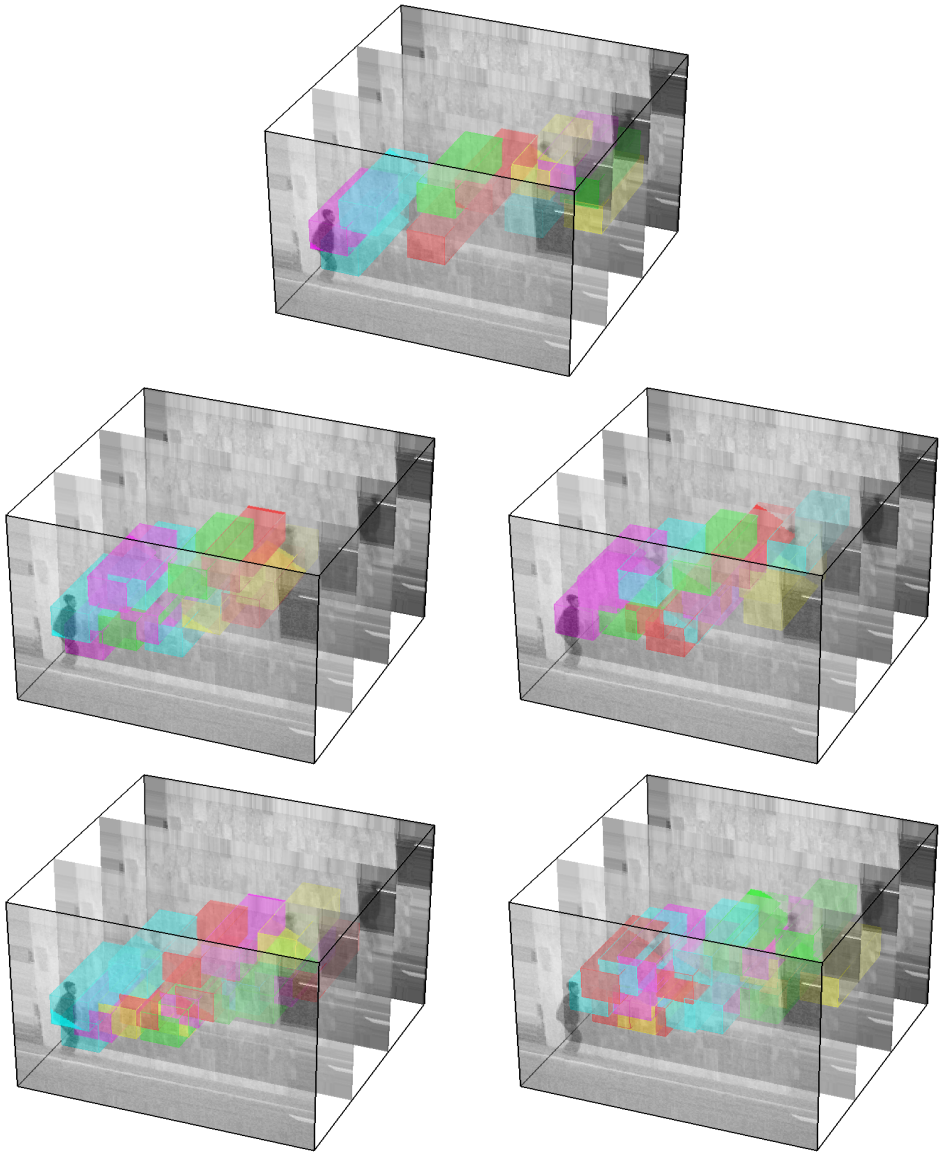


Figure 3.19: 3D local volumes of the motion features at scale $\sigma = 4$ and $\tau = 4$ detected using different temporal filtering (Gabor (first-row), log Gaussian (second-row, left), scale-derivative Gaussian (second-row, right), Poisson (third-row, left), and asymmetric sinc (third-row, right)) in the video of "jumping" from the Weizmann data set [11]. The volumes show the spatio-temporal extension of the features. Note that different features are localized at different locations in the video volume.

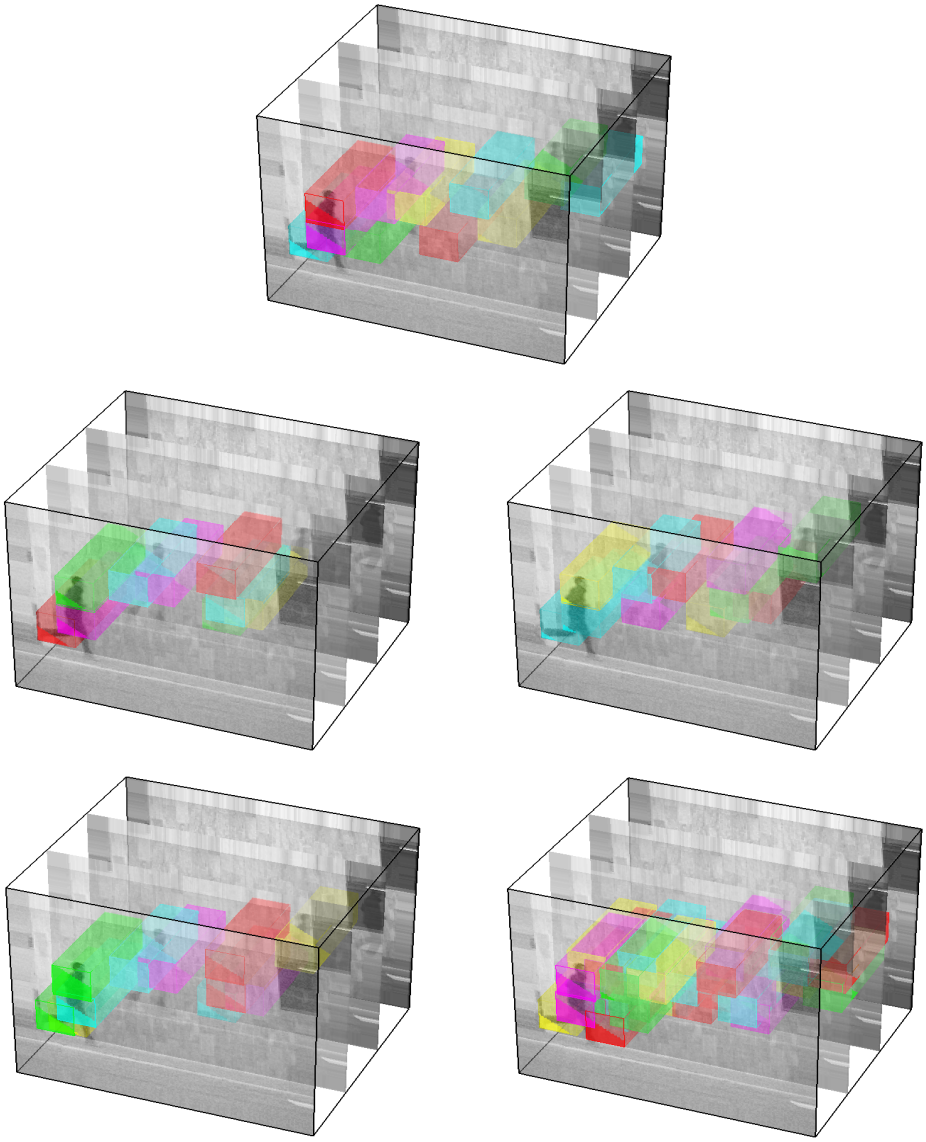


Figure 3.20: 3D local volumes of the motion features at scale $\sigma = 4$ and $\tau = 4$ detected using different temporal filtering (Gabor (first-row), log Gaussian (second-row, left), scale-derivative Gaussian (second-row, right), Poisson (third-row, left), and asymmetric sinc (third-row, right)) in the video of "running" from the Weizmann data set [11]. The volumes show the spatio-temporal extension of the features. Note that different features are localized at different locations in the video volume.

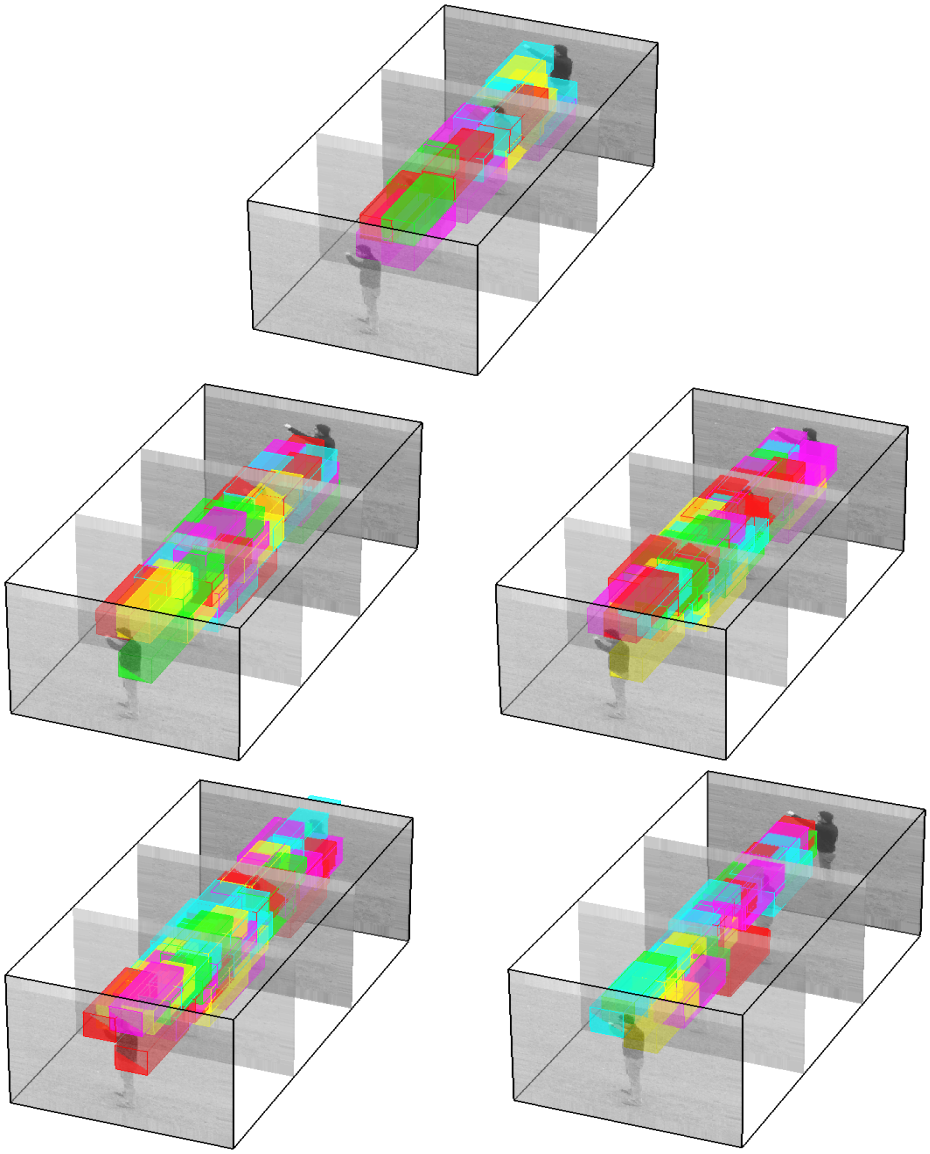


Figure 3.21: 3D local volumes of the motion features at scale $\sigma = 4$ and $\tau = 4$ detected using different temporal filtering (Gabor (first-row), log Gaussian (second-row, left), scale-derivative Gaussian (second-row, right), Poisson (third-row, left), and asymmetric sinc (third-row, right)) in the video of "boxing" from the KTH data set [110]. The volumes show the spatio-temporal extension of the features. Note that different features are localized at different locations in the video volume.

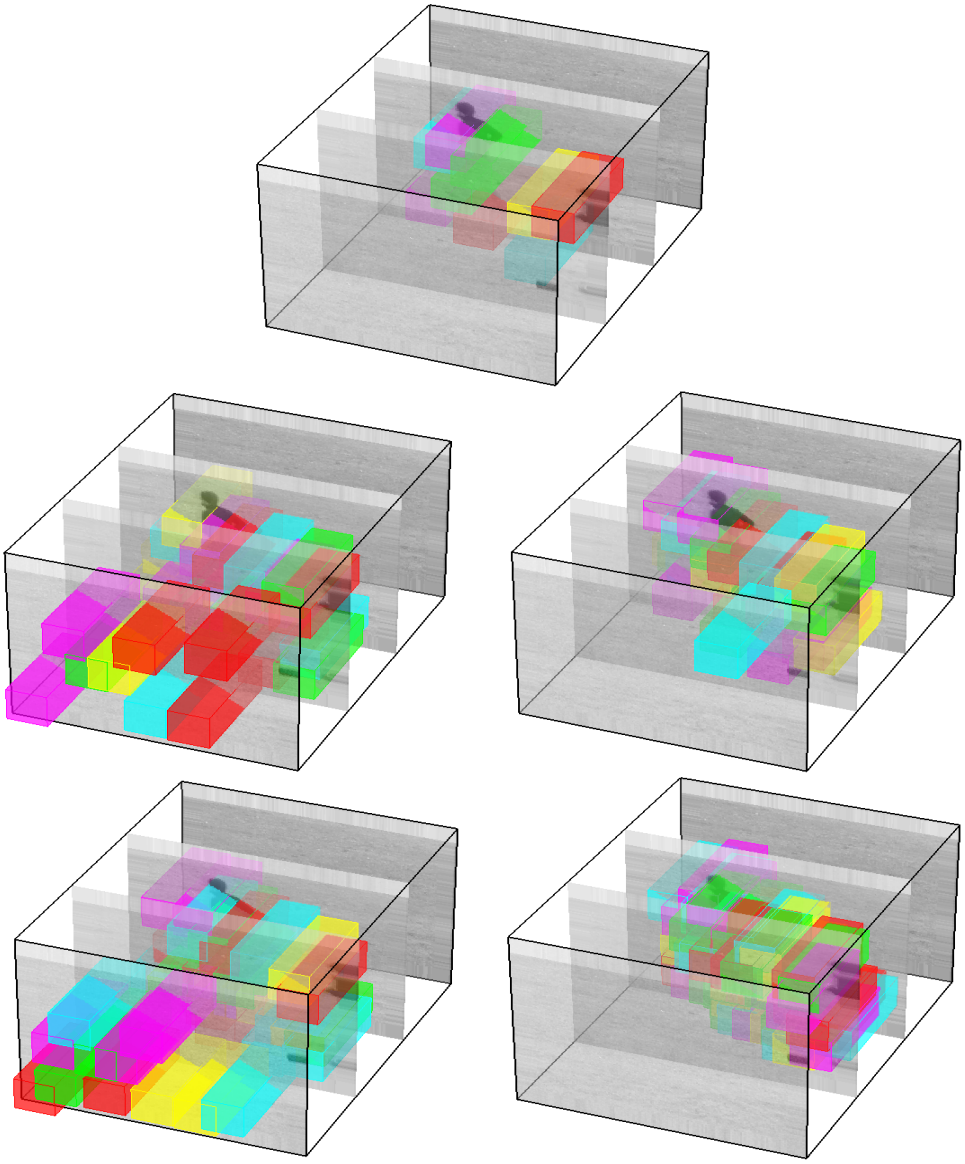


Figure 3.22: 3D local volumes of the motion features at scale $\sigma = 4$ and $\tau = 4$ detected using different temporal filtering (Gabor (first-row), log Gaussian (second-row, left), scale-derivative Gaussian (second-row, right), Poisson (third-row, left), and asymmetric sinc (third-row, right)) in the video of "running" from the KTH data set [110]. The volumes show the spatio-temporal extension of the features. Note that different features are localized at different locations in the video volume.

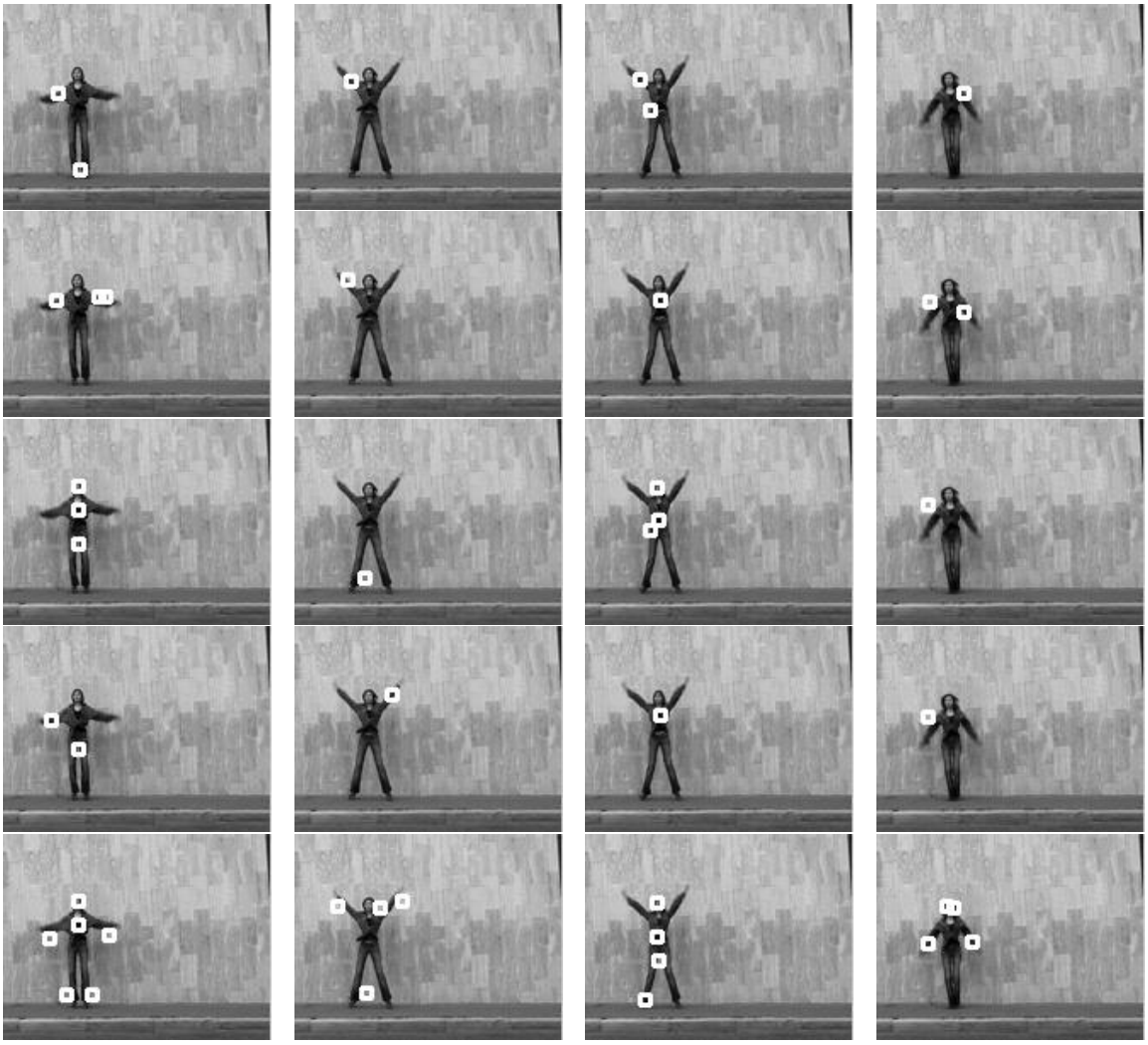


Figure 3.23: 2D projection of the local motion features at scale $\sigma = 2$ and $\tau = 2$ detected using different temporal filtering (from top row to bottom row: (a) Gabor, (b) log Gaussian, (c) scale-derivative Gaussian, (d) Poisson, and (e) asymmetric sinc) in the video of "jumping jack" from the Weizmann data set [11].

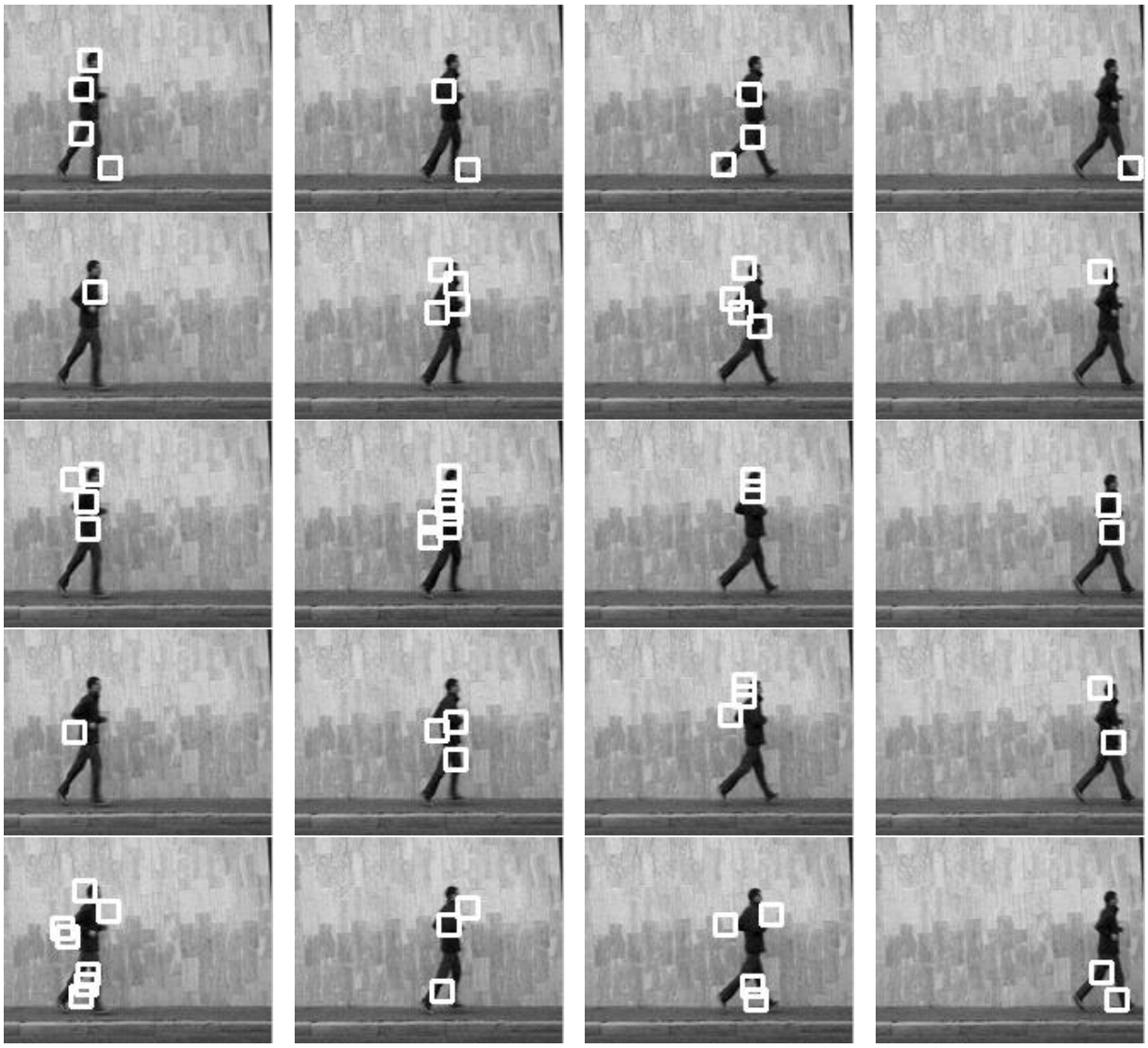
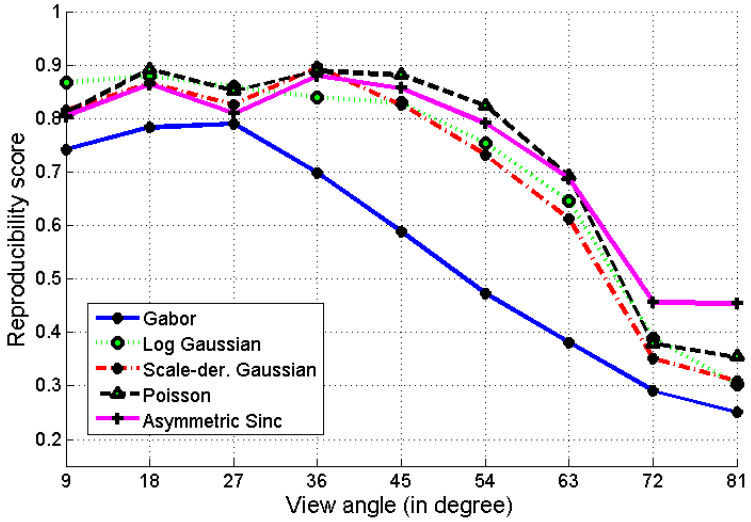
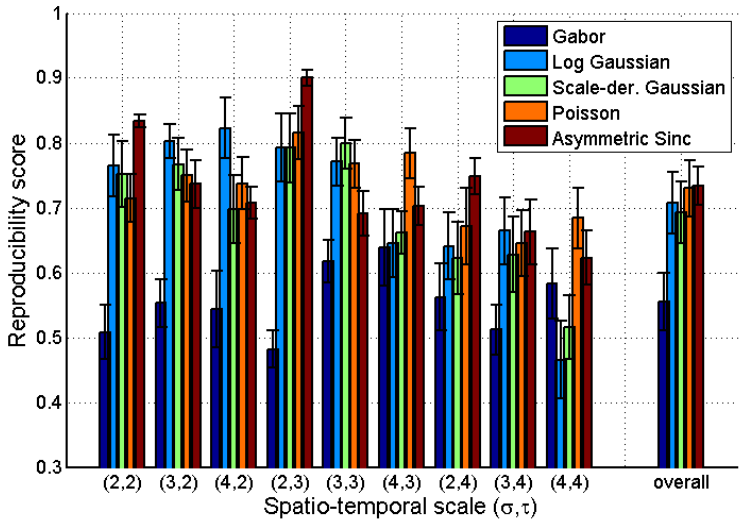


Figure 3.24: 2D projection of the local motion features at scale $\sigma = 3$ and $\tau = 2$ detected using different temporal filtering (from top row to bottom row: (a) Gabor, (b) log Gaussian, (c) scale-derivative Gaussian, (d) Poisson, and (e) asymmetric sinc) in the video of "running" from the Weizmann data set [11].



