Efficient Human Activity Recognition in Large Image and Video Databases

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Muhammad Shahzad Cheema
aus
Sialkot, Pakistan

Bonn 2014

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter:	Prof. Dr. Christian Bauckhage
2. Gutachter:	Prof. Dr. Armin B. Cremers
Tag der Promotion:	01.09.2014
Erscheinungsjahr:	2014

ABSTRACT

Vision-based human action recognition has attracted considerable interest in recent research for its applications to video surveillance, content-based search, healthcare, and interactive games. Most existing research deals with building informative feature descriptors, designing efficient and robust algorithms, proposing versatile and challenging datasets, and fusing multiple modalities. Often, these approaches build on certain conventions such as the use of motion cues to determine video descriptors, application of off-the-shelf classifiers, and single-factor classification of videos. In this thesis, we deal with important but overlooked issues such as efficiency, simplicity, and scalability of human activity recognition in different application scenarios: controlled video environment (e.g. indoor surveillance), unconstrained videos (e.g. YouTube), depth or skeletal data (e.g. captured by Kinect), and person images (e.g. Flicker). In particular, we are interested in answering questions like (a) is it possible to efficiently recognize human actions in controlled videos without temporal cues? (b) given that the large-scale unconstrained video data are often of high dimension low sample size (HDLSS) nature, how to efficiently recognize human actions in such data? (c) considering the rich 3D motion information available from depth or motion capture sensors, is it possible to recognize both the actions and the actors using only the motion dynamics of underlying activities? and (d) can motion information from monocular videos be used for automatically determining saliency regions for recognizing actions in still images?

Our research answers these questions by proposing efficient and scalable approaches. Our methods are distinguished by naive but efficient feature extraction, sparse coding, instance-based learning and latent factor analysis. In particular, we (a) devise an efficient discriminative key poses approach for action recognition in videos that is independent of temporal context (b) present an efficient and scalable nearest affine hull method to HDLSS activity classification based on least squares optimization and QR-factorization (c) present a hierarchical bilinear factorization approach of style and content separation to recognize actions and actors in 3D data (depth, motion capture, motion history volumes) and (d) propose a non-negative matrix factorization based approach to determine action signatures from videos that are later used as saliency maps for classification of images. Our experimental results on a number of popular action datasets show significant achievements in terms of accuracy, scalability and efficiency.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Christian Bauckhage, who has the attitude and substance of a genius. I am much thankful to him for his continuous support of my Ph.D. research, for his patience, motivation, enthusiasm, and immense knowledge. One simply could not wish for a better or friendlier supervisor. I would also like to thank Prof. Armin B. Cremers, my co-advisor, for his valuable feedback and suggestions. His continued support has played a vital role in my thesis. I am grateful to Prof. Jürgen Gall for several insightful discussions at biweekly PhD seminars and for acting as a committee member. Special thanks go to Prof. Lutz Plümer, who was willing to participate in my final defense committee though it was hardship.

I would like to acknowledge the financial support of the German Research Foundation (DFG), the German Academic Exchange Service (DAAD) and the Higher Education Commission of Pakistan (HEC). In particular, this work was carried out in the DFG project "automatic activity recognition in large image databases". The academic and technical support I have received from the Bonn-Aachen International Center for IT (BIT) and the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) has been indispensable.

I am thankful to my colleagues at BIT and IAIS who have contributed immensely to my personal and professional time; especially Abdulrahman Eweiwi for feedback on my thesis, the stimulating discussions, for the sleepless days and nights we worked together before deadlines, and for all the fun we have had in the last three years. The contributions of Dr. Christian Thurau (now at GameAnalytics) have been invaluable for building a sound foundations for my research. I greatly appreciate my friends in Bonn, Lahore, Sahowala, and elsewhere for their continued love and support.

The best outcome from the last four years is having a sweet family: Anisa (my wife), Abdulrahman(my son) and Aroush (my newborn daughter). Thanks for being the source of my eternal happiness. Having you people around me has been key to overcoming any stressful moments. I owe special gratitude to my father, Muhammad Sharif Cheema, for his faith in me and for allowing me to be as ambitious as I wanted.

Dedicated to my beloved mother,
dearest Hamida Bano...
gone but never forgotten...

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Overview/Motivation

Visual perception is the ability to detect and process visible light to interpret the surrounding environment. Computer vision is the field of science that aims to duplicate the abilities of the human visual system by electronically perceiving and understanding real world imagery. It includes methods for acquiring, processing, analyzing, and understanding images, videos, and other high-dimensional data from the real world in order to produce numerical or symbolic information. Typical tasks and applications of computer vision include object or event detection, recognition, segmentation, pose estimation, face recognition, content-based indexing and retrieval, tracking, scene recognition and reconstruction, and image restoration.

Among the many aspects of computer vision, the problem of recognizing human activities in videos and still images has become an increasingly popular due to its demand and applications in a range of areas. Some of its application areas are automated surveillance systems, content based indexing and searching on Web, healthcare monitoring, and human-machine interaction in intelligent environments. The need for automatic activity recognition in multimedia data is a natural consequence of the recent developments in technology and consumer behavior. However, despite significant advancements in the acquisition and the availability of video data, progress towards automatic activity recognition is still rather limited.