(a) Reproducibility score for different views using all scales



(b) Reproducibility score for each scale for all action types

Figure 3.25: View-change reproducibility test. (a) The horizontal axis represents the "walking" action performed at different view angles with the camera. The result for different views is averaged over all samples at nine spatio-temporal scales. As was expected, the performance decreases by an increase in the view angle. Asymmetric filters can reproduce about 70% of the salient features up to 60^o view change, while Gabor cannot reproduce even half of that. (b) The horizontal axis shows nine spatio-temporal scales and the bar shows the variance.

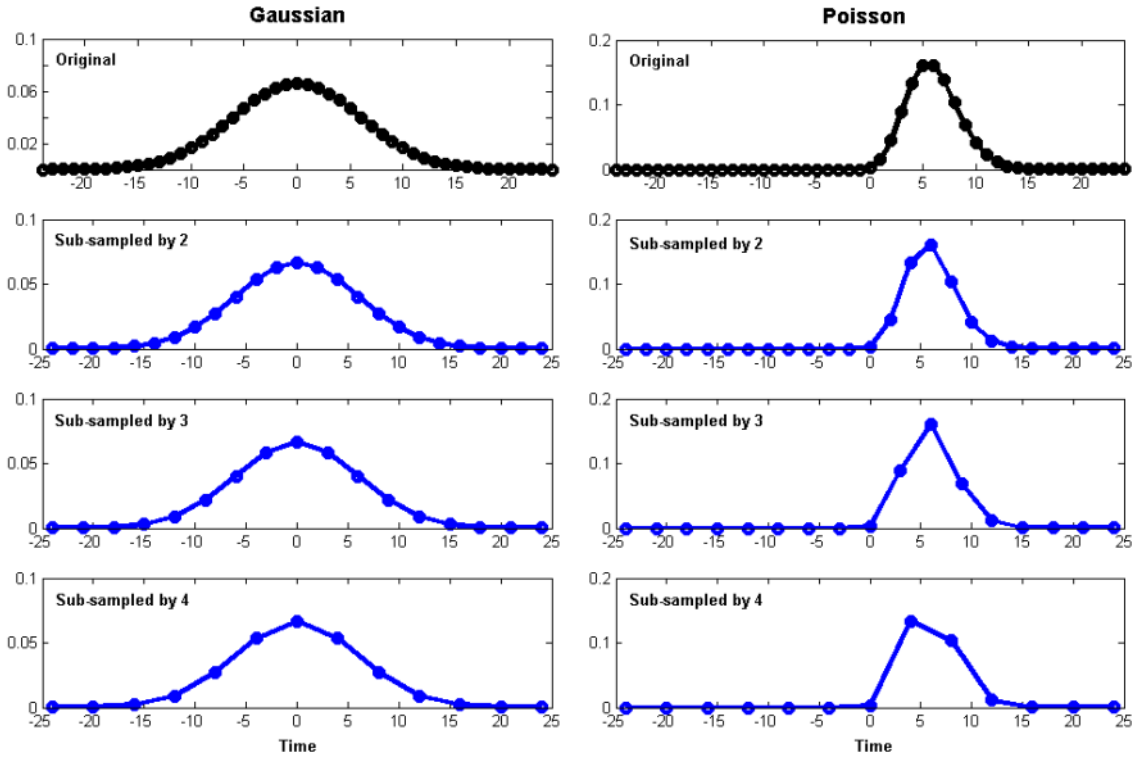
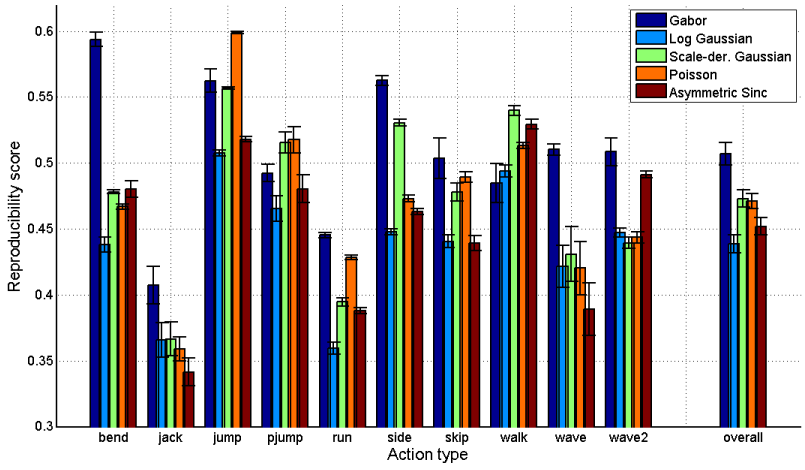
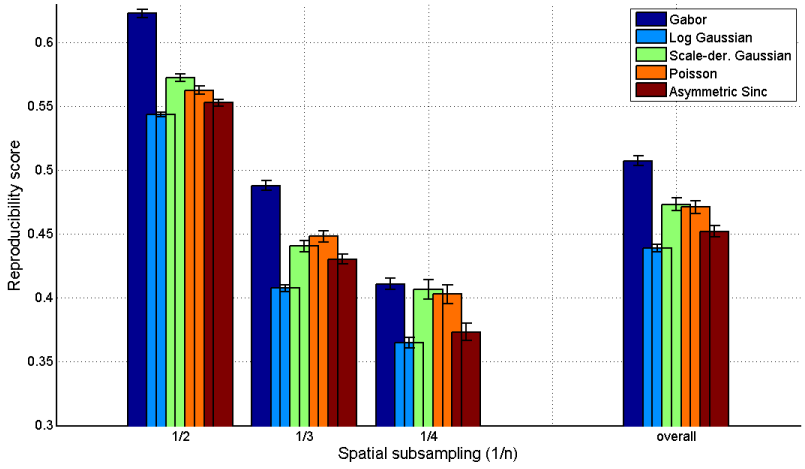


Figure 3.26: Shape change of a symmetric Gaussian kernel and an asymmetric Poisson kernel under sub-sampling. From top to bottom, the kernels are sub-sampled by a factor of 2,3, and 4, respectively. Note that the shape of the Gaussian (left column) is preserved under sub-sampling, while the shape of the Poisson (right column) varies significantly with increase in the sub-sampling scale.

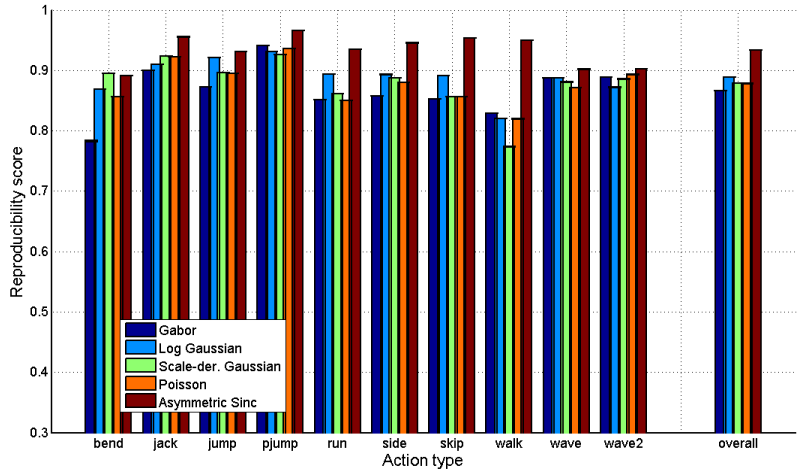


(a) Reproducibility score for each action type using all scales

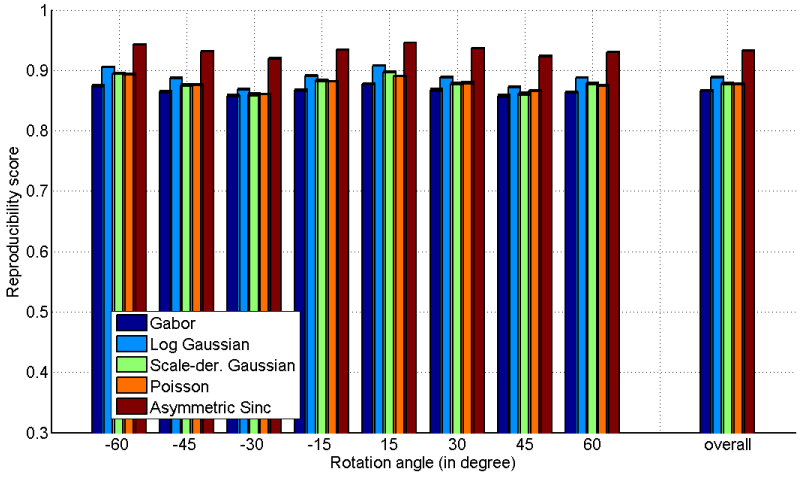


(b) Reproducibility score for each scale for all action types

Figure 3.27: Spatial scale-change reproducibility test. (a) The reproducibility score for each of ten action types is averaged over all samples of the action and over 3 sub-sampling scales $\{1/2, 1/3, 1/4\}$. Each bar shows the associated variance. (b) For each of 3 sub-sampled scales, the reproducibility score is averaged over 93 video samples. Note that for both (a) and (b), the Gabor filtering performs better than the other filters.

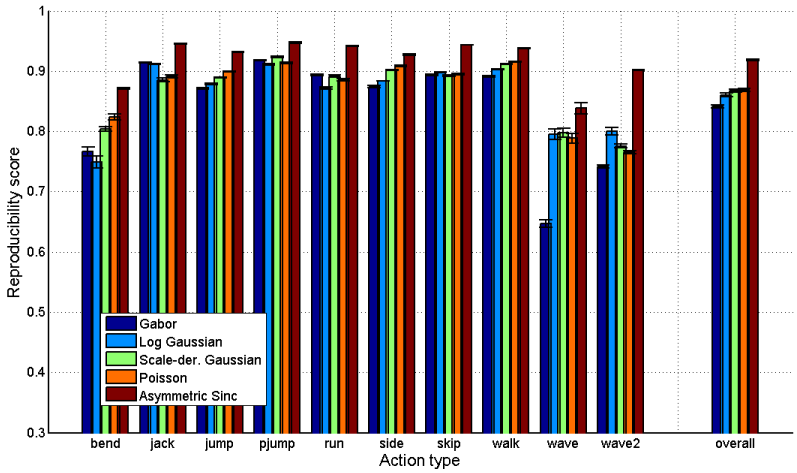


(a) Reproducibility score for each action type using all rotations

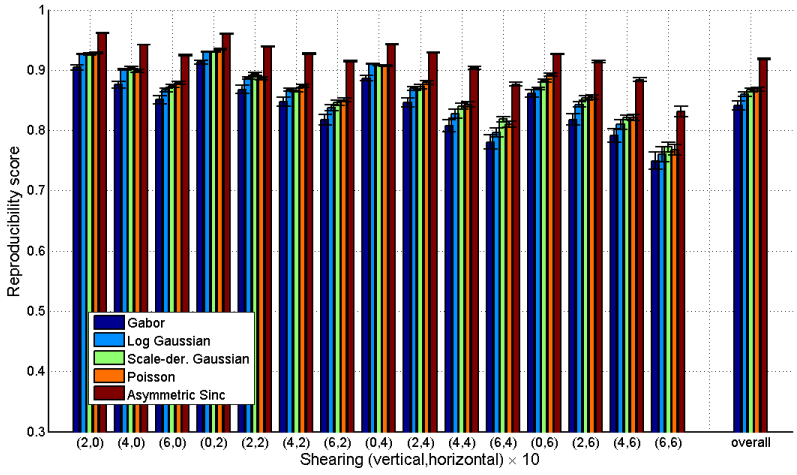


(b) Reproducibility score for different rotation angles for all action types

Figure 3.28: Rotation reproducibility test. This plot shows the reproducibility score (3.7.3) of different temporal filters under a rotation transformation of the video samples of the Weizmann classification dataset [11]. (a) The reproducibility score for each of ten action types is averaged over all samples of the given action and over 8 rotation angles $[-60^o, 60^o]$. The variances are too small to illustrate. (b) The result for each of 8 rotation angles is averaged over 93 video samples. Note that the asymmetric filters, specially the asymmetric sinc, perform better than the Gabor filter under different rotation angles.

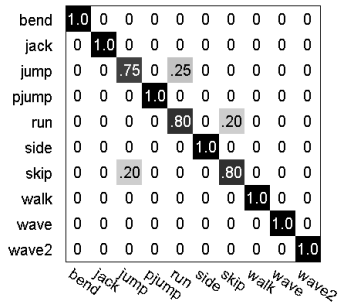


(a) Reproducibility score for each action type using all shearings

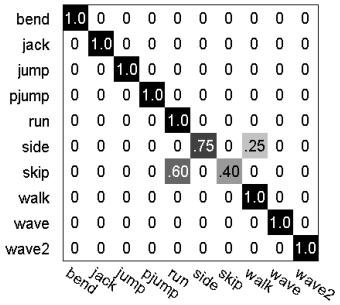


(b) Reproducibility score for different (vertical, horizontal) shearings

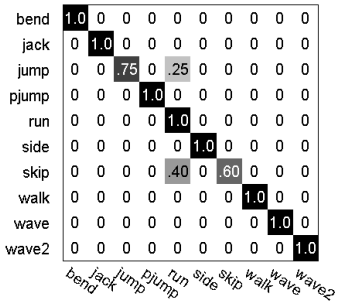
Figure 3.29: Shearing reproducibility test. This plot shows the reproducibility score (3.7.3) of different temporal filters under a vertical and horizontal shearing of the video samples of the Weizmann classification dataset [11]. (a) The reproducibility score for each of ten action types is averaged at 15 shearing transformations (combination of 4 vertical and 4 horizontal shearing)$\{(0.2, 0), (0.4, 0), ..., (0.6, 0.6)\}$. (b) The result for each of 15 shearing transformations (excluding $(0, 0)$) is averaged over 93 video samples. Note that the asymmetric sinc performs the best and the Poisson and scale-derivative Gaussian filters perform slightly better than Gabor under shearing transformation.



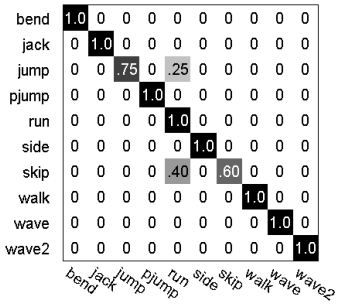
(a) Using Gabor kernel



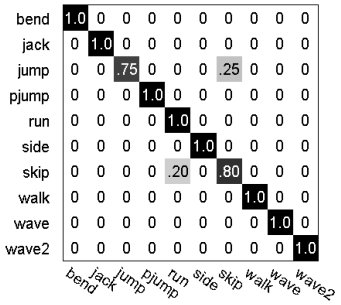
(b) Using log Gaussian kernel



(c) Using scale-derivative kernel

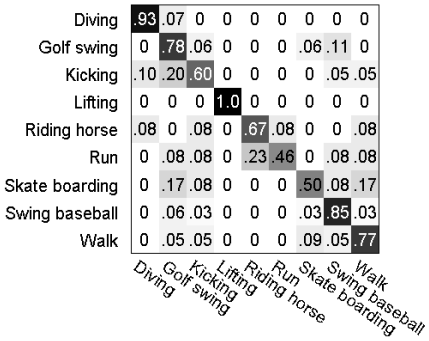


(d) Using Poisson kernel

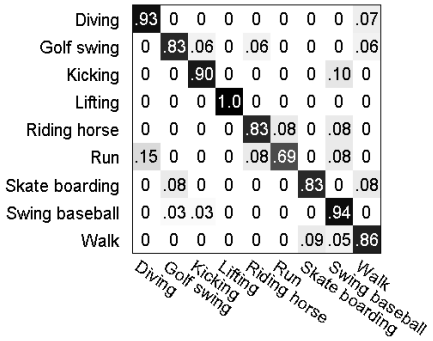


(e) Using asymmetric sinc kernel

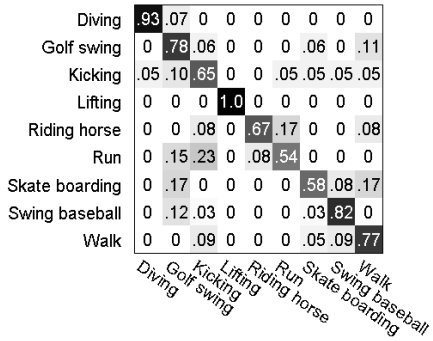
Figure 3.30: Confusion matrices on the Weizmann classification dataset [11]. The scale-space representation for the salient feature detection is obtained using spatial Gaussian kernel and different temporal kernels.Note that symmetric Gabor filter confuses ”skip”, ”jump”, and ”run” with each other and other asymmetric filters have difficulty discriminating ”skip” from ”run”. This is due to the similarity of motion patterns of these fast motions in this dataset. All asymmetric filters perform better than Gabor and the asymmetric sinc is the best.



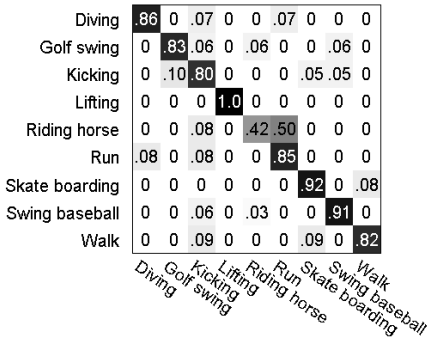
(a) Using Gabor kernel



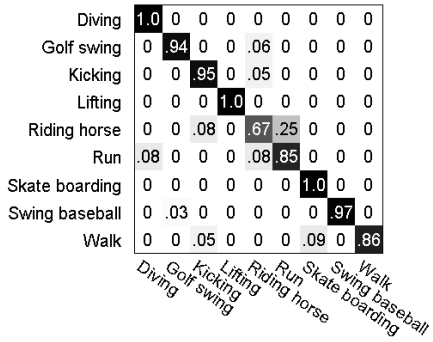
(b) Using log Gaussian kernel



(c) Using scale-derivative kernel



(d) Using Poisson kernel



(e) Using asymmetric sinc kernel

Figure 3.31: Confusion matrices on the UCF sports dataset [107]. The scale-space representation for the salient feature detection is obtained using spatial Gaussian kernel and different temporal kernels. Note that asymmetric since provides higher classification accuracy.

Chapter 4

Action representation

This chapter focuses on the action representation (module B in Fig. 2.2) in a standard discriminative BOW framework. Having detected spatio-temporal salient features at multiple scales (Section 3), the question is how to combine them to obtain a more descriptive and discriminative action representation. This question is important as the classifier uses this representation to discriminate different action types and higher action classification is the objective of a recognition framework. In this chapter, we propose a multi-resolution action signature approach. To this end, the features of all scales are retained. There is therefore motion redundancy among features which occur very close to each other and have significant overlap in their volumetric extensions. The multi-resolution action signature explicitly benefits from the motion redundancy across different scales of salient features. We compare this representation with two existing approaches in which the action signature is computed using scale-invariant features or the combined redundant motion features. We show that the multi-resolution action signature provides a better descriptive and discriminative representation of actions and hence, a higher action classification accuracy.

4.1 Literature review

Action representation is a mid-level modeling of the video contents. Depending on which characteristic of the motion is modeled, this representation might require specific treatment at or after the low-level feature extraction. For example, the scale invariant or redundant features might be of interest for the action representation. Due to the intra-class variations in the motions of a unique action, a robust incorporation of the salient features in the action representation requires that the features be clustered into visual words. The set of visual words constitutes the dictionary or the vocabulary over which an action is represented. A typical action representation is

a histogram representing the frequency of the appearance of the features in a video. This representation shows a global statistical representation of the features and referred to as a BOW representation [68, 82, 118]. There are several methods that incorporate the structural information such as the spatial or the temporal relationships among the features during the clustering stage to obtain a more descriptive representation of the actions. Liu and Shah [47] computed the clusters of the video words by merging the clusters produced by the K-means algorithm and then using the spatio-temporal pyramid matching and spatial correlogram to include the structural information. Jiang and Ngo [133] developed a visual ontology based on the WordNet, a textual ontology which is widely used for text retrieval. An action representation with implicit spatial and temporal localization of features is developed by Marszalek et al. [88] using a multi-channel SVM approach. To this end, multiple channels of the BOW representation with different spatial and temporal extensions are constructed at different locations in the foreground volume. Liu et al. [48] constructed a semantic visual vocabulary. To this end, they used motion salient features of cuboids [25] to obtain a coarse localization of the regions of interest (ROI). They then incorporate the ROIs for the clustering of the static salient features into the vocabulary. Chen et al. [18] revised K-means to incorporate the spatial localization of the features in the construction of the dictionary. Wang and Mori [124] used a hidden conditional random field model to associate the features to implicit body parts. Zhang and Gong [135] used the temporal relationship of the shape features using a structural probabilistic latent semantic analysis framework. Wang et al. [134] developed a dictionary of frame-words by incorporating the global frame features instead of local visual words. Ryoo and Aggarwal [86] used Allen's taxonomy of spatial and temporal relationships between two events such as before, after, and during to describe the structural relationship of salient features.

The incorporation of structural information provides a more descriptive dictionary, but adds to the complexity of the framework. The global BOW is simple, efficient, and general enough to be applied directly to different application domains with minimum changes. In this thesis, a global BOW representation of the video is used. Here, we briefly explain different operations that a method performs to construct a global BOW representation of an action. Two standard approaches for action signature construction are then introduced.

Global BOW action representation

A typical BOW action representation requires the learning of the dictionary of visual words. The inputs to the system are the salient features extracted from the video samples. A salient feature F is localized at (x, y, t) with spatio-temporal scale of (σ, τ) and is described by a feature descriptor such as a 3D SIFT [111] of size $D \times 1$. Fig. 4.1 shows the operations for a typical BOW action representation at the training and testing stages. In the training stage, the systems learn the

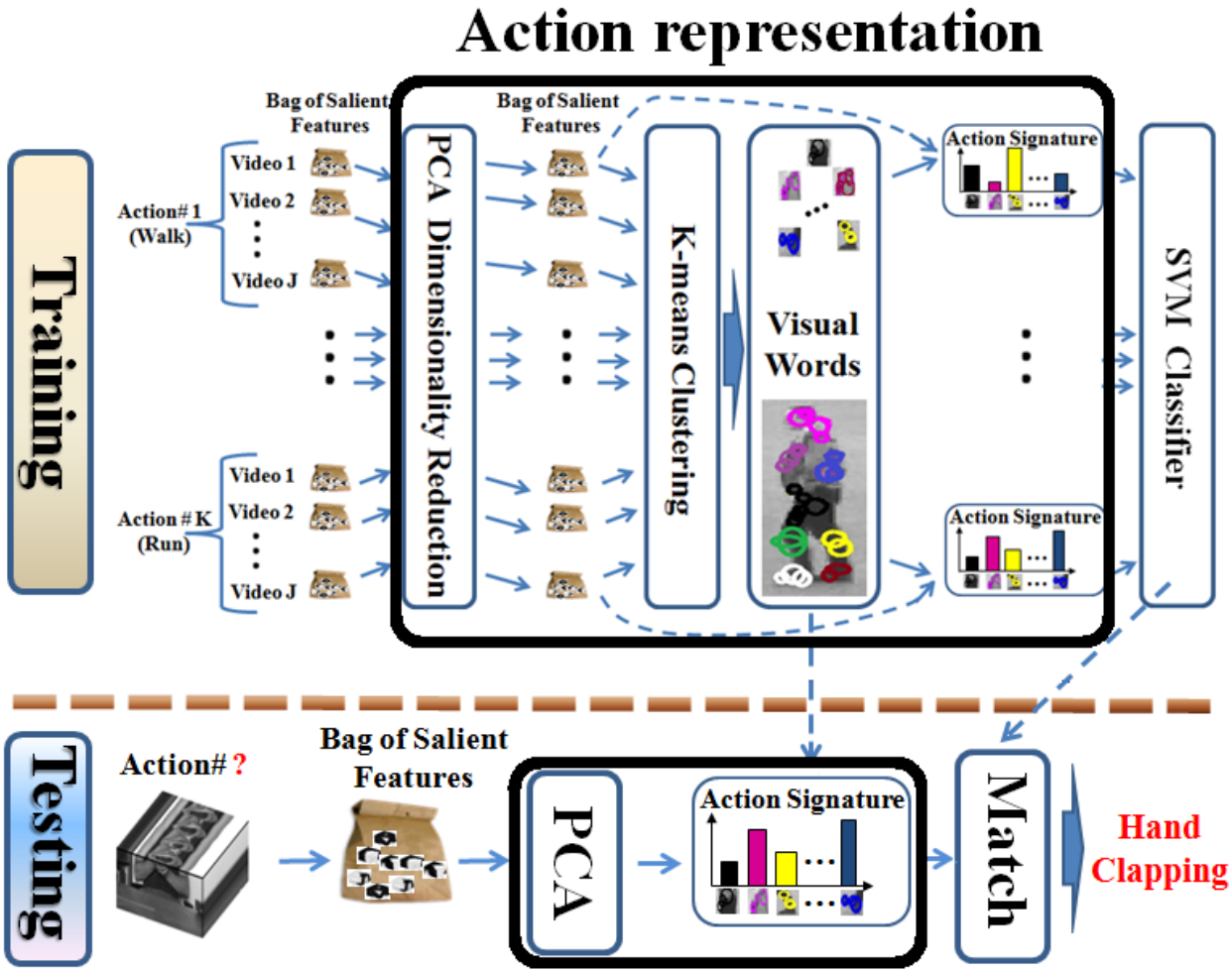


Figure 4.1: The bold dark box shows the typical operations that are performed to obtain a BOW representation of an action. At the training stage, the transformation matrix for projecting the features into a lower dimensional space and the dictionary of visual words are learned. The video of an action is represented by its signature which is the frequency of the salient features. The signatures of different actions are used for the classification.

dictionary from the set of features extracted from all action types. Here, these operations are briefly explained in order.

1. **Dimensionality reduction**. To improve the further matching process of the features and reduce the computation cost [118], the features' dimension should be reduced. In a linear

setting, the transformation matrix A maps each feature descriptor of dimension $D \times 1$ to $d \times 1$ ($d < D$).

$$f_{d\times 1} = A^T_{d\times D} F_{D\times 1} \tag{4.1}$$

Dimensionality reduction requires learning of the transformation matrix A with $d \times D$ dimension using the features from the training video samples. Assume that the feature number i in the video sample $j \in \{1, 2, ..., J_c\}$ of action type $c \in \{1, 2, ..., C\}$ from the training set is denoted by F^c_{ij}. The set of features of a video sample is represented by $V^c_j = [F^c_{1j}, F^c_{2j}, ..., F^c_{n_j j}]$ in which the number of features n_j varies from one video to another depending on the speed of the motion and the length of the video stream. Matrix X contains the collection of all features from all video samples from the training set.

$$X = \left[\; V^1_1, ..., V^1_{J_1}, V^2_1, ..., V^2_{J_2}, ..., V^C_1, ..., V^C_{J_C} \; \right] \tag{4.2}$$

Using principle component analysis (PCA) [13], we obtain a linear projection of the data X to a reduced domain using the eigenvectors of the data distribution. To this end, the mean is first removed from the feature matrix ($\tilde{X} = X - mean(X)$). A singular value decomposition (SVD) [13] can then decompose the data $\tilde{X} = U\Sigma W^T$.

$$[U, \Sigma, W] = SVD(\tilde{X}) \tag{4.3}$$

Each column of the square matrix $U_{D\times D}$ is an eigenvector. For dimensionality reduction, the d dominant eigenvectors form the projection matrix $A_{D\times d} = U(:, 1 : d)$.

2. **Dictionary learning**. The reduced-dimension salient features extracted from the video samples of all actions in the training set are grouped using a clustering method such as K-means [25, 122]. The clustering method splits the features into k clusters with the hope that the features with similar characteristic appearance and motion are grouped together. Each cluster is a visual word (or an action prototype) and the set of these clusters forms the dictionary (or code book) of the visual words. One however should note that the set of action prototypes are not necessarily the genuine set from which the action are generated.

3. **Action signature**. A global BOW representation of a video is the term frequency of appearance of the detected salient features. This signature is a normalized histogram and the bins of the histogram are the learnt visual words. The contribution of each salient feature in the histogram is to the closest visual word with a weight w_{ij} that can be computed using a term frequency measure (4.4) (where N_j is the total number of features in the video j

and N_{ij} counts the number of features from the cluster i). The signature S_j^c of the video j is thus a $k \times 1$-length vector from an action type c.

$$w_{ij} = \frac{N_{ij}}{N_j} \tag{4.4}$$

In the testing stage, the learnt projection matrix A is used to reduce the dimensionality of the features' descriptor. The action signature of the test video is the term frequency of its salient features over the learnt dictionary of visual words. Depending on whether the salient features are scale-invariant or redundant features, a different action signature is obtained.

In a discriminative approach, at the training stage, the action signatures of different actions are fed into a classifier such as a SVM [21] to learn the decision boundaries. At the testing stage, the test signature is categorized as a specific class based on the its statistical similarity with the signatures of the training videos.

4.2 Action signatures

Depending on whether scale-invariant features or a complete set of features at multiple spatio-temporal scales are used for the action representation, different action signatures with different discrimination power are obtained.

4.2.1 Signature of scale-invariant features

A feature maybe best described in its intrinsic scale [59, 81], so one might keep the salient features just at their intrinsic scales [67, 68, 129] and remove them if they appeared in any other scales (Section 3.6). The set of scale-invariant features provides a sparse representation of the video contents. An action signature of a scale-invariant feature might loose its discrimination power [129]. This is due to the fact that the scale selection approach is heuristic [88] and some of the motion features might be incorrectly discarded which may result in the loss of discriminative information.

4.2.2 Signature of redundant features

To avoid the inherent artifact of scale selection [88], one might keep all the salient features [122] and not perform the post-processing scale-invariance operation. This might result in appearance

of features at the same point, but at different spatial or temporal scales. This provides a more redundant representation of the motion information in the video and possibly provide a better reproducibility in the matching of the test videos with those in the training. Note that the contribution of the features at any scale in the action signature can be similar or weighted based on a prior knowledge such as more discriminative power of fine-scale features.

In the next section, we introduce a multi-resolution approach for action signature construction which keeps the motion resolutions of redundant features in the signature to provide a higher discrimination power.

4.3 Multi-resolution action signature

The salient motion features at different spatio-temporal scales encode different characteristics of the motion pattern and might have inherent motion redundancy. For example, the running action contains many motion features occurring simultaneously at different spatio-temporal scales. The entire body gyrates to the running beat while the arms sway in a synchronized fashion. It is quite possible that if only one scale feature was chosen at a spatio-temporal location, these two components, which are essential for defining the action of running, would not be both kept for the classification stage. This is the reason to retain all spatio-temporal features for the action representation and the eventual classification into action recognition. Moreover, the coarse-scale features are more descriptive of average motion patterns, while the fine-scale features capture the details of the motions. For example, actions such as running and jogging have a similar average motion pattern with potentially similar coarse-scale features. The fine-scale features should however provide a better discrimination of these motions (Fig. 4.2). We therefore propose to treat the features of each scale separately and construct an action signature for each individual scale. This choice is motivated by the benefit of motion redundancy across different scales, the elimination of heuristics in scale selection of features with inherent artifacts [88], and the usefulness of multi-resolution analysis [45].

The multi-resolution action representation has drawn inspiration from the texture recognition literature [49, 99, 105]. In this literature, the features generated as part of a multi-resolution decomposition of the image provided promising results in improving the classification accuracy. These techniques benefit from the redundancy of information across different scales for better discrimination of different textures. Similarly, as the redundancy of motion information is utilized in the multi-resolution signature, we expect to obtain higher classification accuracy. Fig. 4.3 shows the concatenation of the action signatures to obtain a multi-resolution action signature. Note that the dictionary of a given scale is constructed from only the features of that specific scale.

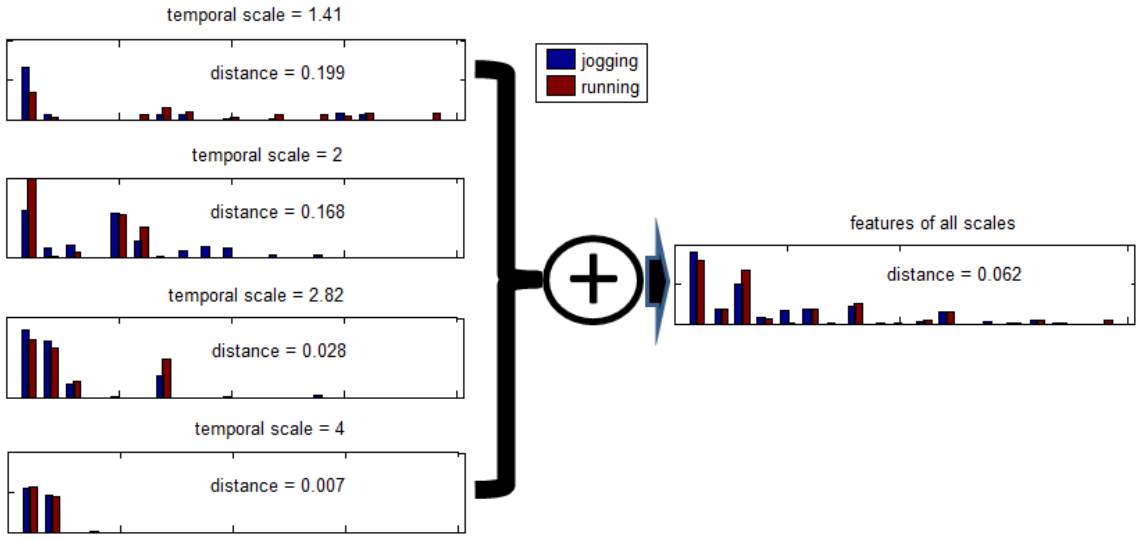


Figure 4.2: Comparing the action signature of a sample video of "jogging" and a sample video of "running" from the KTH dataset [110]. As can be seen (from top to down), the χ^2 distance (3.38) decreases as temporal scale of the features increases (for conciseness, the spatial scale is fixed to $\sigma = 2$ in this figure). The signature of combined features gives a small difference which might result in confusion. In a multi-resolution analysis approach, one can use the signature of all spatio-temporal scales (See Fig. 4.3).

4.4 Experiments

4.4.1 Methodology

We use the cuboids feature detector [25] (Section 3.1.1) and 3D SIFT descriptor [111] (Section 3.1.2) which have shown promising performance for action representation [111, 122]. Similar to the experimental setting in Section 3.7.1, the salient features are detected at nine different spatio-temporal scales. To quantize the salient features into visual words, we perform the K-means clustering with random seed initialization ten times, for each k, and keep the result with the lowest error. The number of clusters k is a parameter of choice from 100 to 1000 with interval of 100. We then construct the corresponding action signature as the L_1-normalized frequency of the occurrences of the features in an action. Note that the Weizmann classification dataset [11] has temporal scale change, but not the spatial scale change since the person does not noticeably change their size relative to the camera frame. The KTH dataset [110] has both spatial scaling

and temporal scaling. The temporal scaling is due to the change in duration of performing an action by different people. A nearest neighbor (NN) classifier is used on the Weizmann dataset [11] and an SVM classifier is used on the KTH dataset [110]. In the implementation, we consider the multi-resolution action signature as a single component for matching of the action signatures. Due to possible spatial and/or temporal scale change in the motion pattern of a test video with that of the training videos, the signature of given scale of test video might be better matched with the signature of another scale of a training video sample. One might therefore perform the search for the best match for the signature of each scale and then find the best overall match of a multi-resolution signature.

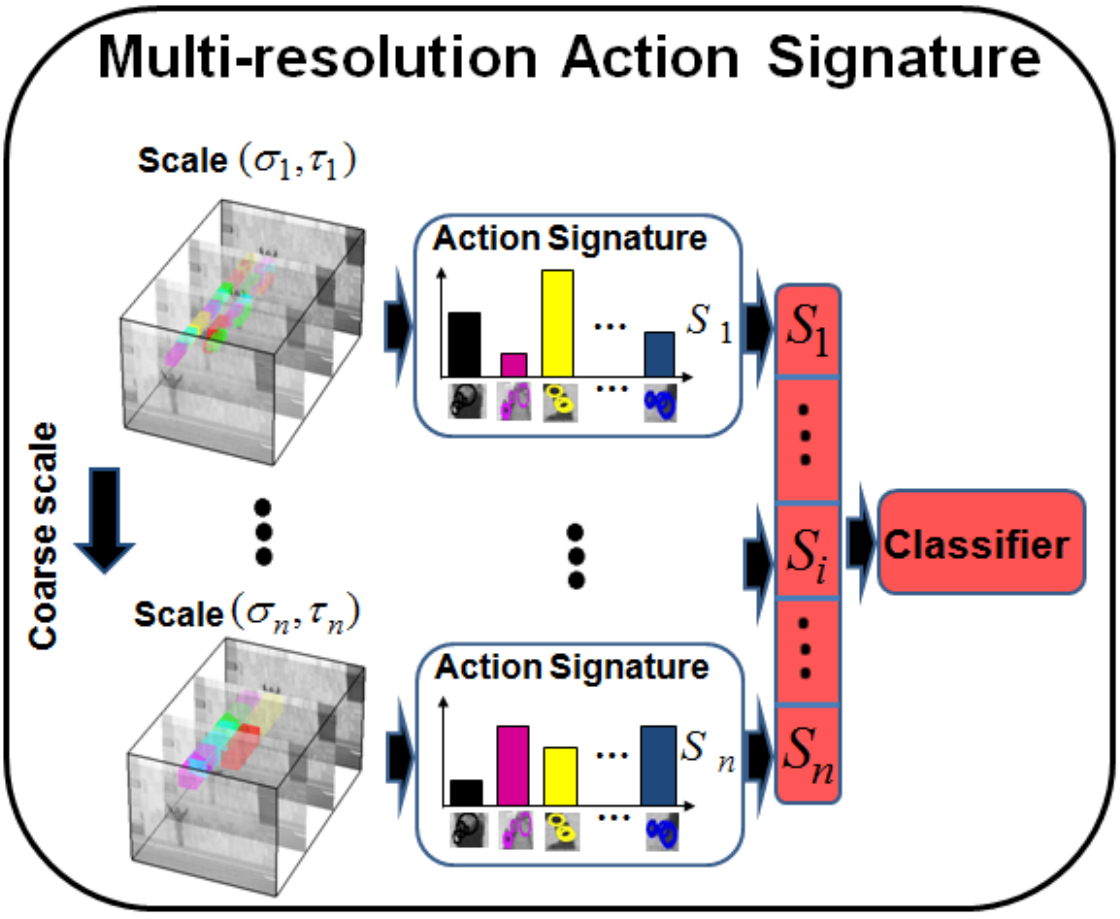


Figure 4.3: In a multi-resolution action signature, the action signature of all scales are concatenated in order. The combined feature space is then fed into a single classifier.

4.4.2 Results

This section provides the experimental results of action classification using different action signature approaches.

Figure 4.4 shows the average classification accuracy as function of number of visual words in the dictionary on the (a) Weizmann and (b) KTH data sets using multi-resolution action signature. The problem with a multi-resolution action signature is that with an increase in the number of scales or the length of the action signatures, their serial concatenation results in a high dimensional action signature with a potential decrease in the classification. That is, for the Weizmann dataset, consider a dictionary with 500 clusters of visual words, for example. With nine spatio-temporal scales, the multi-resolution action signature is a $9 \times 500 = 4500$ dimensional vector. As can be seen in Fig. 3.18(b), on average, there is less than 30 salient features per second and hence, less than 60 features in a typical two-second video sample. And, there is at most ten video samples for each class of actions. The classification accuracy might degrade due to the curse of dimensionality problem [49], as there might not be enough sample features for each cluster of visual words at the training stage. The multi-resolution action signature is also quite sparse and a proper distance metric plays a significant role in the matching of the signatures. Reducing the dimensionality of the feature space might not be however helpful as the motion discrimination power of the feature space might reduce, and consequently, the classification accuracy might degrade. In fact, there is a trade off as the classification accuracy degrades due to both the curse of dimensionality problem and the low dimensional feature space. This problem is less obvious in the KTH dataset as there more samples features from different clusters due to longer video sequences (between 10 to 15 seconds) and more number of videos (64 samples for the training) from each action classes.

Table 4.1 shows the average action classification accuracy on different data sets for three different strategies in utilizing multi-scale salient features. As can be seen, the multi-resolution action signature performs better than the action signature of using redundant features. Moreover, the performance of using an action signature of scale-invariant features is the lowest. This shows that removing the motion redundancy across different features to obtain a sparser representation of the video contents does not help the discrimination of motions.

Fig. 4.5 shows the confusion matrices for action recognition on the Weizmann classification dataset [11] using different action signatures. Note that all methods have difficulty discriminating "skip" from "run" or "jump". This is due to the similarity of motion patterns of these fast motions in this dataset. The multi-resolution action signature performs the best. Fig. 4.6 shows this comparison for the KTH dataset [110]. As can be seen, the multi-resolution action signature performs better in this dataset as well. More specifically, the multi-resolution action signature significantly performs better than the other type of action signatures, providing more evidence

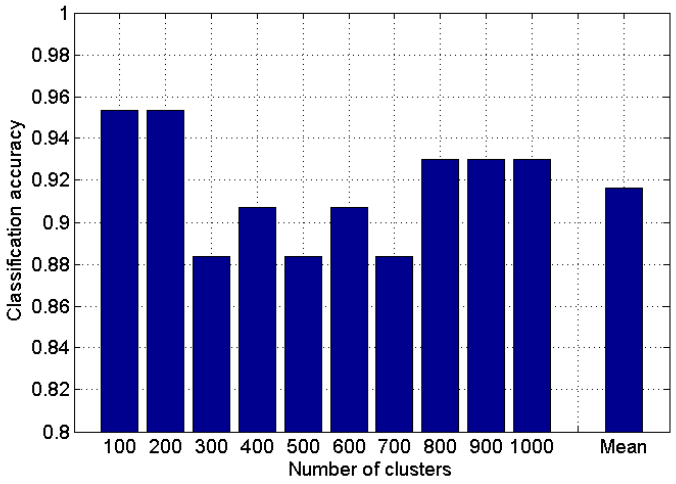
why we should retain all the features and use a multi-resolution analysis approach. Note that there exist both temporal scaling and spatial scaling in the KTH dataset. Fig. 4.7 shows the confusion matrix for the UCF sports dataset [107] for the case in which the redundant features, the scale-invariant features, or a multi-resolution action signature has been used for the action representation. Again, the multi-resolution action signature provides the highest classification accuracy.

Table 4.1: Average classification accuracy on different data sets. Note that the multi-resolution action signature outperforms the approaches which use the scale-invariant features or combined redundant features. The average is over 30 runs. The standard deviations are in order of $10^{-}3$ and are not reported here.

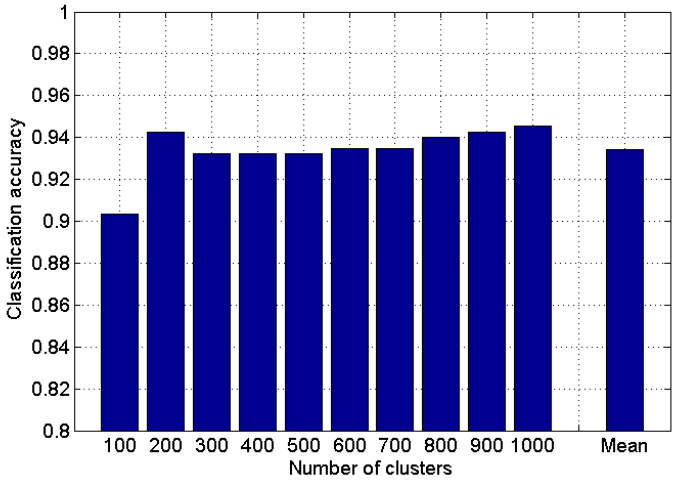
Classification	Type	KTH	Weizmann	UCF sports
Single classifier	with scale-invariant features (Sec. 4.2.1)	89.1 %	90.2 %	71.3 %
	with redundant features (Sec. 4.2.2)	89.7 %	91.1 %	73.3 %
	with multi-resolution action signature (Sec. 4.3)	93.4 %	91.6 %	75.6%

4.5 Conclusion

In this chapter, we introduced the concept of multi-resolution action signature to provide a more discriminative action representation. We showed that retaining the multi-scale features and encoding the motion resolution in the action signature results in an improvement of action classification. The high dimensional multi-resolution action signature may potentially suffer from the curse of dimensionality, when there is not enough samples for each visual words at the training stage. In the next chapter, we propose the multiple classifier systems approach to solve this problem.

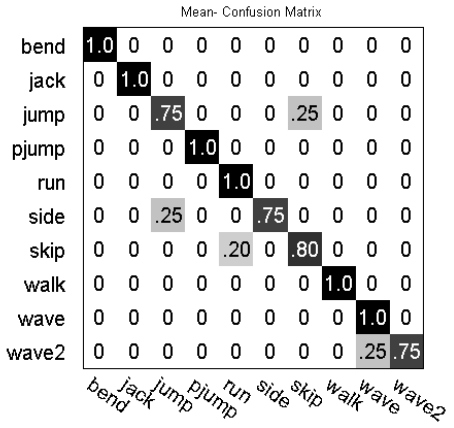


(a) Classification accuracy on the Weizmann dataset

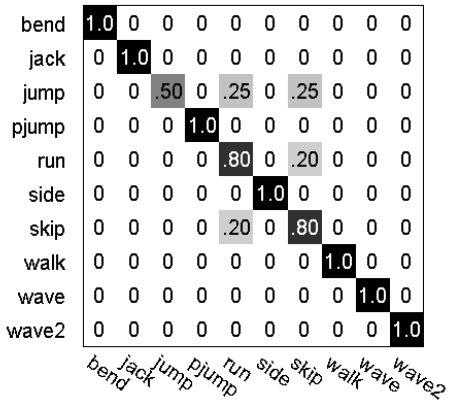


(b) Classification accuracy on the KTH dataset

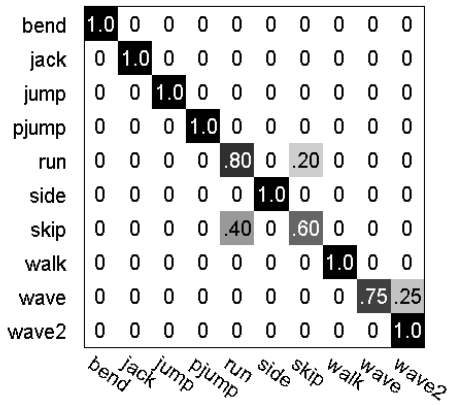
Figure 4.4: The classification accuracy using multi-resolution action signature versus the number of visual words in the dictionary on the (a) Weizmann and (b) KTH data set. The last bar shows the average accuracy over all classifiers. The highest accuracy on the Weizmann datast is obtained with a dictionary with 100 or 200 visual words. A possible explanation of decrease in accuracy at 300 or 500 is that there might not enough samples of non-zero clusters of visual words, and hence, the curse of dimensionality might occur.