Consider, for instance, large multimedia portals and social media sites such as YouTube, Facebook, and Flicker, which have become significantly more popular in our daily lives. On YouTube alone, over a hundred hours of video is uploaded every minute[1]. Flickr hosts over 8 billion images and more than 3.5

[1] http://www.youtube.com/yt/press/statistics.html accessed on 15/11/2013.

million images are uploaded daily[2]. Video or image retrieval in these large-scale archives is currently possible only at the cost of expensive manual annotation.

CCTV video surveillance is another scenario in which a lot of data is generated but computer vision research lacks the capability to provide large-scale recognition. One such case is the CCTV network of London, which has over one million CCTV cameras; yet, there are hardly any cases of crime prevention or automatic detection of crimes. Often investigators have to record and analyze the data manually, which is an obvious bottleneck for large-scale event monitoring.

Furthermore, the advances in sensor technology, such as the invention of Microsoft Kinect sensors[3], has boosted low-cost imagery capture in the form of different modalities. These systems provide multimodal data (depth, RGB, sound, skeleton) that offer a rich perception of the environment and human activities. Such advancements have also inspired researchers to think of out-of-the-box applications of human activity recognition such as physiotherapy exercises, interactive gaming, and smart homes.

1.2 Context/Problem Statement

The evolution of data and explosion of applications have motivated researchers to develop novel techniques to better solve the activity recognition problem. Surveying the literature on human activity recognition, one notes an intriguing trend towards developing ever more sophisticated representations and complex algorithms. Often, basic but important issues such as efficiency, scalability, and simplicity are overlooked. In this thesis, we address these issues across different application areas of human activity recognition.

In particular, we consider efficiency and scalability of human activity recognition in four scenarios of increasing complexity: controlled video scenarios such as video surveillance, uncontrolled and unconstrained video databases such as YouTube, multimodal emerging environments such as those captured through Kinect sensory, and still images. Fig.1.1 shows example images of different

[2] http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer accessed on 15/11/2013

[3] http://www.xbox.com/en-US/kinect

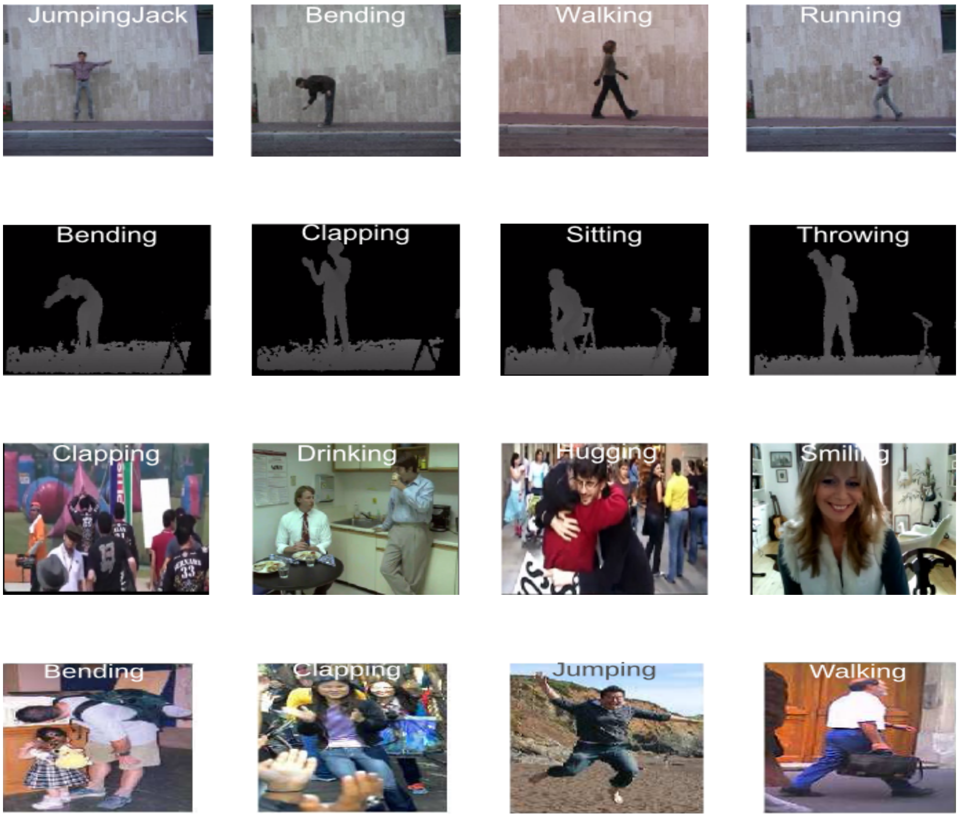


Figure 1.1: Example screen shots and action images from different data sources: RGB videos in a controlled environment (top) , Kinect depth videos in an interactive setup (second row), clips from a typical consumer video database (third row), and still images from a popular action dataset (last row)

actions taken from representative image and video datasets. The frames in the first row are taken from the Weizmann dataset [Blank et al., 2005] that contains videos with a rather static background. The gray-scale frames in the second row are representative of the depth imagery in the Berkeley Multimodal Human Action Dataset (MHAD) [Ofli et al., 2013] that contains multimodal RGBD and skeletal data acquired by using Kinect sensor imagery. Frames from the HMDB database [Kuehne et al., 2011] of unconstrained consumer and commercial videos are shown in the third row. Finally, images in the fourth row are taken from the popular PASCAL-VOC2011 and H3D image databases [Everingham et al., Bourdev and Malik, 2009].