(a) Using scale-invariant features

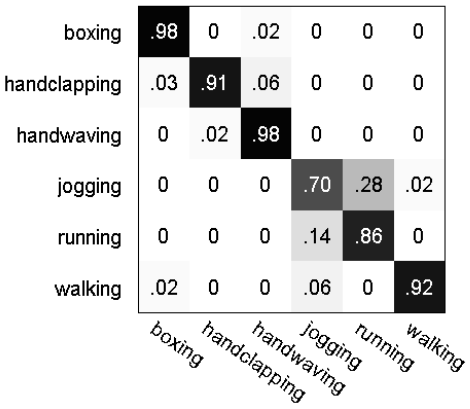


(b) Using combined redundant features

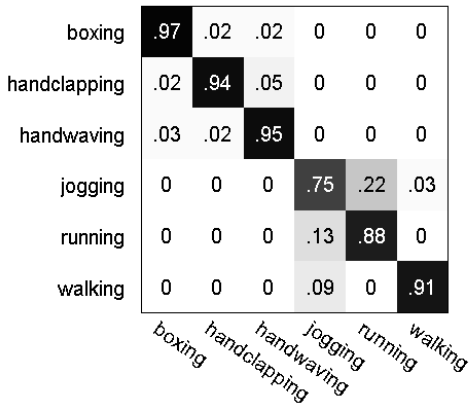


(c) Using multi-resolution action signature

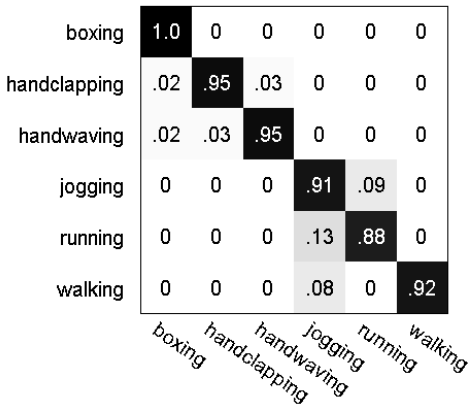
Figure 4.5: Confusion matrices for action classification on the Weizmann classification dataset [11] using (a) an action signature of scale-invariant features, (b) an action signature of combined redundant features, and (c) a multi-resolution action signature. Note that the multi-resolution action signature performs the best.



(a) Using scale-invariant features

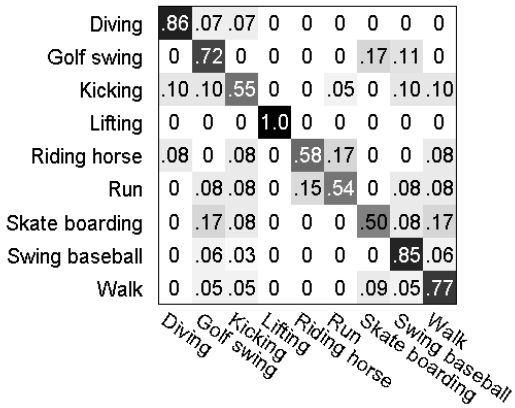


(b) Using combined redundant features

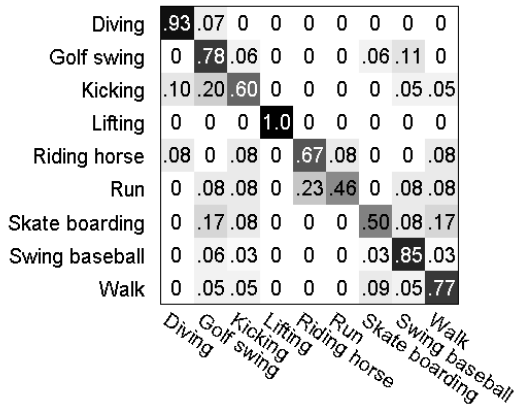


(c) Using multi-resolution action signature

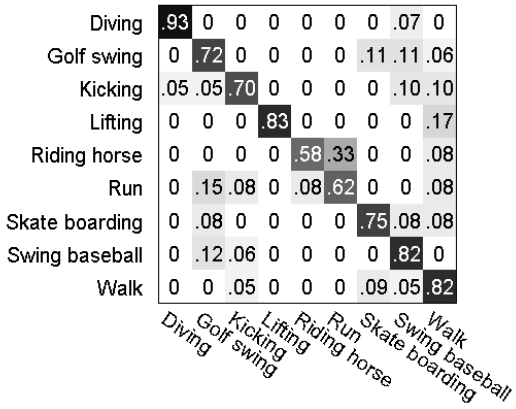
Figure 4.6: Confusion matrices for action classification on the KTH dataset [110] using (a) an action signature of scale-invariant features, (b) an action signature of combined redundant features, and (c) a multi-resolution action signature. Note that the multi-resolution action signature performs the best. Moreover, it discriminates similar actions such as ”jogging” and ”running” much better than the other action signature types.



(a) Using scale-invariant features



(b) Using combined redundant features



(c) Using multi-resolution action signature

Figure 4.7: Confusion matrices for action classification on the UCF dataset [107] using (a) an action signature of scale-invariant features, (b) an action signature of combined redundant features, and (c) a multi-resolution action signature. Note that the multi-resolution action signature performs the best.

Chapter 5

Action classification

This chapter introduces a multiple classifier systems (MCS) framework as an alternative to existing classification approaches which use a single classifier [25, 68, 122] for human action recognition (module C in Fig. 2.2). In this framework, the features of each spatio-temporal scale are treated independently and fed into a separate classifier. The final decision is then made by combining the results of all classifiers. This way, we hypothesize that the MCS provides a better approach for the combination of multi-scale salient features. We then demonstrate that the MCS framework can address the problem of choosing the proper length for the dictionary of visual words at the action representation level and the choice of a distance metric for the action signature matching at the classification stage. These considerations are based on the reformulation of the problem of choosing a single appropriate parameter for the recognition task to use a set of parameters and then combine the results. Consequently, we introduce the multi-stage MCS framework as a parameter-free multi-resolution analysis action recognition scheme.

5.1 Literature review

In a systematic design of a discriminative action recognition framework, the choice of the classifier scheme affects the choice of the features and/or the choice of the action representation approach. Most of the existing discriminative bottom-up action recognition approaches use a single SVM classifier to discriminate different actions from each other [25, 68, 112]. In this setting, the salient features of different spatio-temporal scales are combined together at the beginning of the action representation level. Consequently, a single action signature is fed into the classifier. In Chapter 3, we showed that using robust salient features can improve the final classification accuracy. In Chapter 4, we showed the importance of the multi-resolution representation

of human motions for a better classification. In this chapter, we mainly focus on the choice of the classification approach.

In the existing method [25, 68, 112, 122], the support vector machine (SVM) classifier has been widely used for human action recognition. Here, we briefly explain this classifier.

5.1.1 Support vector machine (SVM)

SVM has been widely used for classification of different visual categories [25, 68, 82]. The original SVM has been proposed for a binary classification problem. In the case of multiple-class classification, the efficient approach of one-against-all can be used [21, 119]. In this approach a separate SVM is learnt for each individual class. The problem then breaks down into a binary classification problem in which the training samples of a given class are the positive samples and the rests are the negative samples. During testing, a test sample is fed into all SVMs and the one with higher classification accuracy determines the class label. Here, we briefly explain the binary SVM for the classification of two classes.

Consider the problem of separating a set of training samples $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i \in \Re^N$ is a feature vector (in our case, $\mathbf{x}$ is the action signature S) and each sample is assigned to one of two classes $y_i \in \{+1, -1\}$. Let's assume that the two classes can be separated using a hyperplane[1] $\mathbf{w}.\mathbf{x} - b = 0$, in which $\mathbf{w}$ is the normal vector, perpendicular to the hyperplane, and $b/|\mathbf{w}|$ is the offset of the hyperplane from the origin along the normal vector. In absence of any prior knowledge about the distribution of the data, the optimum hyperplane is the one which has the longest distance to the closest data of both classes (so-called functional margin) [21, 119]. The maximum-margin hyperplane assures the lowest bound on the expected generalization error [2] of the classifier. To maximize the distance $(2/|\mathbf{w}|)$, we need to minimize $|\mathbf{w}|/2$. The objective is then to find the optimal values of $\mathbf{w}$ and b using the following constrained optimization problem.

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \qquad subject\ to \quad y_i(\mathbf{w}.\mathbf{x}_i - b) \geq 1 \quad \forall i = 1, ..., n. \tag{5.1}$$

Using the non-negative Lagrange multipliers α_i, the problem is stated as

$$\min_{\mathbf{w},b} \max_{\alpha_i} \{\frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{w}.\mathbf{x}_i - b) - 1]\} \tag{5.2}$$

[1] A hyperplane separates the feature space into two parts. For example, a line is a hyperplane in a 2D space or a plane is a hyperplane of a 3D space.

[2] The generalization error is usually defined as the expected value of the square of the difference between the learned function and the exact target (mean-square error).

The corresponding α_i must be set to zero for those points that can be separated as $y_i(\mathbf{w}.\mathbf{x}_i - b) - 1 > 0$. This include most of the points. The points with non-zero α_i value are referred to as the support vectors. By standard quadratic programming [21, 119], the normal vector is computed as the linear combination of the training data.

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{5.3}$$

The support vector lies exactly on the margin $y_i(\mathbf{w}.\mathbf{x}_i - b) - 1 = 0$. The offset b can be computed accordingly.

$$\mathbf{w}.\mathbf{x}_i - b = 1/y_i = y_i \Longleftrightarrow b = \mathbf{w}.\mathbf{x}_i - y_i \tag{5.4}$$

A more robust estimation of the offset is however obtained as the average over all N_{SV} support vectors.

$$b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (\mathbf{w}.\mathbf{x}_i - y_i) \tag{5.5}$$

In the case where two classes are not linearly separable, one can map the data from $\Re^N$ to a high dimensional data space H by $\mathbf{x} \rightarrow \Phi(\mathbf{x})$ in which the data become linearly separable. Computing such a mapping is not always possible as the feature space could be an infinite-dimensional space. The representer theorem [119] shows that $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \Phi(\mathbf{x}_i)$. Instead of direct minimization over $\mathbf{w}$, the optimization can be done for the α. The decision rule is thus $f(\mathbf{x}) = sgn(\sum_{i=1}^{n} \alpha_i y_i \Phi(\mathbf{x}_i).\Phi(\mathbf{x}) - b)$. In return, as we are interested in computing just the inner product, we can use a Mercer kernel[1] K such that $K(\mathbf{x}_i, \mathbf{x}) = \Phi(\mathbf{x}_i).\Phi(\mathbf{x})$. This turn around is usually referred to as the kernel trick. A non-linear SVM finds an optimum hyperplane in the feature space as

$$f(\mathbf{x}) = sgn(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) - b) \tag{5.6}$$

Most of α_i's are zero; those with non-zero value are the support vectors. Among different Mercer kernels, the linear, RBF (Gaussian), sigmoid, (histogram) intersection [87], and exponential χ^2 kernels (5.7-5.11) have shown promising performance for the comparison of two samples $\mathbf{x}$ and $\mathbf{z}$ for different visual recognition tasks [22, 25, 46, 52, 68, 112, 110]. One however should note that choosing the right kernel is application dependant and the fine tuning of the parameters might be

[1] A Mercer kernel is positive semi-definite, meaning that all of the eigen values of its kernel matrices are non-negative. The use of a Mercer kernel insures that the optimization problem is convex and there is a unique solution for that [21, 119].

required. Instead of learning these parameters [120], we choose another approach based on the MCS. Here, we list the Mercer kernels that will be used in Section 5.2.3 for this purpose.

$$K_{Linear}(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T\mathbf{z} = \sum_i x_i z_i \tag{5.7}$$

$$K_{RBF}(\mathbf{x}, \mathbf{z}) = e^{-\gamma|\mathbf{x}-\mathbf{z}|^2} = e^{-\gamma\sum_i |x_i - z_i|^2} \tag{5.8}$$

$$K_{Sig}(\mathbf{x}, \mathbf{z}) = tanh(a\mathbf{x}^T\mathbf{z} + r) \tag{5.9}$$

$$K_{Inter}(\mathbf{x}, \mathbf{z}) = \sum_i min(x_i, z_i) \tag{5.10}$$

$$K_{\chi^2}(\mathbf{x}, \mathbf{z}) = e^{-\gamma\chi^2(\mathbf{x},\mathbf{z})} \ , \qquad \chi^2(\mathbf{x}, \mathbf{z}) = \sum_i \frac{|x_i - z_i|^2}{|x_i + z_i|} \tag{5.11}$$

5.1.2 Design of the classification system

There are several factors that directly affect the design of a classification scheme. Among those, the dimensionality of the feature space (i.e., the number of clusters in our case), the type of the classifier, the type of distance metric, and the number of training samples play a significant role in the final classification result. There are some generally accepted priors such as a better performance of SVM with a high-dimensional (nonlinear) feature space, non-suitability of a Euclidean distance for a non-Euclidean feature space such as normalized histograms, or possibly low performance of a SVM with very few training samples. These considerations help in the design of a classification system, but they cannot be generalized for all data types.

For the configuration of a bottom-up human action recognition system, we focus on three parameters that affect the final classification accuracy. These parameters are aside from the direct effect of the robust salient features on the classification accuracy.

1. **The choice of the classification approach**. Most existing methods use a single classifier [25, 68, 112, 122]. In Section 4.3, we argued why a multi-resolution analysis on

the multi-scale motion features should provide a more discriminative action representation. More specifically, we introduced the multi-resolution action signature which shows improvement over the (single-level) action signatures. The potential curse of the dimensionality problem of the multi-resolution action signature might degrade the classification accuracy (Fig. 4.4). The multi-resolution action signature is also fed into a single classifier. Here, the objective is to determine whether a single classifier or a set of classifiers can better characterize the multi-resolution nature of the human motions.

2. **The choice of the dictionary length**. The feature vector which is fed into the classifier is the action signature with the same dimension as the number of visual words (i.e., action prototypes) in the dictionary. Actions are better represented using a bigger dictionary with more detailed motion information. A typical classifier might however be sensitive to the dimensionality of the feature space considering the number of training samples, for example. Therefore, it is not definite that a dictionary with a larger/smaller number of action prototypes provides a better action representation and classification for different actions performed in different scenarios. This is due to the fact that the learnt dictionary is not necessarily the genuine set of action prototypes that has been used in performance of an action. In addition, in discrimination of similar motions such as running and jogging, a bigger dictionary is more discriminant as it has action prototypes with more detailed motion characteristics. In the discrimination of actions such as walking and running, a small set of action prototypes should be sufficient. Considering the computational cost and the constraint on the number of training samples, the proper choice of the dictionary length is an open problem. Finding the right number of action prototypes is an open problem in pattern recognition for clustering methods [91]. There are few approaches such as gap statistics [116], jump method [114], or the prediction strength method [117] to automatically find the number of clusters. The problem however is that no specific algorithm perform well on all data types and with all clustering methods, and hence, makes them application dependant.

3. **The choice of a distance metric**. A good classifier requires a distance metric that properly captures the statistical difference between the representations of similar/dissimilar actions. The choice of distance metric is not trivial as it directly affects the learning of the decision boundaries and hence, the classification performance. Knowledge of the data statistics helps in choosing the right distance metric. For example, the Euclidean distance is not a proper distance metric to compare two normalized histograms [17]. In action recognition literature, the exponential χ^2 and the radial basis function (RBF) are among the most widely used distance metrics for the comparison of the BOW action signatures using a SVM classifier [25, 68, 122]. The linear kernel is a special case of the RBF kernel [52].

The sigmoid function is a popular kernel due its wide use in neural networks. The sigmoid kernel is however quite sensitive to the parameter tuning [74]. The behavior analysis of this kernel in [74] shows that a specific range of parameters ($a > 0$ and $r < 0$ in equation 5.9) is more suitable for this kernel. More specifically, the kernel is not positive semi-definite, but it is conditionally positive definite when r is small enough. In this case, when a is close to zero, the sigmoid behaves like an RBF kernel. However, the RBF kernel typically performs better and hence, it is usually recommended as a typical kernel for the SVM. The (histogram) intersection kernel has also shown promising performance in the visual recognition [87]. The RBF, sigmoid, and exponential χ^2 kernels have a scale parameter to be tuned, while the linear and intersection kernels are parameter free and hence, attractive from the computational cost point of view.

For clarification, we use both terminologies of distance metric and similarity kernel according to the context. In essence, the distance metric is the opposite of the similarity kernel. For matching of the patterns in a classifier, two patterns are from the same category if they have the least distance or the maximum similarity, statistically. Moreover, we use the term "dictionary length" and the "number of clusters" of the visual words interchangeably.

In the next section, we introduce an MCS approach to address the above-mentioned problems of choice in the existing discriminative approaches.

5.2 Multiple classifier systems (MCS)

An MCS framework uses a combination of classifiers to provide a decision which potentially outperforms the decision of the best classifier. In fact, there are two reasons for using a combination of classifiers: efficiency and accuracy. In a large scale classification problem with complex patterns, one requires an efficient classification approach. In a hierarchical classification setting, for example, one uses a coarse-to-fine approach for the gradual reduction of the set of possible classes. To classify a test pattern, one starts with a simple and cheap classifier with possibly a low-dimension feature space to reduce the number of possible classes to which the pattern might belong. Using a high dimensional feature space with a more complex classifier, the set of possible classes is then narrowed down to one. Higher accuracy is another objective of designing a pattern recognition system using combination of classifiers. One classifier might be designed to perform very well for a subset of patterns using a given feature set. Another classifier might perform better for another subset of patterns with another feature set. This suggests that a combination of classifier designs with potentially complementary information about different patterns should provide higher accuracy [56].

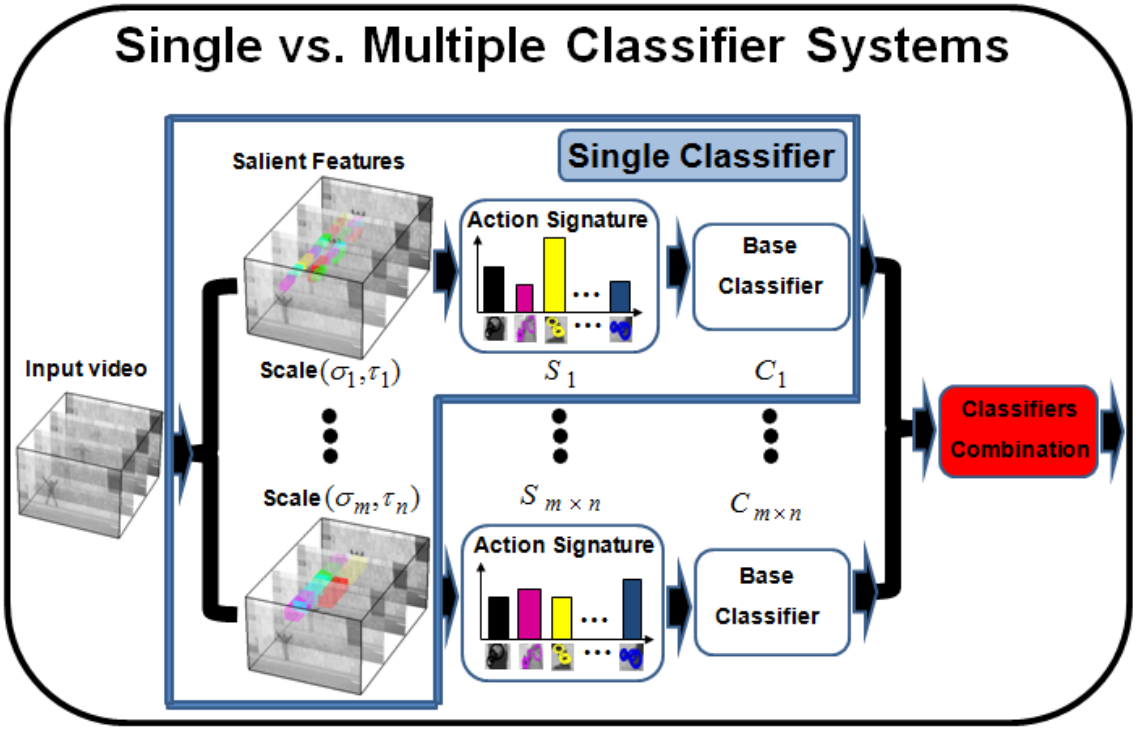


Figure 5.1: In existing bottom-up action classification methods [68, 129, 122], the features of all scales are fed into a single classifier (solid red block). In this chapter, we propose the MCS framework to benefit from the motion redundancy across different scales for better action classification. In our MCS framework, the most common class label across the classifiers of different scales is the final decision.

There are two main factors in the configuration setting of an MCS. First, the classifiers are complementary, which means that the set of miss-classified actions by one classifier is different from that of the other classifiers. Moreover, the classifiers which use different feature sets obtained from different measurement modalities are more complementary. Second, the combination rule must be carefully selected to most effectively fuse the classifier decisions.

The design of an MCS is based on a decision optimization or a coverage optimization approach [65]. The decision optimization approach finds the proper combining rule for a given ensemble of classifiers. The coverage optimization finds a proper construction of an ensemble of classifiers to be used with a given combining rule. The approaches for the combination of classifiers [56, 65, 101, 108] are different in terms of (a) the type of output they combine (abstract

level, ranking level, or measurement level), (b) the system topology (serial, parallel, or hybrid), and (c) the degree of a priori knowledge they use. Fig. 5.2 shows the taxonomy from which different approaches might be derived for the classifiers combination.

- **Output type**. In an abstract-level combination method, the top candidate estimated by each classifier is used. In a ranked-level combination method, the entire ranked list of candidates is used. A measurement-level combination method uses the confidence level of each candidate in the ranked list.
- **System topology**. A combination method might use one of three topologies of serial, parallel, or hybrid. In a serial topology, the classifiers are arranged in cascade and the output of each classifier is the input to the classifier at the next stage. In a parallel topology, the outputs of all classifiers are combined in parallel. A hybrid topology uses both configurations.
- **Prior knowledge**. A non-parametric classifier combination method does not require any kind of a priori information on the set of classifiers. In contrast, a parametric method requires information at the level of each individual classifier or on the behavior of the entire set of classifiers.

The MCS has shown promising performance in a variety of applications such as texture analysis [49], identity recognition using different biometric modalities [56], and remote sensing [7]. When the input pattern of the classifiers are independent, a MCS is expected to perform better than each individual classifier in the set. This is due to the fact that the collective behavior of the classifiers has more information [101]. For details on the bounds on the the generalization error of combined classifiers refer to [61, 62, 66].

In the next subsections, we introduce an approach to use MCS framework for the combination of multi-scale features, the choice of dictionary length, and the choice of distance metric. We investigate the performance of a MCS classification approach knowing that the feature sets which are feed into individual classifiers in the framework are not independent across spatio-temporal scales, different dictionaries, or distance metrics.

5.2.1 MCS for the combination of multi-scale features

The salient features at different spatio-temporal scales capture different motion characteristics. Instead of combining these features at the action representation level, a MCS framework keeps the motion information all the way up to the classification stage. In this framework, the features

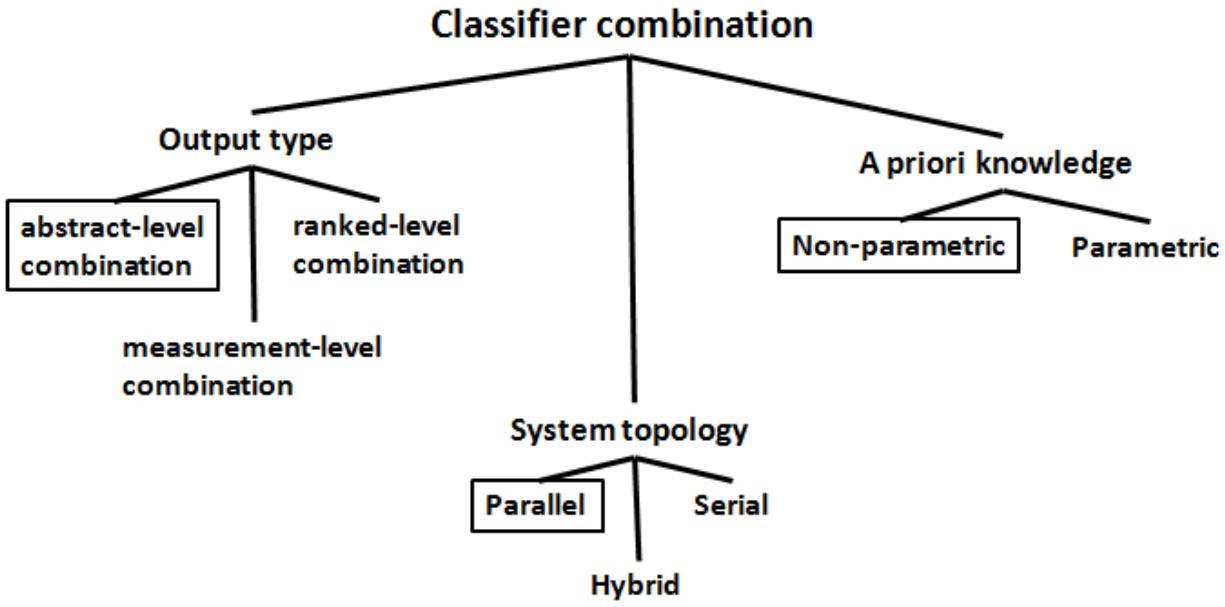


Figure 5.2: A taxonomy for the classifier combination can be expressed in terms of (a) the type of outputs they use, (b) the system topology, and (c) whether they require any prior information. The solid boxes show the approaches that we consider in this thesis: a non-parametric abstract-level classifier combination method in a parallel topology.

of each scale are treated as a separate observation of an action and a separate classifier provides an estimation of the class label accordingly. The classifiers at different scales are then combined for the final decision. The observation is that the set of actions which are misclassified by the classifier of a given scale is probably different from those misclassified by a classifier from another scale. This suggests that the set of all classifiers should provide a complementary information about different actions and hence, a consensus decision should provide a higher accuracy.