Activity recognition in different setups poses challenges of different natures and scales. For example, most approaches to activity recognition in controlled video or 3D data depend on reliable pose estimation. Usually, these approaches benefit from global representations of human shape or motion in segmented video data [Bobick and Davis, 2001, Blank et al., 2005, Ni et al., 2011]. In realistic videos, human pose estimation is rather challenging due to factors such as dynamic background, occlusion, poor illumination conditions, and camera motion. It is also known that the local representations of the motion components and scenery are more beneficial in realistic videos [Laptev et al., 2008, Kliper-Gross et al., 2012, Wang et al., 2013]. As a result, there is no single solution that works best in all scenarios. Still, our overall scientific investigation is based on general principles such as simple but efficient feature extraction, sparse coding, instance-based learning, and latent factor models. For each scenario, our research takes into account the complexity of action recognition, provides answers to intuitive questions, and proposes efficient and scalable solutions. Below we briefly describe significant challenges for action recognition in different environments.

Recognizing human actions in a controlled environment has been extensively studied for the last decade. Most existing research in this domain has focused on pose estimation, temporal modeling, and spatio-temporal feature extraction [Bobick and Davis, 2001, Blank et al., 2005, Cheema et al., 2011]. These approaches apply background subtraction and represent human activity in the temporal context in terms of global (e.g. spatio-temporal volumes) or local (e.g. interest-point based representation) features. Large scale activity recognition in such domains (e.g. surveillance) faces several challenges such as missing frames, observational latency, and online classification. Therefore, methods that rely heavily on temporal information may suffer due to these factors. Interestingly, there is hardly any investigation about the extent to which the temporal cues are important when using naive features such as silhouettes or human contours. In fact, there are only a few methods of action recognition in controlled setups that rely only on human pose and work without exploiting temporal information.

In the context of unconstrained videos, most research has focused on developing more informative features to obtain a single representation of a video. The most common approach is to extract local features around sparse space-time interest points or dense trajectories [Laptev et al., 2008, Kliper-Gross et al., 2012, Wang et al., 2013]. After extracting local features around those key points, a bag-of-words (BoW) approach is adopted. The BoW approach employs (a) a clustering algorithm to determine a *code book* of representative *words* (centroids) in the data and (b) the quantization of the video features according to the code book. Other approaches, such as [Sadanand and Corso, 2012], use a large number of templates to spot certain types of motions. In either case, the resulting feature descriptors are (a) high dimensional because of a large number of code words, (b) sparse due to a high degree of variation within classes, and (c) scarce due to a large number of classes. For example, the action-bank features [Sadanand and Corso, 2012] proposed for large-scale activity recognition in unconstrained videos, such as HMDB, are nearly 15,000 dimensional; where HMDB contains nearly 100 labeled samples for each of the 51 action categories. Such High Dimension Low Sample Size (HDLSS) data is prone to *neighborliness* – the lack of neighborhood among the instances in a very high dimensional space [Donoho and Tanner, 2005, Ahn et al., 2007]. Asymptotic studies reveal a tendency for HDLSS data to lie at the vertices of a regular simplex [Hall et al., 2005, Donoho and Tanner, 2005]. Existing methods of human activity recognition in the wild, however, overlook the underlying distribution of the data and employ off-the-shelf classifiers such as Support Vector Machines (SVM) and Nearest Neighbor (NN) algorithms. Such naive application may cause over-fitting, e.g. in the case of SVMs that are observed to use nearly all the training data as support of the decision function; or it may result in under-fitting, e.g. in the case of NNs that assume that nearby points have the same label (high inductive bias).

We also deal with human activity recognition in (controlled) 3D environments where data is acquired using depth or motion capture. Since the introduction of the Kinect sensor, there have been numerous applications for such environments, including online exercises, multi-player games, and smart homes. Recent research on human action recognition in such scenarios has shown promising results [Wang et al., 2012, Ofli et al., 2013] – thanks to the rich pose and motion information provided by these sensors. In this context, we *extend* the problem of human action recognition by proposing an approach that allows identification of both an action and the performing actor solely using the dynamics of the activity.

Throughout this thesis, we evaluate our algorithms using a number of public datasets. These datasets vary with respect to such factors as size, modality (video, silhouette, RGB, motion capture, depth), number of persons involved (individuals or groups), number of view points or cameras, and type of environment (controlled, uncontrolled). The availability of public datasets serves two purposes. Firstly, it enables researcher to focus on algorithms and test their ideas on existing data. Secondly, it provides benchmarks for a fair comparison of the merits and demerits of different methods.

1.3 Contributions

Our research investigates several issues related to scalable action recognition, purposes new algorithms, and provides extensive empirical evaluation – often outperforming state-of-the-art techniques. Below we give an overview of our contributions.

1.3.1 Efficient Action Recognition in Controlled Environment by Learning Discriminative Key Poses

For action recognition in a constrained or surveillance environment, most existing techniques model human activities using motion cues such as motion energy images [Bobick and Davis, 2001], space-time volumes [Blank et al., 2005], and motion history volumes [Weinland et al., 2006]. A natural question to ask is *is it possible to recognize human actions by only looking at a few key frames?* or in other words *is it possible to recognize human actions in videos without explicit modeling of temporal cues?* Such a system would be robust with regard to missing data and observational latency and may prove more efficient than methods based on temporal cues.

We propose a novel instance-based approach that learns non-temporal discriminative key poses for actions in videos. Given a set of training videos, we first extract a scale-invariant contour-based pose feature from silhouettes of each video. Then, we cluster the features in order to build a set of prototypical key poses. Based on their relative discriminative power for action recognition, we learn weights (latent components) that favor distinctive key poses.

Finally, classification of a novel action sequence is based on a simple and efficient weighted voting scheme that augments results with a confidence value which indicates recognition uncertainty. The proposed approach is efficient and delivers real-time performance. In experimental evaluations for single- and multi-view action datasets, it shows high recognition accuracy. The same contour-based feature and a similar non-temporal approach achieves state-of-the-art recognition rate when applied to large scale multi-view gait recognition problem.

1.3.2 Efficient Instance-based Classification of Large Scale Unconstrained Image and Video Data

With the advancement of Web and social media, recognizing activities in consumer videos has become a vital area of research in the past few years. Emerging large-scale unconstrained video datasets such as HMDB [Kuehne et al., 2011] and UCF50 [Reddy and Shah, 2013] are rather challenging as compared to classical clean and controlled datasets such as KTH [Laptev and Lindeberg, 2003], IXMAS [Weinland et al., 2006], and CASIA [Yu et al., 20006]. The dimensionality of spatio-temporal features for unconstrained videos often ranges into the tens of thousands whereas the number of classes and the number of labeled instances per class usually ranges from ten to a hundred. The resulting High Dimension Low Sample Size (HDLSS) data often suffers from lack of neighborhood due to the curse of dimensionality [Hall et al., 2005, Donoho and Tanner, 2005]. Existing approaches to human activity recognition in the wild often use off-the-shelf techniques such as Nearest Neighbor classifiers or Support Vector Machines without paying attention to the geometry of the underlying feature or instance space. It becomes important to investigate and determine suitable classifiers that could efficiently recognize human actions in such data.

In this thesis, we first investigate the performance of different classifiers to HDLSS action videos and, through extensive empirical evaluation, affirm the lack of proper neighborhood in such data. As the lack of neighborhood undermines the proximity-based methods (e.g. kNN) and may cause over-fitting for discriminant methods (e.g. SVM), we propose a novel least square- and QR factorization-based approach to Nearest Affine Hull (NAH) classification which remedies the HDLSS dilemma and noticeably reduces the time and memory requirements of existing methods. We show that the resulting non-parametric models provide smooth decision surfaces and yield efficient and accurate solutions in multi-class HDLSS scenarios. On several action recognition benchmarks, the proposed NAH classifier outperforms other instance-based methods and shows competitive or superior performance compared to SVMs. In addition, for online settings, the proposed NAH method is faster than online SVMs.

1.3.3 Simultaneous Recognition of Style and Contents in 3D Videos using Bilinear Tensor Factorization

The introduction of low-cost image capturing e.g. Kinect has led a paradigm shift from action recognition in 2D to rich action recognition in a 3D environment. A lot of research efforts are now devoted to recognizing human actions in such skeletal or depth data. Most successful approaches in this domain build on principles of existing 2D pose- or part-based methods [Wang et al., 2012, Ofli et al., 2013] and achieve high action recognition performance. Low-cost access to such 3D data and annotated poses has motivated researchers to think of new applications such as healthcare, multi-player virtual games, and advanced home security. Autonomous recognition of people based on their individual ways of performing different actions can be of great importance in these scenarios. At first, it may look too optimistic but there are psychophysics studies that suggest that human have different styles to perform different actions [Cutting and Kozlowski, 1977, Thoroughhman and Shadmehr, 1999]. In this line, we exploit the rich information content from depth/skeletal imagery and latent bilinear tensor factorization to empirically validate these studies. Specifically, we investigate a novel question: *Is it possible to recognize both the actions and the actors in surveillance videos using only motion dynamics of underlying activities?*

To this end, we present a hierarchical approach that is based on conventional action recognition and asymmetrical bilinear modeling. In particular, we employ bilinear factorization on the tensorial representation of action videos to characterize styles of performing different actions. The proposed approach is solely based on the dynamics of the underlying activity. The model is evaluated on the IXMAS [Weinland et al., 2006] and the Berkeley-MHAD [Ofli et al., 2013] datasets using different modalities based on optical motion capture, Kinect depth videos, and 3D motion history volumes. In each case, high recognition accuracy is achieved in comparison to a too strict symmetric bilinear modeling and a too loose Nearest Neighbor classification. Consequently, our approach extends motion-based person identification to multiple common actions and shows that the identification is not limited to walking or running actions.