Fig. 5.3 shows the confusion matrices on the KTH dataset when a nonlinear (RBF) SVM classifier used the features of only one spatio-temporal scale. For each scale, a dictionary with 400 visual words was learnt from the features of that specific scale. As can be seen, each classifier has different performance on different actions. For example, the classifier which used the features at scale (4,3) perfectly recognized the ”hand clapping” action while its performance was the worst for the classification of ”walking”. The classifier that used the features at scale (4,4) had the highest accuracy in the recognition of ”walking”, but it had difficulty recognizing ”hand clapping”. Moreover, it had the least performance recognition ”running”, while the classifier of

scale (3,2) performed much better on this action. This observation is in support of the evidence that different classifiers perform differently for different actions, and hence, a collection of the classifiers should perform better. In fact, as Fig. 5.4 shows, indeed the classifier combination of a MCS with a simple majority voting scheme outperforms all other nine classifiers.

Fig. 5.1 shows our MCS framework in a parallel topology which combines the classifiers at their abstract-level. Note that the salient features at different scales have overlap and hence, their action signatures are not independent. It is not therefore guaranteed that MCS always performs better than the best classifier. In Section 5.2.1, we show the experimental results showing that the MCS still preforms better than the existing methods which use a single classifiers. Note that there is an overlap among the salient features at different scales, and hence, their action signatures are not independent. It is not therefore guaranteed that the MCS always provides higher accuracy than the best individual classifier.

The independent treatment of the features of each of the scales in our MCS framework differs from the boosting approaches that use multiple classifiers [85]. For example, an adaptive boosting algorithm (AdaBoost [132]) uses a set of weak classifiers on the whole set of features to improve the classification accuracy. In this algorithm, the subsequent classifier focuses on the misclassified instances. AdaBoost is sensitive to noisy data and outliers. Our MCS approach is quite general and can use an AdaBoost classifier as its base classifier.

5.2.2 MCS for the combination of multi-length dictionaries

The choice of dictionary and its length determines the descriptiveness of the action representation. Most of the existing methods set the length of the dictionary (i.e., the number of clusters of the visual words) experimentally. As can be seen in Fig. 5.5 and Fig. 5.6, different classifiers with different distance metrics perform differently with dictionaries with different lengths. We reformulate the problem of choosing a proper dictionary length to using a set of dictionaries with different lengths and then combining them using a MCS.

5.2.3 MCS for the combination of multiple distance metrics

The choice of distance metric directly affect the classification accuracy. Learning or selecting the best distance metric reduces the generality of the classifier. We reformulate the problem of choosing the proper distance metric to use a set of distances and then use a MCS to combine the classification results of different distance metrics. We perform separate classification using each individual distance metric and then combine the results using a MCS to make the final decision about the class label.

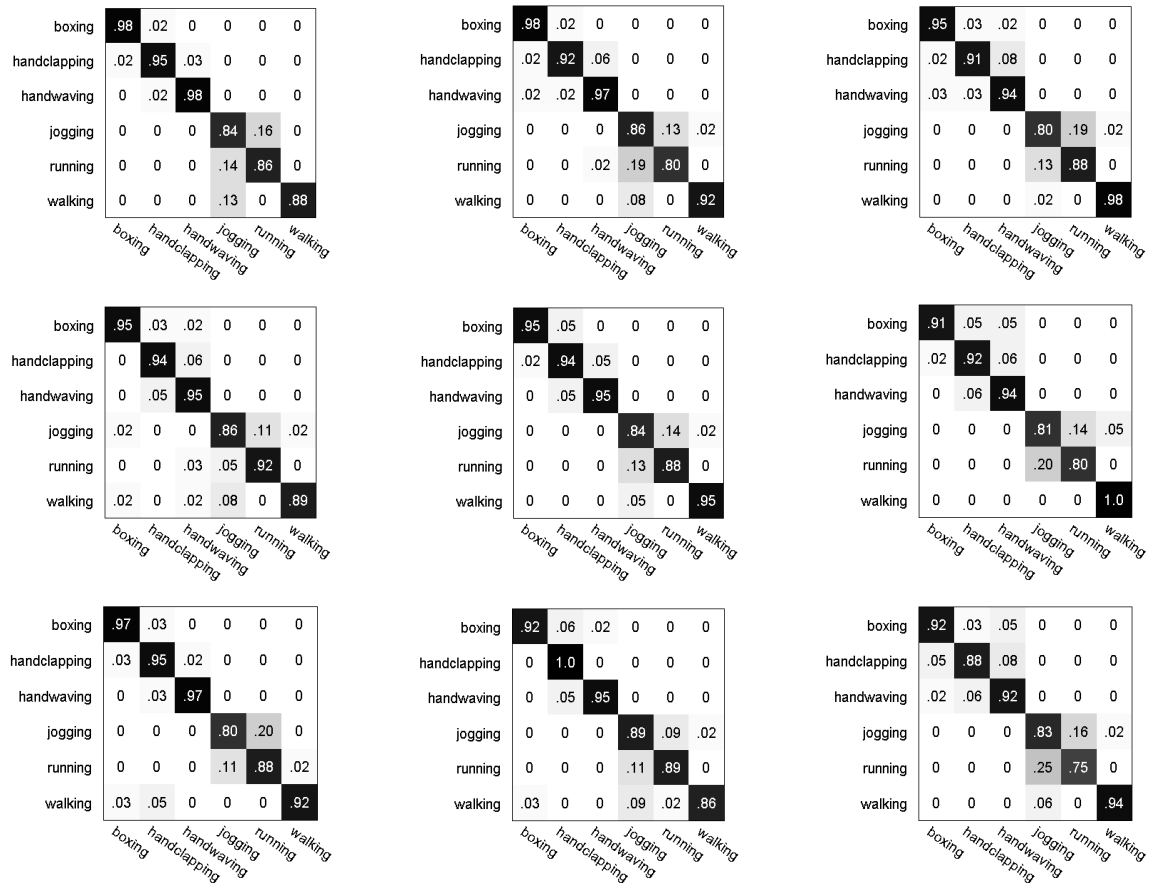


Figure 5.3: Confusion matrices for classification on the KTH dataset when a nonlinear (RBF) SVM classifier uses the features of just a single spatio-temporal scale (From top to bottom, left to right: nine spatio-temporal scales are (2,2), (2,3), (2,4), (3,2), (3,3), (3,4), (4,2), (4,3), (4,4)). Note that each classifier performs differently for different actions. For example, note the two last confusion matrices (bottom left). The classifier which uses the features of scale (4,3) perfectly recognizes the ”hand clapping” action, while its performance is worse for the recognition of ”walking”. In contrast, the classifier of scale (4,4) has the highest accuracy rate in recognition of ”walking”, but it is not that successful to recognize the ”hand clapping” action. As the MCS uses a combination of classifiers, the collective behavior of the classifiers should provide higher accuracy.

Note that our approach for combination of kernels is an alternative approach to the multiple kernel learning (MKL) approach [120]. The MKL framework learns a linear combination of a

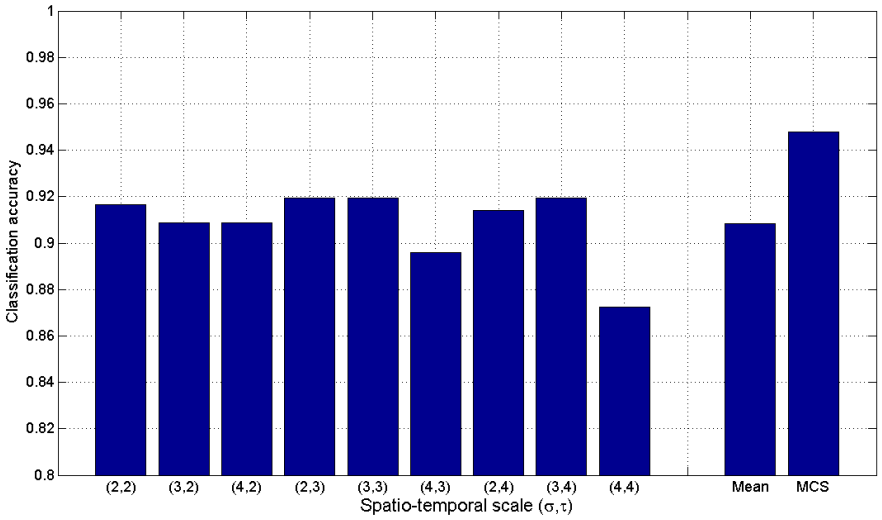


Figure 5.4: This plot shows the average classification accuracy on the KTH dataset when a nonlinear (RBF) SVM classifier uses the features of only one spatio-temporal scale. The two last bars show the mean of the nine classifiers and the performance of a MCS. The MCS performs the majority voting on the prediction of all nine classifiers. Note that the MCS outperforms all the individual classifiers.

predefined finite set of kernels to obtain the final kernel. For a classification task, our approach treats each kernel independently and uses a separate classifier for each kernel. The results of the classifiers are then combined for the final decision. In contrast, the final kernel of MKL feed into a single classifier. Moreover, the MKL requires learning of the weights α_i to determine the contribution of each kernel ($K(\mathbf{x}, \mathbf{z}) = \sum_{m=1}^{M} \alpha_m K(\mathbf{x}, \mathbf{z}; \theta_m)$). The MKL might be susceptible to the overfitting problem similar to any learning algorithm, specially when the feature space of the testing pattern is quite different from the training ones.

5.3 Multi-stage MCS

In the previous section, we introduced the classification combination to address three parameters of an action recognition system. More specifically, classification combination is introduced for (a) proper combination of multi-scale features, (b) the choice of the length of dictionary in an action representation, and (c) the choice of distance metric for the classification stage. We argue that a MCS which combines the decision of multiple classifiers at each stage can improve the final classification accuracy. Intuitively, a framework which integrates all of the MCSs of different stages should provide a more comprehensive information about different actions, and hence, it should provide higher classification accuracy. This multi-stage MCS framework is a more general framework for MCS which does not require explicit parameter setting or learning.

5.4 Experiments

5.4.1 Methodology

As a multi-resolution classification approach, the MCS gains its higher discrimination power by fully exploiting the multi-resolution characteristics of the motion patterns. In this thesis, we consider a simple but an efficient classifier combination approach. Our MCS framework uses a non-parametric abstract-level combination method in a parallel topology with a SVM or NN classifier as a base classifier (Fig. 5.1). The abstract-level combination method uses a voting scheme such as majority voting and it is less complex than the methods which use a complicated weighting scheme. This way the MCS's computational cost is linear with the number of classifiers. Moreover, due to the correlation of the salient features of different scales, they are not independent input features, and hence, the majority voting is more appropriate than other combination rules such as product rule [56]. The parallel topology opens an avenue towards a

sequential prediction-correction approach for an efficient analysis of human actions. In this approach, the classifier that uses the fine-scale salient features can provide an initial prediction of the class label quickly. As the system now has more time available, it extracts more features at coarser scales and corrects the previous estimation. The non-parametric combination strategy is more general and applicable for our framework as we did not consider any priors about the classifiers. From now own, we consider the majority voting of the parallel classifiers as the baseline and leave the other choices as the future research. Note that the majority voting however requires odd number of classifiers, to ensure there is no deadlock. One can instead use the confidence level of the classifiers to turn around this problem.

We used the KTH dataset [110], the Weizmann classification dataset [11], and the UCF sports dataset [107] for the evaluation of different salient features for the action classification. The salient features of cuboids [25] are detected at nine different spatio-temporal scales as described in Section 3.7.1. Consequently, a MCS combines the decision of nine base classifiers for the final decision. For each scale, a separate dictionary is learnt using the features of just that scale. The length of the dictionary (i.e, the number of clusters of visual words) varies from 100 to 1000 with an interval of 100. For the NN classifier, we used a χ^2, a Euclidean, or an $L1$ (Manhattan) distance metric. The SVM classifier uses a linear, an RBF, a sigmoid, an intersection, or a χ^2 kernel.

5.4.2 Results

In this section, we present the results of action recognition accuracy using a MCS framework to address the question of multi-scale feature combination, the choice of dictionary length, and the choice of distance metric for action matching.

MCS for the combination of multi-scale features

Fig. 5.5 shows the average action classification accuracy on the KTH dataset [110] when multi-scale features are fed into a MCS framework. The results are shown for different dictionary lengths and different kernels. Generally speaking, a SVM classifier with the RBF, χ^2, or the intersection kernel perform slightly better than a SVM classifier with a linear or a sigmoid kernel. The general theme on the performance of the MCS on this dataset is that the larger dictionaries might perform better. More specifically, to obtain an accuracy above the mean, a dictionary with at least 600 visual words is required.

Fig. 5.6 shows the average action classification accuracy on the Weizmann dataset [110] when multi-scale features are fed into a MCS framework. The results are shown for different

dictionary lengths and different distance metrics. Generally speaking, a NN classifier with the χ^2 or the $L1$ distance metric performs better than the once which uses the Euclidean distance. The low performance of the Euclidean distance is due to the fact the action signature which are normalized histogram are not lie in a Euclidean space. The general theme of the NN classifier on this dataset is that the smaller dictionaries might perform better.

One important point from comparison of Fig. 5.5 and Fig. 5.6 is that the observation of one experiment should not be generalized to another experiment. More specifically, while the bigger dictionary with a SVM base classifier provides a better classification on the KTH dataset, a smaller dictionary with an NN classifier performs relatively better on the Weizmann dataset. Several factors explain this observation. (1) The temporal and spatial variations in the sample videos of the KTH dataset are much more than those at the Weizmann dataset. More specifically, the motion patterns of "jogging" and "running" in the KTH dataset are relatively similar. A bigger dictionary can better capture finer motion patterns and hence, it can give a better discrimination power to an action model. Moreover, there exists fast zooming in the sample video sequence for the KTH dataset. The video samples of the Weizmann data set are captured in front of a fixed camera which removes the challenges due to spatial variations. (2) The NN classifier typically performs better with low-dimension feature space. The SVM classifier however works fine for a high-dimension nonlinear feature space. (3) The SVM might not perform well under small training sample sizes. That is the reason why we choose a NN classifier for the Weizmann dataset with at most ten video samples of each action.

Table 5.1 compares the average classification accuracy using different classification approaches on the KTH dataset [110], the Weizmann dataset [11], and the UCF sports dataset [107]. Note that the MCS obtains higher classification accuracy by combining the base classifiers of the features at different scales. The other approaches use a single classifier with scale-invariant features, redundant features, or a multi-resolution action signature (see Section 4.2 for details on single classification approaches).

MCS for the combination of multi-length dictionaries

We reformulate the problem of choosing or learning the proper length of the dictionary (i.e., the right number of clusters of the visual words) into using a set of different lengths of dictionaries and then combine the results. Fig. 5.7 shows the classification accuracy on the KTH dataset when a classifier combination is performed on the results of classifiers with different dictionary lengths. The figure shows this experiment when different kernels have been utilized by the SVM base classifier. Using the classifier combination approach, one can obtain a performance of at least equal to the mean which is independent of the choice of the dictionary length. Higher

Table 5.1: Average classification accuracy on different data sets. Note that the MCS obtains higher classification accuracy than the other approaches which use a single classifier. For the KTH and the UCF sports datasets, the base SVM classifiers use the RBF kernel and for the Weizmann dataset, the base NN classifiers use a χ^2 distance metric.

Classification	Type	KTH	Weizmann	UCF sports
Single classifier	with scale-invariant features (Section 4.2.1)	89.1 %	90.2 %	71.3%
	with redundant features (Section 4.2.2)	89.7 %	91.1 %	73.3%
	with multi-resolution action signature (Section 4.3)	93.4 %	91.6 %	75.6%
MCS	for combination of multi-scale features (Section 5.2.1)	**94.2** %	**96.4** %	**77.3** %

classification accuracy is obtained when the base SVM classifier uses a sigmoid or an RBF kernel, compared to when the SVM uses an intersection kernel. The performance of χ^2 and linear kernels is in between.

Fig. 5.8 shows the classification accuracy of combining MCS classifiers with different dictionary lengths on the Weizmann dataset. Note that the results when the base NN classifiers use a $L1$ distance metric is always greater than or equal to the mean. With a χ^2 distance, the classifier combination approaches 100% accuracy, but it then drops to just above 95%. With a Euclidean distance, the classifier combination accuracy could exceed the mean, but its performance is still lower than those which use a $L1$ or a χ^2 distance.

MCS for the combination of multiple distance metrics

The classification performance is directly affected by the choice of the distance metric it uses to compare different patterns. Fig. 5.9 shows the average classification accuracy on the KTH dataset when the SVM base classifiers of the MCS use five different kernels. As can be seen, there is not a single distance metric (i.e., similarity kernel in the case of SVM classifier) that performs the best at all dictionary lengths and for all classification approaches. In fact, with a given dictionary, one distance metric might perform the best while it performs the worst with another dictionary. For example in this figure, the RBF kernel provides the lowest accuracy using a dictionary with 100 clusters of visual words, while it provides the highest classification accuracy with a 400-

cluster dictionary. It is therefore not trivial to determine which distance metric performs best with a given length dictionary and a given classifier. In the same figure, the accuracy when a MCS combines the results of multiple MCS classifiers with different distance metrics are shown as well. The last bar on the horizontal axis shows the average accuracy over the dictionary length. This bar shows that the combination of the results of different kernels performs better than each individual kernel.

Fig. 5.10 shows the average classification accuracy for the Weizmann dataset when a NN classifier uses different distance metrics including a χ^2 (3.38), a Euclidean (equation 5.12 with $p = 2$), or an L_1 (Manhattan) metric (equation 5.12 with $p = 1$). As the last bar shows, for this dataset, the χ^2 kernel performs slightly better than the combination scheme. The low performance of the Euclidean distance is in support of the fact that this distance is not a proper metric for the histogram comparison (see Section 5.1.2).

$$D_p(\mathbf{x}, \mathbf{z}) = |\mathbf{x} - \mathbf{z}|^p = \sum_i |x_i - z_i|^p \tag{5.12}$$

Multi-stage MCS

Table 5.2 shows the average classification accuracy on different data sets using multi-stage MCS. For the KTH and the UCF sports data sets, with five different kernels and ten different dictionaries, the multi-stage MCS combines fifty MCS classifiers which use the SVM base classifier. For the Weizmann dataset, with three different kernels and ten different dictionaries, the multi-stage MCS combines thirty MCS classifiers which use the NN base classifier. Note that the multi-stage MCS obtains higher classification accuracy than the average of the MCSs for all three data sets.

Table 5.2: Average classification accuracy on different data sets using multi-stage MCS. Note that the multi-stage MCS obtains higher classification accuracy than the average of the MCSs.

Data set	Base classifier	Scheme	Classification accuracy (%)
KTH	SVM	**Multi-stage MCS**	**95.05** %
		Mean of MCSs	93.94 %
Weizmann	NN	**Multi-stage MCS**	**95.35** %
		Mean of MCSs	94.11 %
UCF sports	SVM	**Multi-stage MCS**	**78.6** %
		Mean of MCSs	77.6%

To show an example of the efficiency of a multi-stage MCS, Fig. 5.11 shows the action classification accuracy using this approach when just two kernels are combined. The combination is performed for two kernels of linear and RBF for the base SVM classifier. In this figure, the horizontal axis shows the number of MCSs that have been combined. Up to ten in the horizontal axis shows the results of combining the MCSs which use a linear SVM base classifier with ten different dictionaries. The next ten show the classification accuracy when the MCSs with the RBF SVM base classifier are also included in the combination. Note that when all the MCSs are combined, the classification accuracy reached to the maximum accuracy that all the MCSs could provide. This is a fundamental improvement in terms of independency of the multi-stage MCS from the choice of distance or the dictionary length. Moreover, the multi-stage MCS could reach to the accuracy of the best classifier even though the feature sets fed into the classifiers are not independent.

Fig. 5.12 shows the confusion matrix of different MCSs and their combination on the KTH dataset. Two sets of MCSs with ten different dictionaries use linear and RBF kernels in their SVM base classifier. Note that when the multi-stage MCS is performed to combine the decision of all twenty MCSs (Fig. 5.12(c)), the final decision is better than each set of MCSs (Fig. 5.12(a), 5.12(b)). More specifically, this combination helps better discrimination of actions such as ”jogging” which is quite similar to the ”running” in the KTH dataset.

5.5 Conclusion

In this section, we proposed a MCS approach for multi-resolution action classification. In this approach, the features of each individual spatio-temporal scales are treated separately and consequently fed into a separate base classifier. A simple but efficient classifier combination based on majority voting provided us with a higher classification accuracy than the existing approaches that use a single classifier. In fact, the collective behavior of the features are better captured using a MCS framework. We further extended the idea of classifiers combination to address the problems of the dictionary's length and the choice of distance metric. In fact, instead of selection or learning, we performed separate classification for each individual dictionaries with different lengths. We then combined the results for the final decision. A similar approach has been used for the combination of classifiers which use different distance metrics. This approach provided us with a classification accuracy higher or equal to the mean of all classifiers. We also proposed the multi-stage MCS framework which unifies different MCSs at different stages of human action recognition framework for further improvement of the classification accuracy. The multi-stage MCS is a parameter-free multi-resolution action recognition framework.

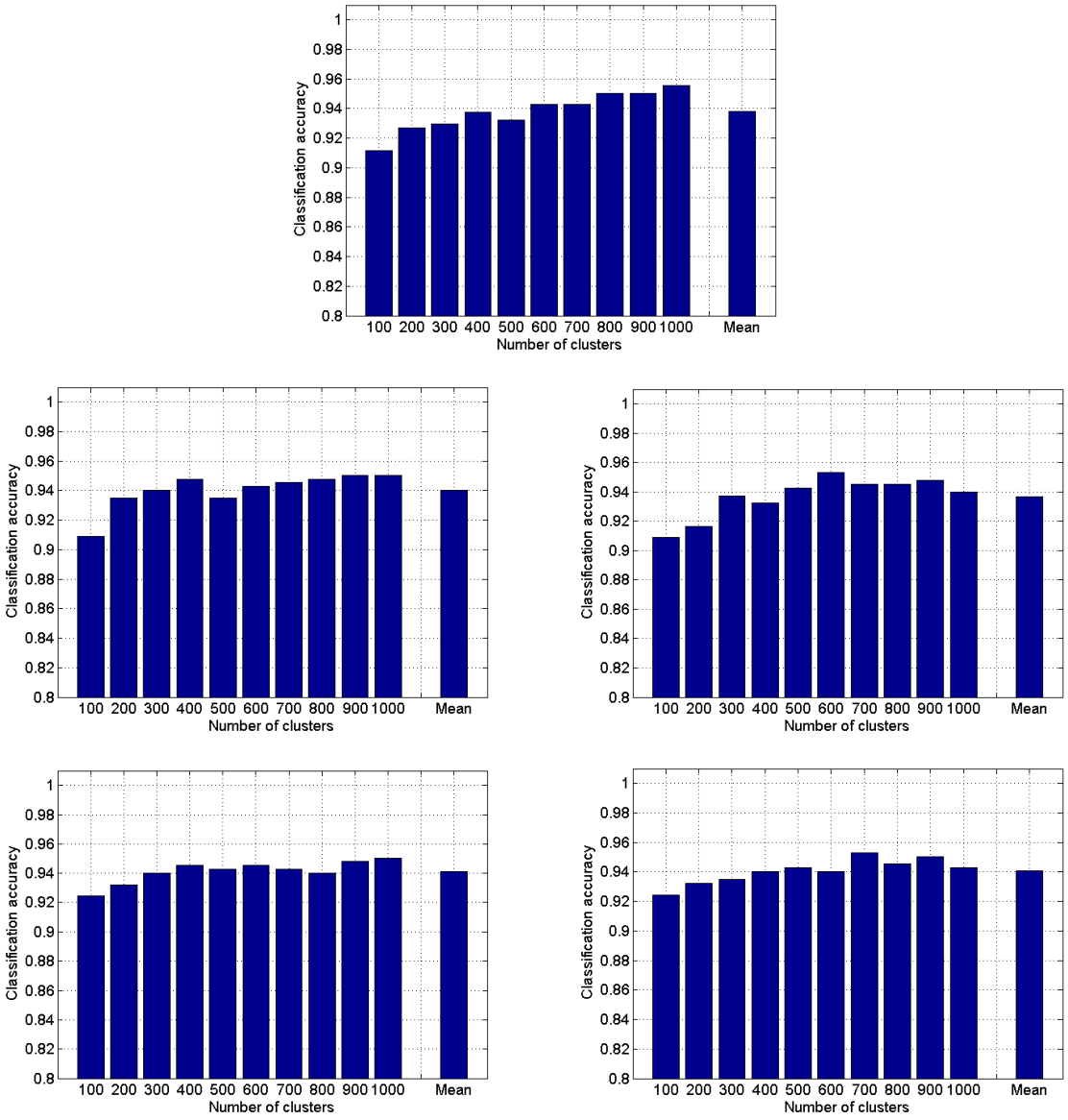


Figure 5.5: MCS on the KTH dataset using SVM base classifiers with different kernels (from left to right, top to bottom: linear, RBF, sigmoid, intersection, and chi-squared). The dashed line shows the mean over different dictionary lengths. The main observation here is that the choice of kernel and dictionary length is not trivial. See Table 5.1 for the comparison of a MCS with other classification approaches.