1.3.4 Determining Salient Regions for Action Recognition in Still Images using Videos

Recognizing human actions in still images is a challenging problem due to challenges such as occlusion, texture, and lack of any motion information. A common approach to human action recognition from still images consists in computing local descriptors for classification [Bay et al., 2008, Harris and Stephens, 1988, Jhuang et al., 2007]. Typically, these descriptors are computed in the vicinity of key points which either result from running an appearance-based key point detector or from dense random sampling of pixel coordinates. Such key points are not a-priori related to human activities and thus might not be very informative with regard to action recognition. Other saliency-based approaches determine regions of interest by tracking human visual attention [Vig et al., 2012, Mathe and Sminchisescu, 2012]. Alternatively, many recent approaches are based on detecting *poslets* [Bourdev and Malik, 2009] (manually labeled image patches). Determining attention-based saliency regions and constructing poselets both involve significant manual effort in the training phase. On the other hand, appearance-based approaches to determine key points suffer severely from background and texture factors.

Since an action is often described in terms of articulations of different body parts, we address the issue: *can motion information from simple videos be used for determining saliency regions or important body parts for recognizing actions in still images?* If so, efficient and large-scale image classification can be obtained by focusing on feature extraction from salient regions. This is in contrast to sparse or dense sampling of image patches, as they do not regard task specific objectives in key point localization and typically assume key points to be independent and therefore fail to explain characteristic spatial and temporal layouts.

In this line, we investigate the possibility and applicability of identifying action-specific points or regions of interest in still images based on information extracted from controlled video data. We propose a novel method for extracting spatial interest regions or action signatures where we apply Non-negative Matrix Factorization (NMF) to optical flow fields extracted from videos in person bounding boxes. The resulting basis flows are found to indicate image regions that are specific to certain actions and therefore allow for an informed sampling of key points for feature extraction. We thus present a generative model for action recognition in still images that allows for characterizing joint distributions of regions of interest, local image features (visual words), and human actions. Experimental evaluation shows that our approach is able to extract interest regions that are highly correlated to those body parts most relevant to different actions. As a result, it achieves higher action recognition accuracies than recent baseline methods.

1.4 Thesis Organization

Chapter 2 provides an overview of related research on activity recognition in videos. Most approaches with relevance to ours are discussed in appropriate detail in subsequent chapters as well. In Chapter 3, we describe an efficient approach of action recognition in controlled videos by learning discriminative key poses. Application of key pose based classification is further extended to gait recognition in Chapter 4. The issue of classification of HDLSS data is discussed in Chapter 5, where we investigate different instance-based methods and propose an efficient affine hull based method. Chapter 6 lays out a novel approach to simultaneously recognize actions and actors in depth and skeletal data. Chapter 7 sets forth our approach to obtain action signatures from action videos in order to use them for action classification in still images. Finally, Chapter 8 concludes findings of this research and discusses some possible future directions.

1.5 Related Publications

The list of accepted and submitted articles and their contribution to this thesis is given below.

[1] Cheema, M.S., Eweiwi A., Thurau C., and Bauckhage C., Action Recognition by Learning Discriminative Key Poses, PERHAPS workshop at 13th IEEE International Conference on Computer Vision (ICCV), Spain, 2011

[2] Eweiwi A., Cheema, M.S., Thurau C., and Bauckhage C., Temporal Key Poses for Human Action Recognition, PERHAPS workshop at 13th IEEE International Conference on Computer Vision (ICCV), Spain, 2011

[3] Cheema, M.S., Eweiwi A., and Bauckhage C., Gait Recognition by Learning Distributed Key Poses, 19th IEEE International Conference on Image Processing (ICIP), Florida, October 2012

[4] Cheema, M.S., Eweiwi A., and Bauckhage C., Who is Doing What? Simultaneous Recognition of Actions and Actors, 19th IEEE International Conference on Image Processing (ICIP), Florida, October 2012

[5] Cheema, M.S., Eweiwi A., and Bauckhage C., Human Activity Recognition by Separating Style and Content, Pattern Recognition Letters Journal, 2013 (In press), DOI http://dx.doi.org/10.1016/j.patrec.2013.09.024

[6] Cheema, M.S., Eweiwi A., and Bauckhage C., High Dimensional Low Sample Size Activity Recognition Using Geometric Classifiers, under review at Digital Signal Processing, 2014

[7] Eweiwi A., Cheema, M.S., and Bauckhage C., Action Recognition in Still Images by Learning Spatial Interest Regions from Videos, Pattern Recognition Letters (Accepted), 2014

[8] Eweiwi A., Cheema, M.S., and Bauckhage C., Discriminative Joint Non-negative Matrix Factorization for Human Action Classification, German Conference on Pattern Recognition (GCPR/DAGM), 2013

[9] Cheema, M.S., Eweiwi A., and Bauckhage C., A Stochastic Late Fusion Approach to Human Action Recognition, submitted to German Conference on Pattern Recognition (GCPR/DAGM), 2014

[1] and [2] respectively present novel template- and pose-based methods for action recognition in controlled video environments. Chapter 3 explains our methodology used in [1] in detail. [3] is partially included in Chapter 4. Chapter 6 is based on [4] and [5] that present our bilinear modeling approach to

multi-factor action recognition. [6] is based on Chapter 5 that proposes an efficient classifier suitable to large-scale unconstrained videos and images. [7] is based on Chapter 7 where author's major contribution is determining the salient regions in videos using NMF and designing a generative framework to classify images. [8] proposes a joint NMF approach for action recognition in still images using multiple features. In [9], a principled late fusion approach is presented that also utilizes multiple features to enhance action recognition in videos and still images.

Chapter 2

A Review of Vision-based Action Recognition in Videos

The demand of automatic activity recognition in different scenarios has led to significant research efforts. Given the diversity of these areas, researchers have worked on different aspects of the problem. Accordingly, the followed approaches vary significantly. A number of review and survey papers have been published in the last decade [Turaga et al., 2008, Aggarwal and Ryoo, 2011, Poppe, 2010, Ke et al., 2013, Chaquet et al., 2013, Jiang et al., 2013]. In this chapter, we provide a review of most relevant literature on human action recognition in videos.

Typical components of an action recognition system include: the target domain and environment, modality and representation of the data, and the classification approach. Figure 2.1 shows a general block diagram. In this chapter, we discuss these components in an incremental and inclusive manner. First, we give an overview of the context of the action recognition problem and the relevant datasets in Section 2.1. Popular feature representations and their characteristics are discussed in Section 2.2. Finally, different classification approaches to action recognition, in accordance with the context and feature representations, are described in Section 2.3.

2.1 The Context and The Data

Understanding target application and characteristics of the input data is most critical to envisioning the subsequent scientific challenges. There are various environmental factors that affect the complexity of automatic activity recognition. Some of those are: complexity of motion dynamics, number of individuals and objects in the scene, type of environment, type of interactions among ob-

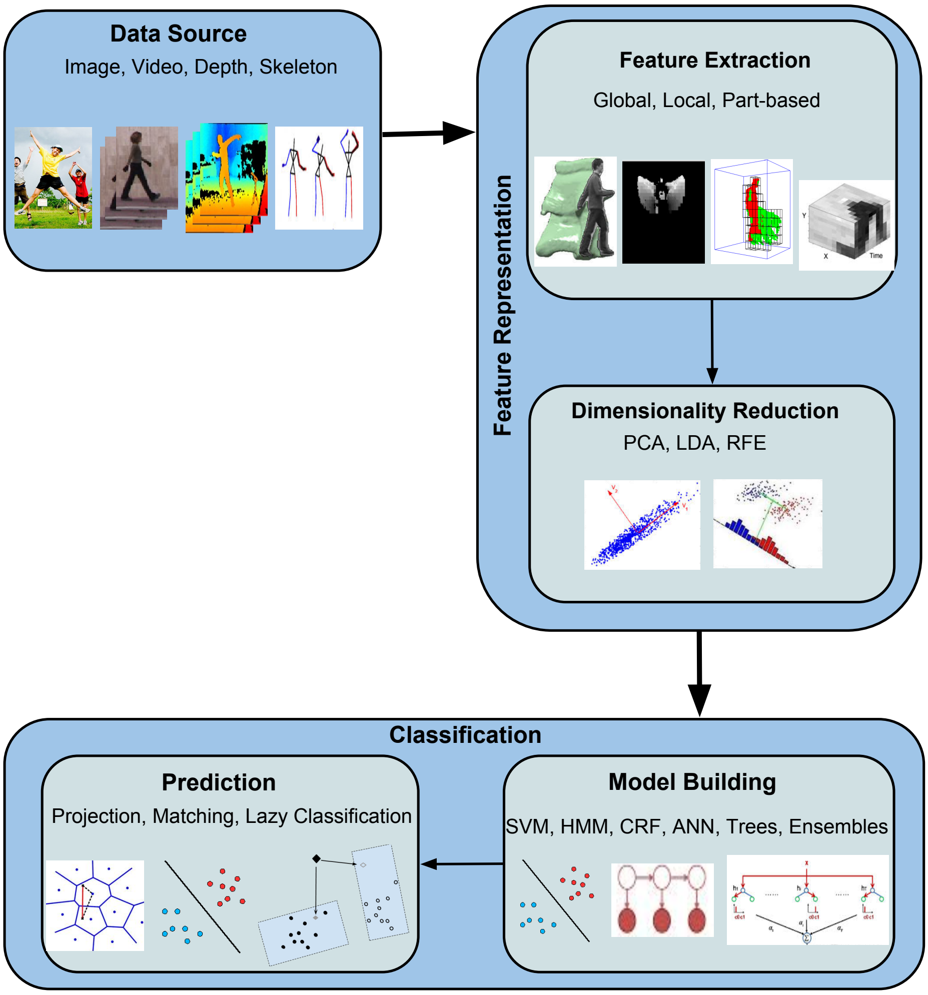

Data Source
Image, Video, Depth, Skeleton
Feature Representation
Feature Extraction
Global, Local, Part-based
Dimensionality Reduction
PCA, LDA, RFE
Classification
Prediction
Projection, Matching, Lazy Classification
Model Building
SVM, HMM, CRF, ANN, Trees, Ensembles