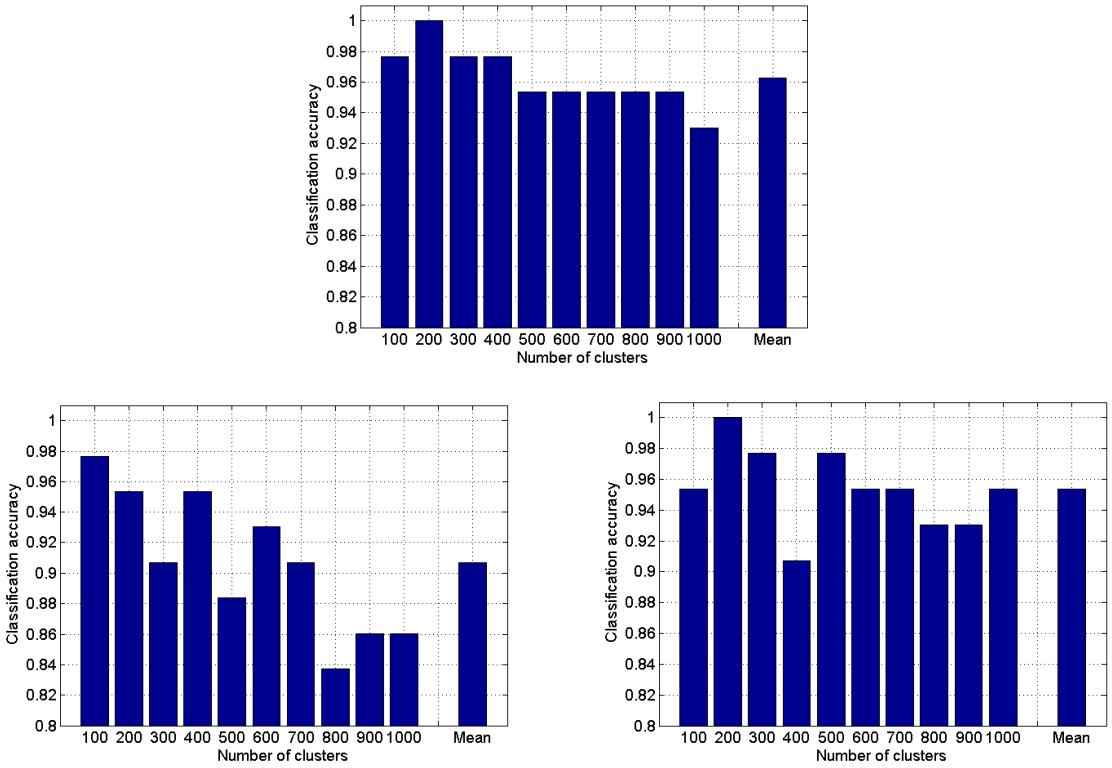
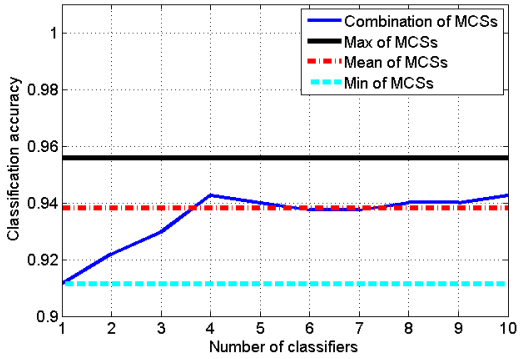
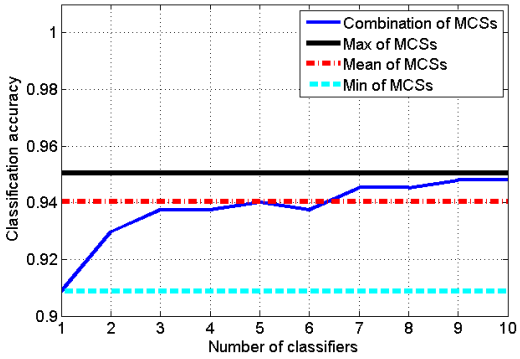


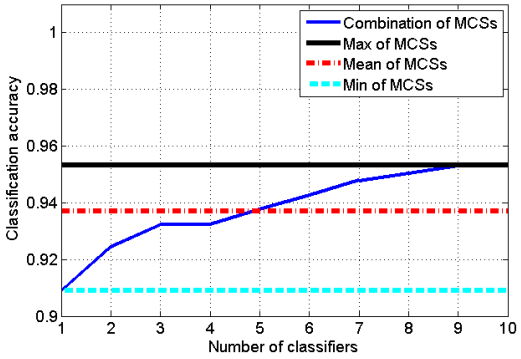
Figure 5.6: MCS on the Weizmann dataset using NN base classifiers with different distance metrics (χ^2, Euclidean, and L1). The dashed line shows the mean over different dictionary lengths. The main observation here is that the choices of distance metric and the dictionary length are not trivial. See Table 5.1 for the comparison of a MCS with the other classification schemes.



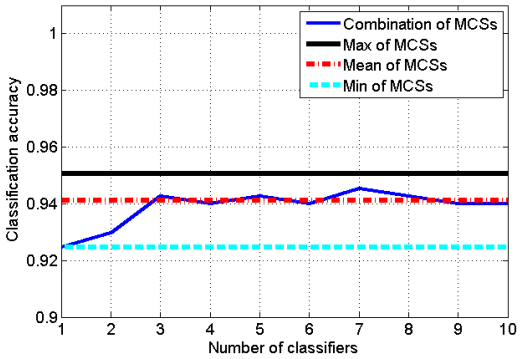
(a) Linear kernel



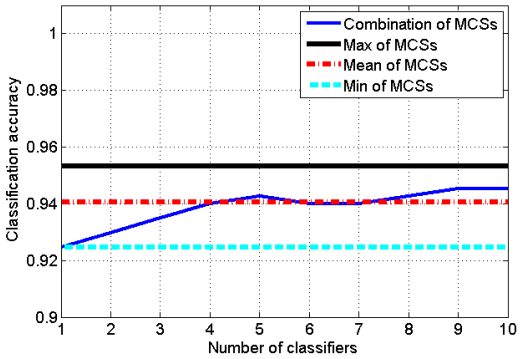
(b) RBF kernel



(c) Sigmoid kernel

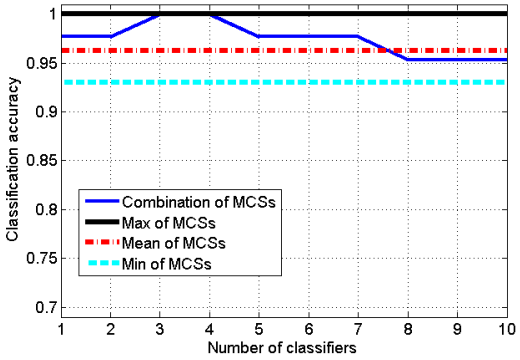


(d) Intersection kernel

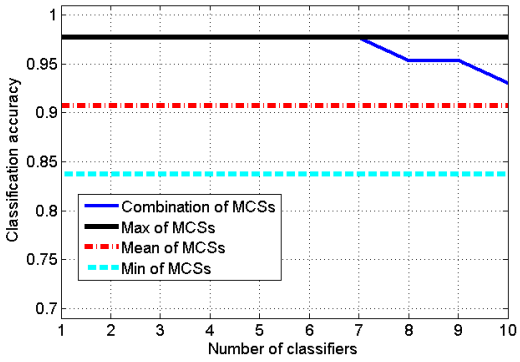


(e) χ^2 kernel

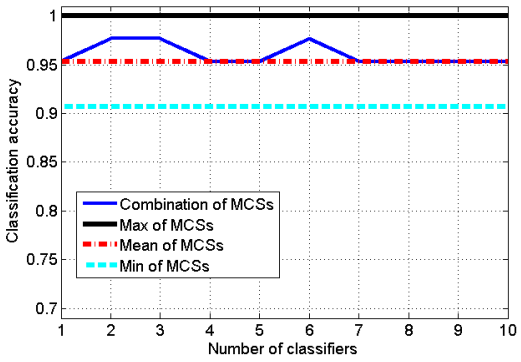
Figure 5.7: Combination of MCS classifiers with multi-length dictionaries on the KTH dataset [110]. Note that a sigmoid and then an RBF kernel (second row) has a better discrimination power than the other kernels.



(a) χ^2 distance



(b) Euclidean distance



(c) L1 distance

Figure 5.8: Combination of MCS classifiers with multi-length dictionaries on the Weizmann dataset [11]. Note that a L1 kernel have a better discrimination power than the other kernels as it keeps the final results greater or equal to the mean. Note that the mean of χ^2 and $L1$ is higher than that of Euclidean distance.

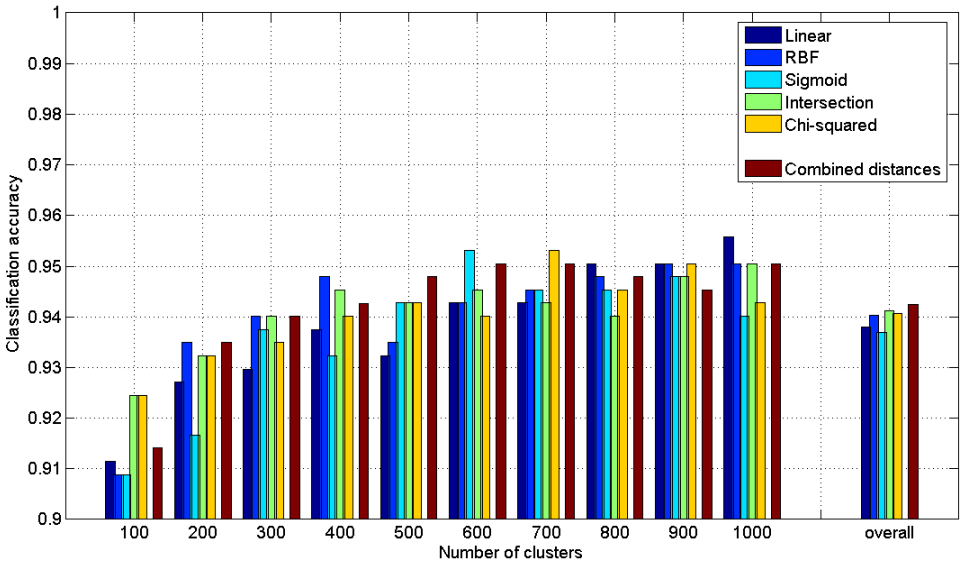


Figure 5.9: Average classification accuracy using MCSs with different kernels on the KTH dataset. The last bar shows the mean over the clusters. Note that the combination of kernels provides a better overall performance.

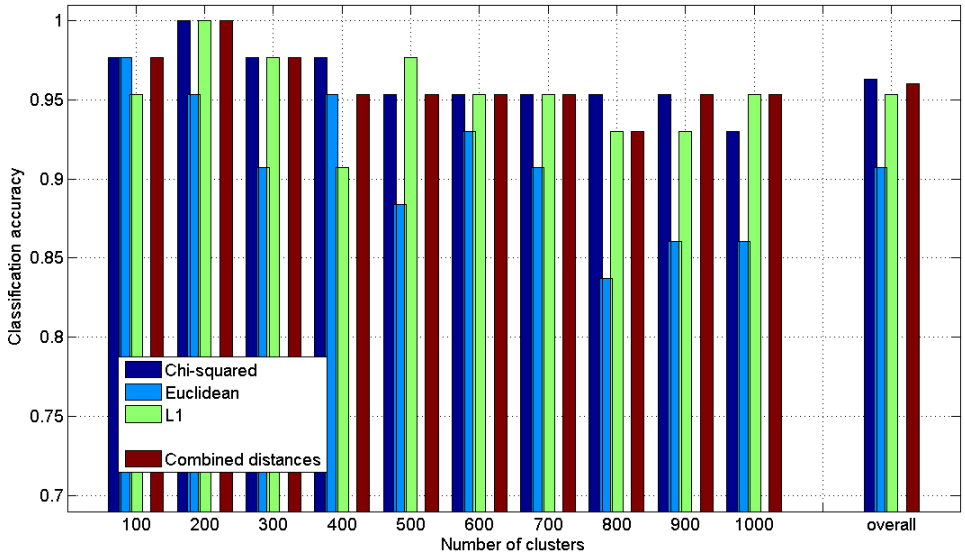


Figure 5.10: Average classification accuracy using MCSs with different distance metrics on the Weizmann dataset. Here, the base NN classifier uses a χ^2, a Euclidean, or an L_1 distance metric. The last bar shows the mean over the clusters. The combination of distances provides a higher overall classification accuracy than the Euclidean and L_1 distances, but slightly lower than the χ^2. Note to the low performance of the Euclidean distance.

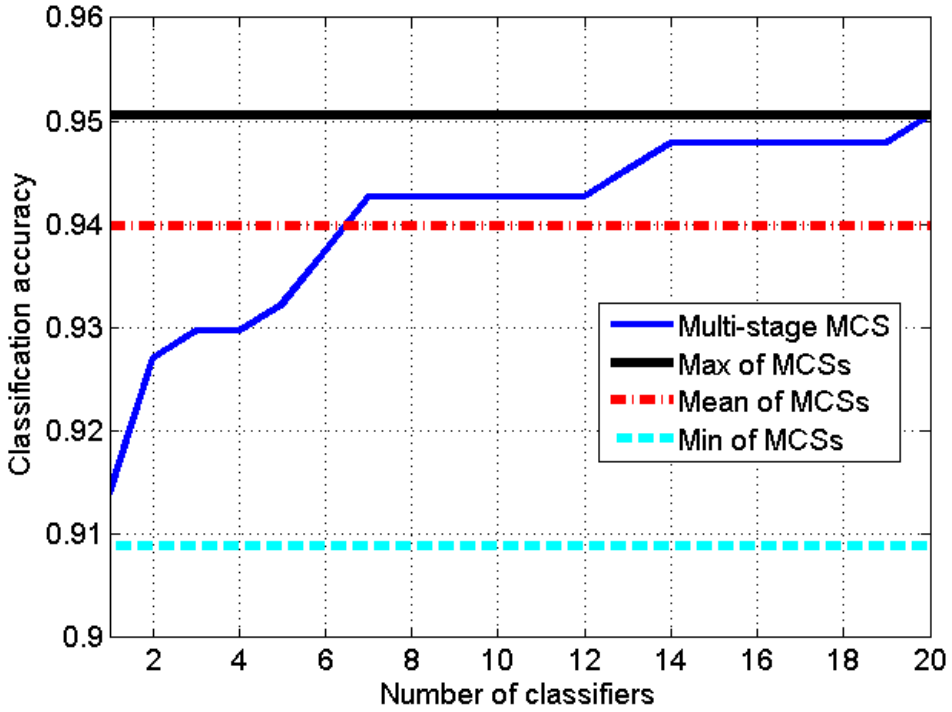
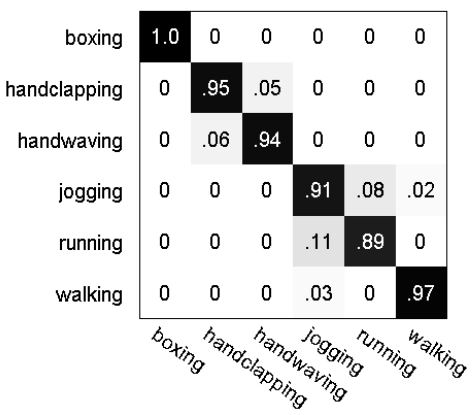
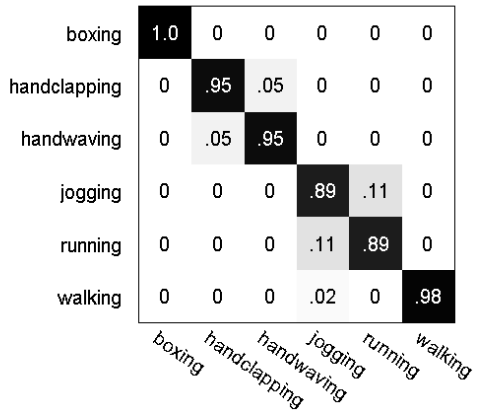


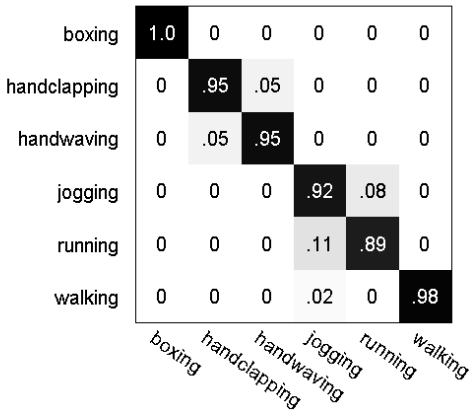
Figure 5.11: Multi-stage MCS on the KTH dataset. The results of two sets of MCSs are combined using a multi-stage MCS. The first ten MCSs use ten different dictionary lengths with a linear kernel for the SVM base classifier. The next ten MCSs use the RBF kernel. The sold blue line shows the increase of accuracy when all the MCSs are combined using a multi-stage MCS. Note that the multi-stage MCS could reach to the performance of the best MCS.



(a) Linear kernel



(b) RBF kernel



(c) Combined linear and RBF kernels

Figure 5.12: Confusion matrices of using different MCSs and a multi-stage MCS on the KTH dataset. The base SVM classifiers use a linear kernel (a) or an RBF kernel (b). The MCSs of different dictionaries are combined. In (c), using a multi-stage MCS, the MCSs with different dictionaries and two kernels of linear and RBF kernel are combined. Note that the multi-stage MCS performs better and it can better recognize the ”jogging” which is similar to ”running”.

Chapter 6

Multi-stage MCS with new features

This chapter integrates the approaches which are introduced in the previous chapters for different components of an action recognition system. It is of interest to see the performance of an action recognition system which uses a set of robust salient features, a proper action representation, and an appropriate classification approach. This requires a comprehensive comparison of different approaches introduced in the previous chapters. In Chapter 3, we introduced a new temporal multi-resolution filtering approach for robust local salient feature detection in video. In Chapter 4, we introduced the concept of multi-resolution analysis of human motion patterns using a multi-resolution action signature. In Chapter 5, we investigated the design of a new multi-stage MCS as a parameter-free multi-resolution framework. In this chapter, we briefly highlight the findings of the previous chapters and then provide the experimental results of using a multi-stage MCS framework for the human action recognition using multi-scale features extracted from both symmetric and asymmetric video filtering.

6.1 New salient features

Chapter 3 introduced the asymmetric temporal filtering for a multi-resolution representation of a video signal for the salient feature detection. More specifically, four asymmetric temporal filters of Poisson, asymmetric sinc, scale-derivative Gaussian, and log Gaussian are compared with the symmetric Gabor filter. Using a common local feature detection framework, we showed that an asymmetric temporal filtering provides salient features which are more precise and more robust compared to those obtained from the symmetric Gabor filtering. Moreover, the features detected using the asymmetric filters provided us higher classification accuracy when they are fed into a

standard bottom-up discriminative recognition framework. In fact, our novel asymmetric sinc filtering performed the best. Poisson and scale-derivative Gaussian performed better than the log Gaussian. All of these asymmetric filters perform better than the symmetric Gabor filter.

6.2 Multi-stage MCS

In Chapter 4, we argued that a multi-resolution motion analysis is required to better represent an action. A multi-resolution action signature provides us a higher classification accuracy using a standard single classification approach. Due to dimensionality problem of the multi-resolution signatures, we introduced the MCS framework (Section 5.2.1) which improves the classification accuracy. In Section 5.3, the MCS has been further extended to multi-stage MCS to address the proper approach for the combination of the multi-scale features, the length of the dictionary by which an action is described, and the choice distance metric for comparing the action signatures. The multi-stage MCS is a parameter-free framework which improves the overall classification accuracy compared to those methods which use a single classifier.

6.3 Experiments

6.3.1 Methodology

We use a multi-stage MCS framework (Section 5.3) to compare the performance of different salient features. The salient features are detected after a multi-resolution temporal filtering of the video using a symmetric Gabor, an asymmetric sinc, a Poisson, a scale-derivative Gaussian, or a log Gaussian filter. The spatial filter is a 2D Gaussian. The video signal is filtered at three spatial and three temporal scales. The salient features are therefore detected at nine spatio-temporal scales similar to the setting in Section 3.7.1. Similar to the configuration setting in Section 4.4.1, a dictionary of visual words for each spatio-temporal scale is computed. The number of clusters of the visual words varies from 100 to 1000 with interval of 100. In the design of the multi-stage MCS, the base SVM classifiers are used for the KTH and UCF sports datasets and the NN classifiers are used for the Weizmann dataset. The SVM might use five different similarity kernels including a linear, the RBF, a sigmoid, an intersection, or a χ^2 kernel (equations 5.7-5.11). The NN classifiers might use three different distance metrics of χ^2, Euclidean, or $L1$.

6.3.2 Results

Table 6.1 shows the average classification accuracy using a multi-stage MCS framework with features detected using five different temporal filters of the Gabor, log Gaussian, sclae-derivative Gaussian, Poisson, and an asymmetric sinc filter. The results are shown for the Weizmann, the KTH, and the UCF sports datasets.

From Section 3.7.4, we have seen that the features detected using asymmetric temporal filtering of the video provide higher classification accuracy when they are fed into a single classifier. When the features are fed into a multi-stage MCS framework, we have different observation. From Table 6.1, we see that the asymmetric sinc performs the best and the performance of Gabor exceeds the other asymmetric filters. One possible explanation is that the base classifiers which use the features of an asymmetric filter such as scale-derivative Gaussian on the KTH dataset are not as complementary as those of Gabor. This might be due to high correlation among the features of different scales or the fact that dictionaries with different sizes might not that much complementary to describe different characteristics of an action. Consequently, the gain in using multiple classifiers using asymmetric features of log Gaussian or the scale-derivative Gaussian is not as the asymmetric sinc or the Gabor. As we stated in section 5.2, the MCS performs optimum when the features of the base classifiers are independent and complementary. However, one should note that the multi-stage MCS outperforms single classifier for all five types of features. Using a multi-stage MCS framework, in summary, the asymmetric sinc performs the best in all three datasets. On KTH dataset, the Gabor is the second best and on the Weizmann the Gabor and the scale-derivative Gaussian are the second best. On the UCF sports dataset, the log Gaussian is the second best and the performance of the Gabor is the least.

Table 6.1: Average classification accuracy on the Weizmann, KTH, and UCF sports data sets using multi-stage MCS. The accuracies are shown for features detected using five different temporal filters. Note that the asymmetric sinc has the best performance in all three datasets.

Classification	Type	KTH	Weizmann	UCF sports
Multi-stage MCS	Gabor	95.05 %	95.35 %	78.7 %
	Log Gaussian	93.75 %	93.02 %	90.7 %
	Scale-deri. Gaussian	92.8 %	95.35 %	82 %
	Poisson	93.5 %	93.02 %	85.3%
	Asymmetric sinc	**95.7** %	**98** %	**96.7** %

Fig. 6.1, Fig. 6.2, and Fig. 6.3 show the confusion matrices using multi-stage MCS classification framework on the Weizmann classification dataset [11], on the KTH dataset [110], and

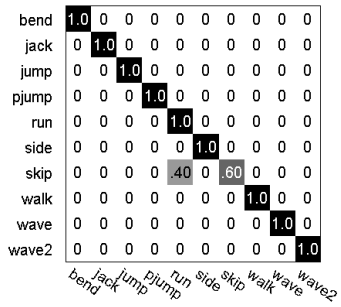
on the UCF sports dataset [107], respectively. The scale-space representation for the salient feature detection is obtained using spatial Gaussian filter and different temporal filters of a Gabor, a log Gaussian, a scale-derivative Gaussian, a Poisson, and an asymmetric sinc. In the Weizmann dataset, the motion pattern of the "skip" and "run" is similar. In the KTH dataset, the motion pattern of "running" and "jogging" are quite similar. In the UCF sports dataset, the most confusion is between motion pattern of "run" and "riding". Generally speaking, all the filters have difficulty discriminating these similar motion patterns. However, in all three dataset, the asymmetric sinc can better discriminate these similar actions and provides the highest classification accuracy.