Figure 2.1: Block diagram of a typical action recognition system

jects, quality of videos or other input sources, and number of viewpoints or cameras. For instance, based on the complexity of motion dynamics, human activities can be broadly categorized into three classes: gestures (atomic poses or short sequences, e.g. raising an arm), actions (single-person activities that may be composed of different gestures in a sequence, e.g. running), and activities (complex actions or interactions involving two or more persons and/or objects, such as hand-shaking or riding bicycle). Throughout this thesis, we will refer to the terms gesture, action and activity in an inclusive manner, i.e. ultimately any sequence is considered an activity. Often, we will use the words action and activity interchangeably. Considering the complexity of activities, scenes, and other factors, we categorize the available video datasets into three broad classes: controlled (e.g. monocular video clips recorded under controlled

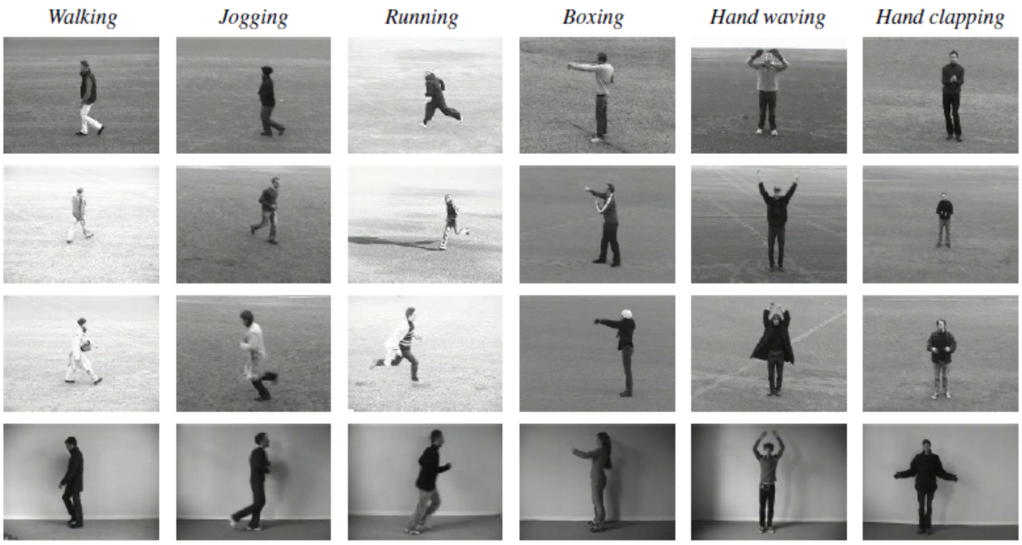


Figure 2.2: KTH action dataset [Schueldt et al., 2004]

conditions), unconstrained (e.g. Youtube or movie clips), and multimodal (e.g. recorded through multiple cameras or multiple sensors such as Kinect). Below we discuss each of these classes with example datasets. For a detailed survey on video datasets for human action recognition, see [Chaquet et al., 2013].

2.1.1 Controlled Video Datasets

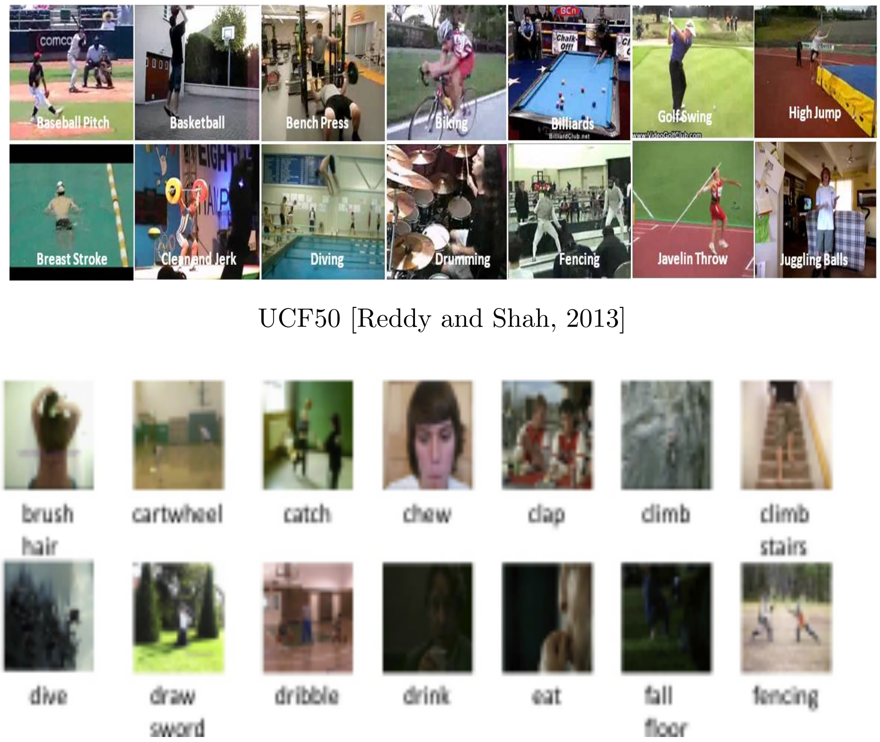
The earliest challenge to vision-based human action recognition was to recognize a single action of a single human in a video from a single viewpoint. To this end, early datasets like KTH [Schueldt et al., 2004] and Weizmann [Blank et al., 2005] are most popular. These datasets contain clips where a single person is performing a single action in a controlled indoor or outdoor environment with a static background. In such cases, background subtraction and human localization are reasonably reliable. Most approaches developed around these datasets focus on pose estimation, holistic feature representations, and state-space modeling (e.g. HMM, CRF).