6.3.3 Comparison with existing methods

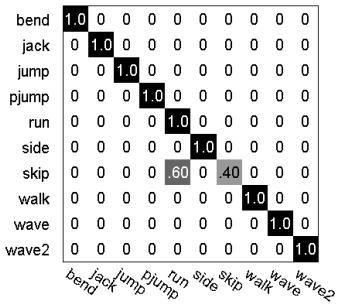
This thesis introduced a new approach to detect robust salient features and a new approach for multi-resolution action classification using the multi-stage MCS. In a common framework for detection and classification, these contributions have been fairly compared with the existing methods in in Table 3.2 and Table 4.1, respectively. However, a direct comparison of the final classification results with other methods in the action recognition literature is not trivial. This is because different publications consider different experimental settings such as the choice of dictionary length, the similarity kernel, or the classification scheme even though they used the same data sets.

Here, we provide Table 6.2 just for showing where our contributions stand in comparison with the state-of-the-art methods for human action recognition on the KTH [110] and Weizmann [11] data sets. The intention is not to show all possible publications in this field. Different methods in Table 6.2 might use (a) different feature detectors or descriptors, (b) the video segmentation, and (c) different classification frameworks. For example, the hidden conditional random fields method utilized by Wang and Mori [124] is much more complex than the simple BOW modeling and it performs background subtraction to compute motion fields. Similarly, the local SVM approach by Weinland et al. [127] requires video segmentation and computation of the 3D HOG at densely distributed locations within the foreground volume. It is not therefore fair to directly compare these methods with our method which uses a simple global BOW representation of sparse local motion features and does not require video segmentation.

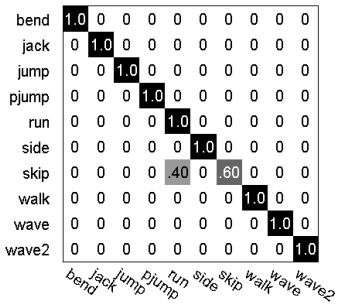
Our intention is not to advocate the report of the best result for a comparison. For example, as Fig. 5.10 shows, we also obtained perfect classification of the actions in the Weizmann dataset using the baseline cuboids [25] (features detected by the temporal Gabor filtering) when a MCS with 200 visual words in the dictionary and a χ^2 or a $L1$, or our combined distance MCS approach



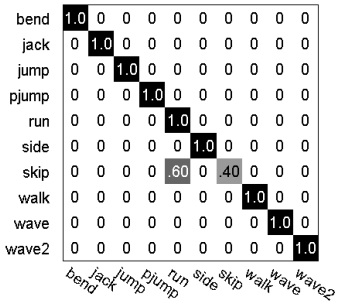
(a) Using Gabor kernel



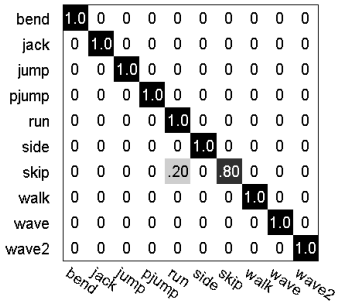
(b) Using log Gaussian kernel



(c) Using scale-derivative kernel



(d) Using Poisson kernel



(e) Using asymmetric sinc kernel

Figure 6.1: Confusion matrices on the Weizmann classification dataset [11] using multi-stage MCS classification framework. The scale-space representation for the salient feature detection is obtained using spatial Gaussian filter and different temporal filters.Note that all the filters have difficulty discriminating ”skip” from ”run”. This is due to the similarity of motion patterns of these fast motions in this dataset. The asymmetric sinc performs the best.

is used. We also obtained perfect classification on this dataset using the features of an asymmetric sinc with a MCS framework with 100 visual words in the dictionary and a χ^2 distance metric.

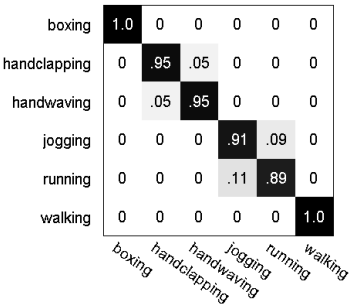
Without video segmentation and with a similar BOW setting, we obtained 93.3% accuracy on the KTH dataset which is comparable with 84.3% using 3D Hessian features [129], 88.3% accuracy using dense features [106], and 92.1% accuracy using 3D Harris features [122]. Moreover, our 96.7% accuracy on the UCF sports dataset is much better than the accuracy obtained using the dense sampling [122] with 85.6%, the unsupervised feature learning [72] with 86.5%, and the dense trajectory [44] with 88.2%. Considering the sparsity of asymmetric sinc filtering (Fig. 3.18), our results are in contrast to the previous observation about better performance of dense sampling in realistic scenario [122], showing that the quality of salient features is more important than the number of salient features. Even though the salient features from background might provide some contextual information, they might be harmful in discrimination of different activities as well. In fact, our results on the KTH and UCF sports shows the importance of sparse salient feature detection and effectiveness of asymmetric temporal filtering for robust and good quality salient feature detection.

6.4 Conclusion

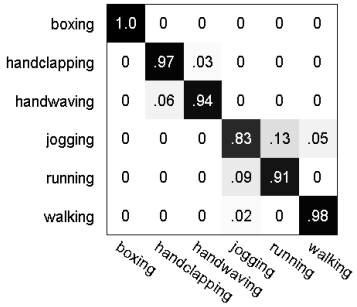
In this section, we compared the performance of different salient features in a new classification framework of multi-stage MCS. The main observation is that for all type of features, the multi-stage MCS provides higher classification accuracy than the existing classifiers (See Table 3.2 and Table 4.1). We also show that the multi-stage MCS provides comparable results with the existing state-of-the-art methods in the literature. The simplicity and being parameter-free are the main advantages of using multi-stage MCS compared to more complex recognition frameworks.

Table 6.2: Comparison of different state-of-the-art methods for human action recognition on the KTH [110], Weizmann [11], and the UCF sports [107] data sets. Note that not all of the methods listed here use a similar experimental setting. For the multi-stage MCS, we report the results when the features detected from the scale-space representation of the video using an asymmetric sinc filtering. We could obtain the best performance of 95.7% on the KTH dataset and 96.7% on the UCF sports dataset. On the Weizmann dataset, the asymmetric since obtains perfect classification using a MCS with 100 visual words in the dictionary and a χ^2 distance metric. The average performance using multi-stage MCS is however 98%.

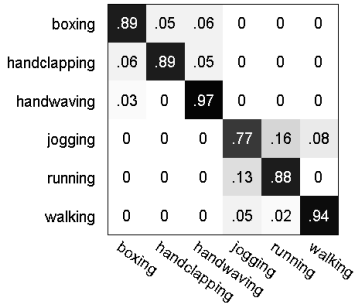
Method	*KTH actions*	*Weizmann actions*	*UCF sports*
Multi-stage MCS	**95.7** %	98 %	**96.7** %
Wang et al. [44]	94.2 %	-	88.2 %
Le et al. [72]	93.9 %	-	86.5 %
Wang et al. [122]	92.1 %	-	85.6 %
Wang and Mori [124]	92.5 %	**100** %	-
Weinland et al. [127]	92.4 %	**100** %	-
Zhang et al. [136]	91.33 %	92.89 %	-
Klaser et al. [57]	91.4 %	84.3 %	-
Liu and Shah [47]	94.16 %	-	-
Zhao and Elgammal [137]	91.7 %	-	-
Rapantzikos et al. [106]	88.3 %	-	-
Willems et al. [129]	84.3 %	-	-
Dollar et al. [25]	81.2 %	-	-
Schuldt et al. [110]	71.7 %	-	-
Ali et al. [5]	-	92.6 %	-
Goodhart et al. [40]	-	83.7 %	-
Scovanner et al. [111]	-	82.6 %	-



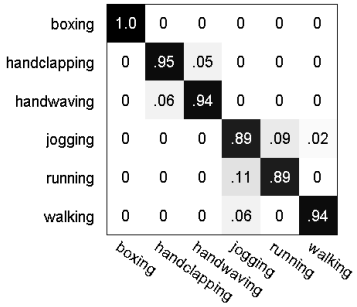
(a) Using Gabor kernel



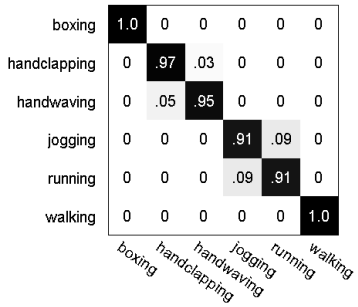
(b) Using log Gaussian kernel



(c) Using scale-derivative kernel

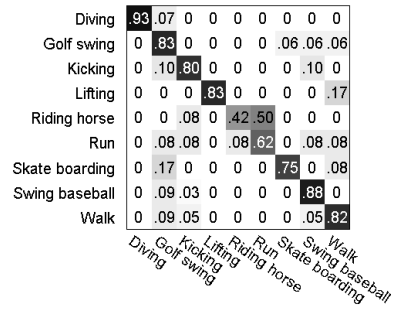


(d) Using Poisson kernel

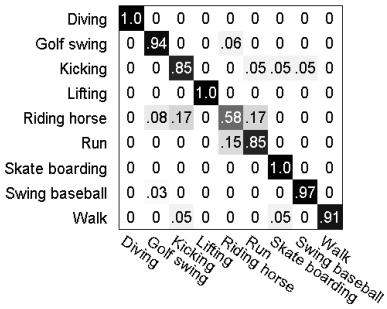


(e) Using asymmetric sinc kernel

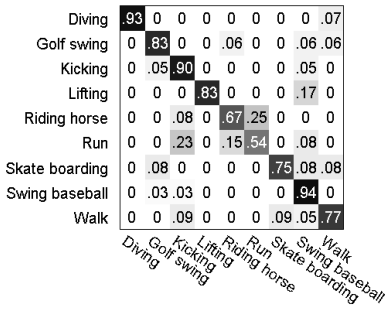
Figure 6.2: Confusion matrices on the KTH dataset [110] using multi-stage MCS classification framework. The scale-space representation for the salient feature detection is obtained using spatial Gaussian filter and different temporal filters. Generally speaking, all the filters have difficulty discriminating "jogging" from "running". This is due to the similarity of motion patterns of these fast motions in this dataset. The asymmetric sinc has the best overall performance and it also could better discriminate these actions.



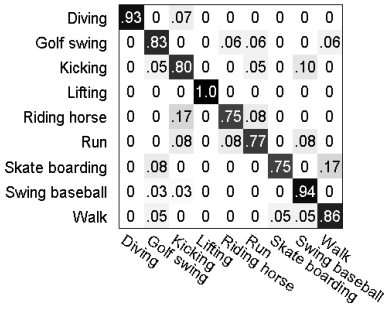
(a) Using Gabor kernel



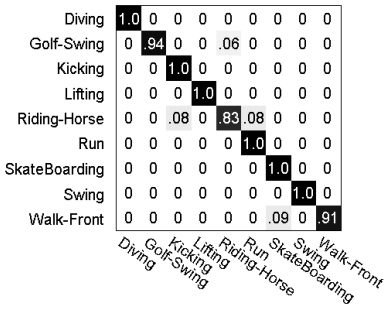
(b) Using log Gaussian kernel



(c) Using scale-derivative kernel



(d) Using Poisson kernel



(e) Using asymmetric sinc kernel

Figure 6.3: Confusion matrices on the UCF dataset [107] using multi-stage MCS classification framework. The scale-space representation for the salient feature detection is obtained using spatial Gaussian filter and different temporal filters. The asymmetric sinc has the best overall performance and it can better discriminate different actions.

Chapter 7

Conclusion and future research

7.1 Conclusion

This thesis introduced three contributions to improve the human action recognition in video using a discriminative bottom-up method. First, an asymmetric temporal filtering (Chapter 3) was introduced for a robust salient motion feature detection. Second, a multi-resolution action representation using a multi-resolution signature was introduced in Section 4. Finally, a multi-stage multiple classifier systems (MCS) (Chapter 5) was introduced as a parameter-free classification framework for a multi-resolution analysis of human actions.

For robust local salient feature detection, three asymmetric temporal filters of log Gaussian, scale-derivative Gaussian, and Poisson from the scale space theory and a brand new filter of asymmetric sinc were compared with the widely-used temporal symmetric Gabor filter. By examining the precision score and the reproducibility scores under geometric deformations such as view change or affine transformations, the features which were obtained using an asymmetric temporal filtering had shown better performance compared with the features obtained from the symmetric temporal Gabor filtering. Moreover, the asymmetric features were more informative than the symmetric features as they provided higher classification accuracy with a standard bag-of-words representation and a single discriminative classifier. In all experiments, except spatial scale changes, the asymmetric filters performed better than the symmetric temporal Gabor filtering. Among the asymmetric filters, our novel asymmetric sinc performed the best for precision, reproducibility, and classification accuracy.

For a proper discrimination of human actions, an efficient MCS framework with a parallel topology has been introduced. This framework better encodes the multi-resolution characteristics

of the motion patterns in different human actions. Using non-parametric classifiers and a simple majority voting for the classifiers combination, the MCS framework has been generalized to provide a better differentiation of different human actions. This general classifier combination approach has been then extended to address two other problems, namely the choice of a distance metric and the number of visual words in the dictionary/codebook. This way MCS had been applied at different stages of the recognition process, from low-level feature combination, to mid-level action representation, and high-level classification. This multi-stage MCS provides improved classification accuracy, and it is free of experimental selection/tweak for the choice of distance metric or dictionary length.

In Section 6.3.2, we compared the performance of the multi-stage MCS using features detected from different asymmetric/symmetic temporal filters. The multi-stage MCS outperforms the single classifier no matter which feature is being used. However, the performance with features obtained from an asymmetric sinc filtering is the highest on all three datasets of the Weizmann, the KTH, and the UCF sports dataset. Based on the experimental results on the feature detection and action classification, the multi-stage MCS framework which uses the multi-scale features obtained from the asymmetric sinc filtering is recommended for the task of human action recognition in video.

We introduced the concept of time-causal and asymmetric temporal filtering for robust local salient feature detection in video. One should note that the spatial and temporal causality of the local events in video is also important for proper modeling of the human motions. In the next section, we highlight a new approach to address this problem and other future directions that one can take to extend this thesis.

7.2 Future directions of this research

Both introduced concepts of asymmetric temporal filtering and the multi-stage MCS classification require more theoretical and experimental analysis. Here, some of the future research directions to extend the approaches introduced in this thesis are explained briefly.

7.2.1 Action modeling using a dynamic bag-of-words

In Section 4.1, we explained that the global BOW representation of an action is an order-less representation and does not take into account any spatial or temporal relationship among the features. The structural constraints and the order of movements of different body limbs are

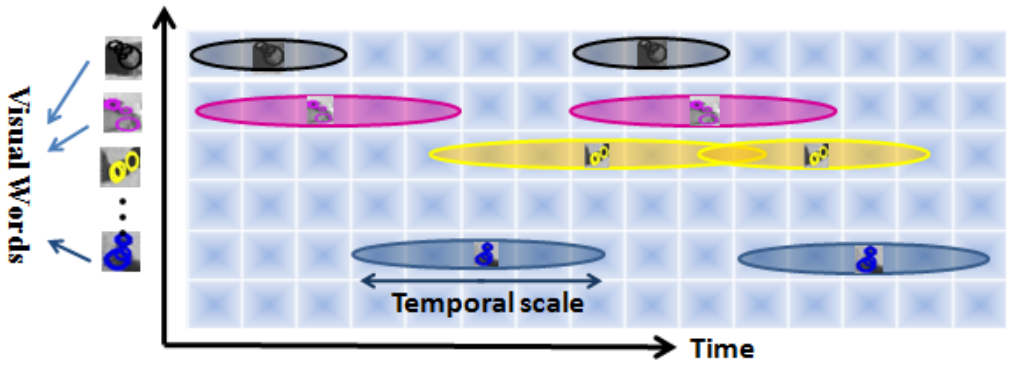


Figure 7.1: The observation matrix Y shows the probability of appearance of the visual words at each time instant in a video. The probability is maximum at the time of feature's occurrence and will degrade according to the temporal scale of the visual words.

important information in modeling the human motions. Intuitively, incorporation of this information in the action modeling should provide a better discriminative action representation and classification. There are some methods such as HCRF [124] which model the spatial structures and the structured pLSA [135] which incorporates the temporal relationships in the action modeling. In general, the dynamic Bayesian networks are the graphical models capable of modeling both spatial and temporal structures in time-series data [64, 93, 131].

Here, we introduce a novel system-theoretic approach for this modeling. In this approach, each action is modeled as a linear/non-linear dynamic system observed through the dynamic BOW representation (Fig. 7.1). Using a dynamic system identification method, one can then differentiate different actions from each other. The input to the system is the sequence of visual words which are extracted using the bottom-up feature extraction approach. The system theoretic modeling of the actions is a top-down approach. Therefore, this framework is a unification of the bottom-up and top-down approaches for human action recognition.

In a system theoretic formulation of the human actions, the temporal evolution of the visual words $Y = \{y_1, y_2, ..., y_t, ..\}$ in an action is modeled using a linear dynamic system. The system is governed by a set of (hidden) states $x \in \Re^n$ and is observed through the visual words $y \in \Re^m$. A linear dynamic system (LDS) is then formulated as follows.

$$x_{t+1} = A\,x_t + \omega_t, \quad \omega \sim \aleph(0, Q) \tag{7.1}$$

$$y_t = C\,x_t + \nu_t, \quad \nu \sim \aleph(0, R) \tag{7.2}$$

in which, the inaccuracy of the linear modeling of the relations between the observation y_t and the hidden states x_t is modeled by the observation noise ν_t. The system noise ω_t models the uncertainty of the linear dynamic modeling of the hidden states. These noise models are assumed to be from normal distributions with different standard deviations. The parameters of the system are the dynamic model $A \in \Re^{n \times n}$, the observation model $C \in \Re^{m \times n}$, and the noises covariance matrices R and Q. The state of the system is initialized at x_0. In a compact form, we can represent an LDS system with $L = (A, C, Q, R, x_0)$. Note that the standard LDS is related to probabilistic graphical models [93].

Having learnt different LDS systems for different actions, the objective of the classification of an unknown action is to find the best LDS match to the LDS of the unknown action. In order to determine the difference between two LDS systems, different distance metrics can be used. Among those, three main distance metrics are: 1) algebraic metrics such as the Binet-Cauchy kernel [121], 2) geometric distances based on subspace angles between the observability subspaces of the LDSs [20] , and 3) the information theoretic metrics like the KullbackLeibler divergence [14].

The Binet-Cauchy kernels, for example, can compare two dynamic systems [121]. From the family of these kernels, the trace kernel K_{Trace} compares the distance between two zero-mean LDS systems $L = (A, C, Q, R, x_0)$ and $L' = (A', C', Q', R', x'_0)$. If two LDS systems have the same underlying and independent noise processes, the closed form of the trace kernel can be defined as [121]:

$$K_{Trace}(L, L') = x_0^T P x_0 + \frac{\lambda}{\lambda - 1} \, trace(QP + R) \tag{7.3}$$

in which $P = \sum_t \lambda^t A^T(t) C^T C' A'(t)$. If $\lambda |A||A'| << 1$, then P is the solution of the Sylvester equation $P = \lambda A^T P A' + C^T C'$. Note that the trace kernel $K_{Trace}(L, L')$ depends on the initial states of the two LDS systems. In action recognition task, however, an initial state independent metric is required as it does not matter when an action starts. An initial state-independent alternative of the Binet-Cauchy kernel such as the maximum singular value kernel ($K_\sigma(L, L') = \sigma_{max}(P)$), or the determinant kernel ($K_d(L, L') = det(P)$), or the the initial state-independent trace kernel ($K_t(L, L') = Trace(P)$) should thus be used [121].

A non-linear dynamic system (NLDS) should better model human actions as the underlying motions of the body limbs are non-rigid. One might also consider a nonlinear relationship between the measurement model and the observation (i.e., $\Phi(y_t) = C \; x_t$). One can thus use the kernel PCA [17] for the decomposition of the variation of the observation matrix and a non-linear form of the similarity kernel for system matching.

7.2.2 Nonlinear scale-space filtering

In Section 3.3, consistent with the existing literature, we used a spatial Gaussian filtering for spatio-temporal salient feature detection. For the time domain, we introduced four different causal and asymmetric temporal filters in addition to the widely used the symmetric Gabor filter [25]. We then showed that the scale-space representation of the video signal which is filtered by an asymmetric filtering is more reliable for the salient feature detection than that of a symmetric temporal filtering. However, all of these spatial and temporal filters are linear, meaning that the conceptually meaningful structures such as edges [126] and salient motions [113] might be blurred and delocalized. The delocalization of these important structures might result in the false positive detection of salient features no matter which saliency criteria is used. To address this problem, one need to develop a non-linear (scale-space) filtering approach which prevents both spatial and temporal dislocations.

For a non-linear scale-space filtering, adaptive non-linear smoothing functions should be used so that they can control the smoothing and hence prevent blurring and delocalization of the important structures. More specifically, instead of a constant diffusion in equation 3.24, the edge-stopping g_s and motion-stopping g_t functions (7.4) can control the energy redistribution in the scale-space filtering. This way the smoothing at high spatial gradients ∇u_s, corresponding to the edges, and at high temporal gradients ∇u_t, corresponding to the salient motions, is minimized. In contrast, the smoothing is maximized at homogeneous regions where the gradient values are lower than a threshold λ. The non-linear scale-space filtering has been widely used in the image processing literature for image restoration, enhancement and feature extraction [126]. Among different choices, the non-linear stopping functions (7.4) in anisotropic diffusion literature [100] or the Tukey error norm [10] in robust estimation are the most widely used ones. One can therefore use these non-linear functions in the spatio-temporal diffusion equation (3.24) by replacing them with the constant conduction terms. Note that this PDE is time-causal meaning that the temporal diffusion is unilateral to just the past frames. It is hence different from the straightforward extension of spatial anisotropic diffusion to the temporal domain [36] which requires both past and future frames. One can also use this extension with the fusion with the Schrodinger's equation [39] to derive a time-causal complex scale-space filter.

$$g_r(|\nabla u_r|) = 1/(1 + (|\nabla u_r|/\lambda_r)^2) \ , \quad r \in \{s, t\} \tag{7.4}$$

Note that in our framework of action recognition the exact location of the features have not been used. Moreover, small dislocation of the features should not be that much important. This is because the descriptor of the salient features are incorporated for the further processing and the descriptor is usually computed in a relatively wide region around the feature. Aside from

the fact that a non-linear filtering should produce fewer false positive detection, in registration tasks such as structure from motion, the well-localized features are essential and directly affect the performance.

7.2.3 Optimized MCS

In Section 5.2, we introduced the MCS as a multi-resolution analysis approach in which the salient features of each individual scale is fed into a separate classifier. The decision of all classifiers are then combined for the final decision making. For testing, the current implementation of the MCS determines the class label $k^* \in \{1, ..., K\}$ for the features of a given scale using the corresponding scale's classifier. That is the action signature S_i computed using the features of scale i are fed into only the classifier f_i and this classifier finds the class label c for this action signature. This implementation works fine when the video samples of a given action are captured at similar distance to the camera and with similar motion speeds, in both training and testing. In real world scenario, the spatial and temporal scales of the testing videos might not be the same as the training video samples or even those in the training set might be captured at different speeds or from different distance from the cameras. More specifically, the action signature S_i might have a better match in another spatio-temporal scale j^* in which $j^* \in \{1, ..., i, .., n\}$. That is the probability that the classifier f_j^* determines for the most probable class label of S_i is the maximum among those probabilities that the other classifiers can find the best match. This way the MCS will be optimized to perform a better combination on the classifiers.

$$j^* = \arg_j \max P(f_j|S_i) \tag{7.5}$$

$$P(f_j|S_i) = \max_k P(f_j^c|S_i) \ , \ \ \forall c \in \{1, ..., C\} \tag{7.6}$$

Since the features of finer scales are computed faster and are typically more discriminative than those of coarser scales (See Fig. 5.4 for the performance decrease with increase in the temporal scale), one might start with a sequential search to find the best match for the action signatures. When the classifier f_j^* finds the best match for the signature of a given scale, it is removed from the list of classifiers. One might consider different priors or heuristics to constraint the search for the best classifier f_j^*, such as the list of classifiers for the S_i which is computed from features of spatio-temporal scale (σ_m, τ_n) can include five neighbor classifiers trained using the features of immediate neighbors ($m = i \pm 1$, $n = i \pm 1$). This policy may avoid the poorest classifier being assigned to the coarsest scale. Moreover, it is more efficient, specially when there are many expensive classifiers to be searched. Note that the idea of searching for the optimum

matching of signatures can be applied to the multi-resolution action signature (Section 4.3) as well.

7.2.4 Combined symmetric-asymmetric features

The local salient features obtained from an asymmetric temporal filtering of the video signal are different from those obtained from a symmetric temporal filtering. Moreover, both type of features had shown promising performance for human action recognition task. Intuitively, a combination of both types of features should provide a more complete representation of the video and possibly better encoding and differentiation. The combination of symmetric and asymmetric features should provide a more dense sampling from the video contents, and thus, it might provide higher reproducibility score. Higher classification accuracy is therefore expected with the cost of more computation time.

7.2.5 Experiments and data sets

A general problem in the literature of human action recognition is that different publications consider different experimental settings even though they used the same data sets. This makes the direct comparison of different methods sometime difficult. Moreover, the problem with a well-defined training/testing setting is that a fair judgement on the performance of different methods is not possible. Possibly a better approach is to use different training/testing proportions in which samples are randomly divided. In addition, there is not that much of attempt to perform recognition across different datasets. That is training is performed on one dataset and testing on another data set. This way, a ranking of datasets might be obtained. To do this, one might want to start with the set of actions which is common across multiple datasets.

The three datasets of the Weizmann [11], the KTH [110], and the UCF sports that have been used in this thesis are among the standard datasets introduced to the literature of action recognition. In the Weizmann dataset, the camera is fixed and the body limbs are visible. The KTH dataset is more challenging. The video samples in this dataset were captured in both indoor and outdoor with variations such as fast zooming, blurring, illumination changes, and occlusion of the body limbs with cloth. In both Weizmann and the KTH datasets, however, the background is simple and there is only one person performing a single action. In addition, both Weizmann and KTH data sets contain multiple cycles of different actions. Test on more challenging data sets with possibly several people and more natural background is an important step towards better understanding of the behavior of different features and classifier approaches. The UCF sports dataset is one realistic example of this type of datasets that we used in this thesis.

There are several data sets designed for different objectives. For example, the INRIA Xmas Motion Acquisition Sequences (IXMAS [128]) contains videos of thirteen daily-live motions performed by a singe actor each three times and captured under different views. The HOHA (HOllywood Human Actions) datasets contain video clips of different human actions from the Hollywood movies. The first version of this dataset (HOHA1 [71]) contains 233 video samples of eight actions and the second version (HOHA2 [88]) contains 12 classes of human actions and 10 classes of scenes. This dataset has over 3669 video clips, approximately 20.1 hours of video in total. There are some clutters, background movements, occlusions, multiple people, and non-periodic actions in the HOHA data sets. In addition to these challenges, the VIRAT dataset [96] has been recently released for the surveillance applications with both ground camera videos and aerial videos captured with wide ranges of spatial and temporal resolutions, with $230Hz$ frame rates and 10-200 pixels in person-height. The VIRAT dataset is a first attempt towards large scale dataset for the human action recognition.

References

[1] E. H. Adelson and J.R. Bergen. Spatio-temporal energy models for the perception of motion. *Optical Society of America*, pages 284–299, 1985. 12, 20, 25, 29

[2] J. K. Aggarwal and S. Park. Human motion: Modeling and recognition of actions and interactions. *Second International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT'04)*, pages 640–647, 2004. 4

[3] J.K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, March 1999. 4

[4] A. Ali and J.K. Aggarwal. Segmentation and recognition of continuous human activity. *IEEE Workshop on Detection and Recognition of Events in Video*, 1:28–35, 2001. 4

[5] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. *IEEE International Conference on Computer Vision*, pages 1–8, 2007. 116

[6] S. Allaire, J. Kim, S. Breen, D. Jaffray, and V. Pekar. Full orientation invariance and improved feature selectivity of 3d sift with application to medical image analysis. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008. 16, 19

[7] J. A. Benediktsson B. Waske. Fusion of support vector machines for classification fo multisensor data. *IEEE Transaction on Geoscience and Remote Sensing*, 45(12):3858–3866, 2079. 91

[8] J. Babaud, A. Witkin, M. Baudin, and R. Duda. Uniqueness of the *G*aussian kernel for scale-space filtering. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 8:26–33, 1986. 20

[9] H. Bay, T. Tuytelaars, and L. Van Gool. **SURF**: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008. 7

[10] M.J. Black, G. Sapiro, D.H. Marimont, and D. Heeger. Robust anisotropic diffusion. *IEEE Transaction on Image Processing*, pages 421–432, 1998. 33, 34, 123

[11] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE International Conference on Computer Vision, Beijing, China*, pages 1395–1402, October 2005. 36, 37, 38, 39, 42, 44, 45, 46, 47, 48, 52, 53, 54, 56, 57, 58, 61, 62, 66, 67, 68, 76, 77, 78, 81, 97, 98, 105, 112, 113, 114, 116, 125

[12] Harris C. and M.J. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, pages 147–152, 1988. 7

[13] T. Calvert and A. Chapman. Analysis and synthesis of human movement. *Handbook of Pattern Recognition and Image Processing*, pages 432–474, 1994. 73

[14] A. Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. *IEEE Conference on Computer Vision and Pattern Recognition*, 1(4):846–851, 2005. 122

[15] C.C. Chang and C.J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 48

[16] I. Chang and C. Huang. The model-based human body motion analysis system. *Image and Vision Computing*, 18:1067–1083, 2000. 4

[17] R. Chaudry, A. Ravichandran, G. Hager, and R. Vidal. Histogram of oriented optical flow and binet-cauchy kernel on nonlinear dynamic systems for the recognition of human actions. *IEEE Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA*, pages 1932–1939, June 2009. 4, 5, 88, 122

[18] D. Chen, H. Wactlar, M. y. Chen, C. Gao, A. Bharucha, and A. Hauptmann. Recognition of aggressive human behavior using binary local motion descriptors. *IEEE Engineering in Medicine and Biology Society, Vancouver, Canada*, page 52385241, Auguest 2008. 71

[19] J. Choi, Y. Cho, T. Han, and H. S. Yang. A view-based real-time human action recognition system as an interface for human computer interaction. *Virtual Systems and Multimedia*, pages 112–120, 2008. 4

[20] K. D. Cock and B. D. Moor. Subspace angles and distances between arma models. *System and Control Letters*, 46(4):265–270, 2002. 122

[21] N. Cristianini and J.S. Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, UK, 2000. 10, 47, 74, 85, 86

[22] S. Dalal, M. Alwan, R. Seifrafi, S. Kell, and D. Brown. A rule-based approach to the analysis of elders activity data: Detection of health and possible emergency conditions. *AAAI: Association for the Advancement of Artificial Intelligence*, pages 1214–1220, 2005. 4, 7, 18, 86

[23] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. *Computer Vision and Pattern Recognition, San Juan, Puerto Rico*, pages 928–934, June 1997. 5

[24] J.W. Davis and A. Tyagi. A reliable-inference framework for recognition of human actions. *IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 169 – 176, 2003. 6

[25] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal filters. *IEEE International Workshop VS-PETS, Beijing, China*, pages 65–72, August 2005. 4, 5, 7, 9, 16, 17, 18, 19, 20, 24, 33, 41, 47, 49, 71, 73, 76, 84, 85, 86, 87, 88, 97, 113, 116, 123

[26] A. Dubrovina and R. Kimmel. Matching shapes by eigendecomposition of the laplace-beltrami operator. *Fifth International Symposium on 3D Data Processing Visualization and Transmission, Paris, France*, May 2000. 23

[27] R. Duits and B. Burgeth. Scale spaces on lie groups. *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 300–312, 2007. 22

[28] R. Duits, L.M.J. Florack, J. de Graaf, and B.M. ter Haar Romeny. On the axioms of scale space theory. *Journal of Mathematical Imaging and Vision*, 20(3):267–298, 2004. 22

[29] D. Fagerstrom. Temporal scale spaces. *International Journal of Computer Vision*, 64(2-3):97–106, 2005. 25

[30] D. Fagerstrom. Spatio-temporal scale-spaces. *International Conference on Scale Space and Variational Methods in Computer Vision, Ischia, Italy*, pages 326–337, June 2007. 12, 21, 25

[31] M. Felsberg, R. Duits, and L. Florack. The monogenic scale space on a bounded domain and its applications. *in Scale Space Conference*, pages 209–224, 2003. 22

[32] M. Felsberg, R. Duits, and L. Florack. The monogenic scale-space: A unifying approach to phase-based image processing in scale-space. *Journal of Mathematical Imaging and Vision*, 21:5–26, 2004. 22

[33] C. Fermller, J. Hui, and A. Kitaoka. Illusory motion due to causal time filtering. *Vision Research*, 50:315–329, 2009. 33

[34] G. Flitton, T. Breckon, and N. Megherbi Bouallagu. Object recognition using 3d sift in complex ct volumes. *British Machine Vision Conference*, pages 1–12, 2010. 16, 19

[35] M. G. F. Fourtes and A. L. Hodgkin. Changes in the time scale and sensitivity in the omatidia of limulus. *Journal of Physiology*, 172:239–263, 1964. 12, 19, 20, 25, 26

[36] I. Galic, J. Weickert, M. Welk, A. Bruhn, A. Belyaev, and H.S. Seidel. Image compression with anisotropic diffusion. *Journal of Mathematical Imaging and Vision*, 31(2-3):255–269, 2008. 123

[37] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7):1–18, 2008. 29

[38] W. Gerstner and W.M. Kistler. *Spiking Neuron Models. Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002. 26

[39] G. Gilboa, N.A. Sochen, and Y.Y. Zeevi. Complex diffusion processes for image filtering. *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 299–307, 2001. 22, 123

[40] T. Goodhart, P. Yan, and M. Shah. Action recognition using spatio-temporal regularity based features. *IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegad, USA*, pages 745–748, March 2008. 50, 116

[41] L. Grady. Random walks for image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(11):1–17, 2006. 26

[42] G. Guerra-filho and Y. Aloimonos. Human activity language: Grounding concepts with a linguistic framework. *1st International Conference on Semantics And digital Media Technology (SAMT), Athens, Greece*, 4306:86–100, 2006. 3

[43] G. Guerra-filho and Y. Aloimonos. Towards a sensorimotor wordnetsm: Closing the semantic gap. *3rd International WordNet Conference (GWC), Jeju Island, Korea*, 2006. 3

[44] C. Schmid C.L. Liu H. Wang, A. Klaser. Action recognition by dense trajectories. *IEEE Conference on Computer Vision and Pattern Recognition, Colorado Spring*, pages 3169–3176, June 2011. 50, 115, 116

[45] E. Hadjidemetriou, M.D. Grossberg, and S.K. Nayar. Multiresolution histograms and their use for recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 26:831–847, 2004. 35, 75

[46] T. Howley and M. G. Madden. The genetic kernel support vector machine: Decription and evaluation. *Artificial intellegence review*, 24:379–395, 2005. 86

[47] M. Shah J. Liu. Learning human actions via information maximization. *IEEE Conference of Computer and Pattern recognition, Anchorage, AK*, pages 1–8, June 2008. 71, 116

[48] M. Shah J. Liu, J. Luo. Recognizing realistic actions from videos "in the wild". *IEEE Conference of Computer and Pattern recognition, Miami, FL, USA*, page 19962003, June 2009. 71

[49] A.J. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000. 75, 78, 91

[50] S.H. Jung, Y. Guo, H. Sawhney, and R. Kumar. Multiple cue integrated action detection. *Computer Vision for Human Computer Interaction*, pages 108–117, 2007. 18

[51] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. *European Conference on Computer Vision, Prague, Czech Republic*, pages 228–241, May 2004. 7

[52] S.S. Keerthi and C.J. Lin. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation*, 15(7):1667–1689, 2003. 86, 88

[53] V. Kellokumpu, M. Pietikainen, and J. Heikkila. Human activity recognition using sequences of postures. *IAPR Conference on Machine Vision Applications*, pages 570–573, 2005. 4

[54] R. Kimmel, R. Malladi, and N. Sochen. Image processing via the beltrami operator. *Asian Conference on Computer Vision,Hong Kong*, January 1998. 23

[55] R. Kimmel, N. Sochen, and R. Malladi. From high energy physics to low level vision. *First Int. Conf. on Scale Space Theory in Computer Vision, Utrecht, Netherlands*, July 1997. 23

[56] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20:226–239, 1998. 89, 90, 91, 96

[57] A. Klser, M. Marszaek, , and C. Schmid. A spatio-temporal descriptor based on 3dgradients. *British Machine Vision Conference*, pages 995–1004, 2008. 7, 19, 116

[58] H. Knutsson and M. Andersson. Loglets: generalized quadrature and phase for local spatio-temporal structure estimation. *13th Scandinavian Conference on Image Analysis, Halmstad, Sweden*, pages 741–748, June 2003. 29

[59] J. J. Koenderink. Scale-time. *Biological Cybernetics*, pages 162–169, 1988. 12, 21, 24, 74

[60] A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, 2002. 9

[61] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30:1–50, 2000. 91

[62] V. Koltchinskii, D. Panchenko, and F. Lozano. Some new bounds on the generalization error of combined classifiers, 2001. 91

[63] P. Kovesy. Phase congruency: A low-level image invariant. *Psychological Research*, 64(2):136–148, 2000. 40

[64] A. Kuenzer, C. Schlick, F. Ohmann, L. Schmidt, , and H. Luczak. An empirical study of dynamic bayesian networks for user modeling. *Workshop on Machine Learning for User Modeling, Sonthofen, Germany*, pages 1–10, 2001. 121

[65] L. I. Kuncheva. *Combining Pattern Classifiers*. Methods and Algorithms, Hoboken (N.J.), Wiley, 2004. 90

[66] P. Bartlett L. Mason and M. Golea. Generalization error of combined classifiers, 2002. 91

[67] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64:107–123, 2005. 5, 7, 11, 15, 16, 74

[68] I. Laptev, B. Caputo, C. Schuldt, and T. Lindeberg. Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, pages 207–229, 2007. 4, 5, 7, 8, 9, 11, 15, 16, 18, 20, 23, 33, 34, 41, 47, 71, 74, 84, 85, 86, 87, 88, 90

[69] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. *ECCV Workshop "Spatial Coherence for Visual Motion Analysis"*, 3667:91–103, 2004. 18

[70] I. Laptev and T. Lindeberg. Velocity adaptation of space-time interest points. *IEEE International Conference on Pattern Recognition, Cambridge, UK*, pages 52–56, August 2004. 11, 24

[71] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *Conference on Computer Vision & Pattern Recognition*, June 2008. 126

[72] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. *IEEE Conference on Computer Vision and Pattern Recognition, Colorado Spring*, pages 3361–3368, June 2011. 50, 115, 116

[73] X. Li. Hmm based action recognition using oriented histograms of optical flow field. *IEEE Electronics Letters*, 43:560–561, 2007. 18

[74] H.T. Lin and C.J. Lin. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. 2003. 89

[75] T. Lindeberg. Scale-space for discrete signals. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12(3):234–254, 1990. 20

[76] T. Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision*, 11(3):283–318, 1993. 20

[77] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994. 10, 20, 22, 23, 45

[78] T. Lindeberg. Linear spatio-temporal scale-space. *International Conference on Scale Space and Variational Methods in Computer Vision, Utrecht, Netherlands*, pages 113–127, July 1997. 25

[79] T. Lindeberg. Generalized gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space. *Journal of Mathematical Imaging and Vision*, 40(1):36–81, 2010. 29, 30, 31

[80] T. Lindeberg, A. Akbarzadeh, and I. Laptev. Galilean-diagonalized spatio-temporal interest operators. *IEEE International Conference on Pattern Recognition*, pages 57–62, 2004. 7

[81] T. Lindeberg and D. Fagerstrom. Scale-space with causal time direction. *European Conference on Computer Vision, Cambridge, UK*, pages 229–240, April 1996. 12, 21, 25, 27, 29, 74

[82] A.P.B. Lopes, J. M. de Almeida, and A. de Albuquerque Araujo. Action recognition in videos: from motion capture labs to the web. *Preprint submitted to Computer Vision and Image Uunderstanding*, 2010. 19, 71, 85

[83] D. G. Lowe. Object recognition from local scale-invariant features. *International Conference on Computer Vision*, 60:1150–1157, 1999. 9

[84] D. G. Lowe. Distinctive image features from scale-invariant key points. *International Journal of Computer Vision*, 60:91–110, 2004. 5, 7, 16, 20, 23, 35, 36, 45

[85] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. *European Conference on Computer Vision, Graz, Austria*, pages 359–372, May 2006. 93

[86] J. K. Aggarwal M. S. Ryoo. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. *International Conference on Computer Vision, Kyoto, Japan*, pages 1593 – 1600, October 2009. 71

[87] S. Maji, A.C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. *IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska*, pages 1–8, June 2008. 86, 89

[88] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. *IEEE Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA*, pages 2929–2936, June 2009. 9, 10, 35, 71, 74, 75, 126

[89] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2006. 10, 16, 18

[90] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, pages 43–72, 2006. 10, 15, 20

[91] B. Mirkin. Choosing the number of clusters. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):252260, 2011. 88

[92] T. Mori, Y. Segawa, M. Shimosaka, and T. Sato. Hierarchical recognition of daily human actions based on continuous hidden markov model. *IEEE International Conference on Automatic Face and Gesture Recognition*, 73:779– 784, 2004. 4

[93] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis. 121, 122

[94] N.T. Nguyen, D.Q. Phung, S. Venkatesh, and H. Bui. Learning and detection activities from movement trajectories using the hierarchical hidden markov model. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:955–960, 2005. 4

[95] J.C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatialtemporal words. *British Machine Vision Conference, Edinburgh, UK*, 3:1249–1258, September 2006. 6, 19

[96] S. Oh, A. Hoogs, A. Perera, N., C.-C. Chen, J.T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. *Computer Vision and Pattern Recognition*, 2011. Dataset available at `http://vision.ics.uci.edu/papers/Sangmin_CVPR_2011/`. 126

[97] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal saliency for human action recognition. *IEEE International Conference on Multimedia and Expo.*, pages 430–433, July 2005. 5, 7

[98] A. Oikonomopoulos, I. Patras, M. Pantic, and N. Paragios. Trajectory-based representation of human actions. *AI for human computing*, pages 133–154, 2007. 15, 18

[99] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 75

[100] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 629–639, 1990. 12, 26, 40, 123

[101] G. Pirlo, C.A. Trullo, and D. Impedovo. A feedback-based multi-classifier system. *International Conference on Document Analysis and Recognition, Barcelona, Spain*, July 2009. 90, 91

[102] R. Polana and R. Nelson. Recognition of motion from temporal texture. *IEEE Conference on Computer Vision and Pattern Recognition, Champaign, IL , USA*, pages 129–134, June 1992. 3

[103] D. A. Pollen and S. F. Ronner. Phase relationships between adjacent simple cells in the visual cortex. *Science*, 212:1409–1411, 1981. 29

[104] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28:976–990, 2010. 4, 5

[105] T. Randen and J.H. Husoy. Filtering for texture classification: A comparative study. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21:291–310, 1999. 75

[106] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1454–1461, 2009. 50, 115, 116

[107] M.D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. *IEEE Conference on Computer Vision and Pattern Recognition, Alaska, USA*, pages 1–8, June 2008. 9, 36, 37, 40, 48, 69, 79, 83, 97, 98, 113, 116, 118

[108] F. Roli and G. Giacinto. *Design of Multiple Classifier Systems*. Hybrid Methods in Pattern Recognition: World Scientific Publishing, 2002. 90

[109] N.C. Rust, E.P. Simoncelli, and J.A. Movshon. How mt cells analyse the motion of visual patterns. *Nature Neuroscience*, 9(11):1421–1431, 2006. 20

[110] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. *IEEE International Conference on Pattern Recognition, Cambridge, UK*, 3:32–36, August 2004. 36, 37, 39, 47, 48, 59, 60, 76, 77, 78, 82, 86, 97, 98, 104, 112, 113, 116, 117, 125

[111] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional *SIFT* descriptor and its application to action recognition. *ACM Multimedia, Augsburg, Germany*, pages 357–360, September 2007. Code available at `http://www.cs.ucf.edu/~pscovann/`. 7, 18, 19, 20, 45, 47, 50, 71, 76, 116

[112] A. H. Shabani, J.S. Zelek, and D.A. Clausi. Human action recognition using salient opponent-based motion features. *IEEE Canadian Conference on Computer and Robot Vision, Ottawa, Canada*, pages 362 – 369, May 2010. 4, 11, 29, 47, 84, 85, 86, 87

[113] H. Shabani, D.A. Clausi, and J.S. Zelek. Towards a robust spatio-temporal interest point detection for human action recognition. *IEEE Canadian Conference on Computer and Robot Vision*, pages 237–243, 2009. 123

[114] C.A. Sugar and G.M. James. Finding the number of clusters in a data set: An information theoretic approach. *Journal of the American Statistical Association*, 98:750–763, 2003. 88

[115] B.M. ter Haar Romeny, L.M.J. Florack, and M. Nielsen. Scale-time kernels and models. *Scale-Space and Morphology in Computer Vision, Vancouver, Canada*, pages 255–263, July 2001. 12, 19, 20, 21, 24

[116] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society*, 63:411–423, 2001. 88

[117] R. Tibshirani, G. Walther, and T. Hastie. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005. 88

[118] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008. 71, 72

[119] V. Vapnik. *Statistical Learning Theory*. Wiley, NY, 1998. 10, 47, 85, 86

[120] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. *International Conference on Computer Vision, Kyoto , Japan*, pages 606 – 613, September 2009. 87, 94

[121] S. Vishwanathan, A. Smola, and R. Vidal. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *IJCV*, 73(1):95–119, 2007. 122

[122] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. *British Machine Vision Conference, London, UK*, pages 1–11, September 2009. 4, 7, 9, 11, 18, 19, 33, 37, 47, 48, 50, 73, 74, 76, 84, 85, 87, 88, 90, 115, 116

[123] J. Wang, B. Thiesson, Y. Xu, and M. Cohen. Image and video segmentation by anisotropic kernel mean shift. *European Conference on Computer Vision*, pages 238–249, 2004. 21

[124] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. *IEEE Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA*, pages 872–879, June 2009. 6, 41, 71, 113, 116, 121

[125] A. B. Watson and A. J. Ahumada. Model of human visual-motion sensing. *Journal of Optical Society of America*, 2(2):322–342, 1985. 12, 20, 25, 29, 30

[126] J. Weickert. A review of nonlinear diffusion filtering. *International Conference on Scale-space theory in computer vision, Utrecht, Netherlands*, pages 3–28, July 1997. 20, 22, 23, 40, 42, 123

[127] D. Weinland, M. Ozuysal, and P. Fua. Making action recognition robust to occlusions and viewpoint changes. *European Conference on Computer Vision, Heraklion, Greece*, pages 635–648, September 2010. 10, 113, 116

[128] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 103(2):249–257, 2006. 5, 126

[129] G. Willems, T. Tuytelaars, , and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. *European Conference on Computer Vision, Marseille, France*, pages 650–663, October 2008. 4, 5, 7, 11, 16, 17, 20, 23, 24, 34, 35, 38, 50, 74, 90, 115, 116

[130] A.P. Witkin. Scale-space filtering. *International Joint Conference Artificial Intelligence, Karlsruhe, Germany*, pages 1019–1022, August 1983. 20

[131] T. Xiang and S. Gong. Discovering bayesian causality among visual events in a complex outdoor scene. *IEEE International Conference on Advanced Video- and Signal-based Surveillance*, pages 177–182, 2003. 4, 121

[132] R.E. Schapire Y. Freund. *A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting*. 1995. 93

[133] C.-W. Ngo Y.-G. Jiang. Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. *Journal of Computer Vision and Image Understanding*, 113(3):405–414, 2009. 71

[134] G. Mori Y. Wang, P. Sabzmeydani. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. *2nd Workshop on Human Motion Understanding, Modeling, Capture and Animation (at ICCV),Rio de Janeiro, Brazil*, page 240254, October 2007. 71

[135] J. Zhang and S. Gong. Action categorization by structural probabilistic latent semantic analysis. *Computer Vision and Image Understanding*, 114(8):857–864, 2010. 6, 41, 71, 121

[136] Z. Zhang, Y. Hu, S. Chan, , and L.-T. Chia. Motion context: A new representation for human action recognition. *European Conference on Computer Vision, Marseille, France*, 4:817–829, October 2008. 116

[137] Z. Zhao and A. Elgammal. Information theoretic key frame selection for action recognition. *Biritish Machine Vision Conference, Leeds, UK*, pages 1–10, September 2008. 116