The Weizmann collection [Blank et al., 2005], recorded in 2005, is one of the earliest and most popular controlled datasets. Most state-of-the-art approaches at that time were based on feature tracking, which could not properly deal with self-occlusions. Moreover, they could only recognize periodic actions such as walking and running. The dataset was constructed to encourage new approaches based on considering actions as space-time shapes (volumes). It contains video samples for 10 different actions performed by 9 actors. These actions are: walking, running, jumping, galloping sideways, bending, one-hand waving, two-hands waving, jumping in place, jumping jack, and skipping. The background is relatively simple, the view is static, and only one person is acting in each video. The KTH human action dataset [Schueldt et al., 2004] is another controlled dataset. It contains six types of actions (boxing, hand clapping, hand waving, jogging, running, and walking) performed several times by 25 subjects in four different scenarios (indoors, outdoors, outdoor with scale variation, outdoors with different clothes). There are a total of $25 \times 6 \times 4 = 600$ videos for each combination of 6 actions, 4 scenarios, and 25 individuals. Example frames of this dataset are shown in Fig. 2.2. It was the first dataset for which features were extracted using space-time interest points.

2.1.2 Large-scale Unconstrained Video Datasets

Access to the internet and the popularity of social media portals such as Facebook and YouTube have revolutionized our daily lives over the last decade. Efficient utilization of massive amounts of multimedia activity data, which is often noisy and unconstrained, requires robust algorithms that can organize, store, analyze, and retrieve this data. This development has posed several challenges to the conventional action recognition approaches that were designed to work in controlled environments. To aid research in this domain, a number of challenging datasets have been proposed so far, and the trend is going on [Laptev et al., 2008, Kuehne et al., 2011, Reddy and Shah, 2013, Soomro et al., 2012]. The realistic videos in these databases pose a number of challenges due to camera motion, different viewpoints, cluttered background, poor illumination conditions, large inter-class variations, occlusions, and poor quality of the medium. These challenges cause most approaches designed for controlled environments to fail. While the problem is largely unsolved, approaches that are based on extracting spatio-temporal features around interest points or motion trajectories have shown promising results [Laptev et al., 2008, Wang et al., 2011, Reddy and Shah, 2013, Kliper-Gross et al., 2012]. Figure 2.3 shows screen shots of different action videos in the two largest datasets.

UCF50 [Reddy and Shah, 2013] is one of the largest activity recognition



UCF50 [Reddy and Shah, 2013]

HMDB [Kuehne et al., 2011]

Figure 2.3: Screen shots of different sample videos from two large unconstrained action datasets

datasets, containing 6,618 videos of 50 activity classes. Most activities contain multiple persons, object interactions, dynamic backgrounds and a high degree of variation. For each activity, there are at least 100 videos split into 25 or more groups. Each group contains clips cropped from the same video, i.e. videos in a group share the same scene context. This grouping allows leave-one-group-out and leave-k-groups-out cross validation so that training and test data do not share videos that are very similar. HMDB51 [Kuehne et al., 2011], or HMDB in short, is another large-scale activity recognition dataset that contains unconstrained videos collected from a variety of sources, ranging from commercial movies to YouTube videos. It contains 6,766 videos belonging to 51 activity classes. The range of activities include: general facial gestures such as smiling, chewing, and laughing; facial actions with object manipulation such as eating, drinking, and smoking; general body movements such as hand clapping, climbing, and diving; body movements with object interaction such as hair brushing, sword drawing, and bike riding; body movements for human actions such as fencing, hugging and hand shaking. A protocol containing three train-test splits is provided in [Kuehne et al., 2011]. For every class and split, there are 70 videos for training and 30 for testing.

While controlled datasets such as KTH and Weizmann are criticized for the lack of dynamic environments, the unconstrained datasets are criticized for the lack of sufficient labeled data to model human activities in context. In the latter case, the videos (and images) usually have to be manually segmented, labeled and preprocessed to obtain the ground truth. Recently, Torralba and Efros analyzed several popular "'object recognition in the wild"' datasets and concluded that such collections can lead to biased results [Torralba and Efros, 2011]. This bias can limit the progress of algorithms developed and evaluated against such datasets. This has been lately realized by the action recognition community and some recent work has focused on providing annotations or bounding boxes. For example, Jhuang et al. [2013] presents the labeled joints database JHMDB for a subset of videos from HMDB. The homepage[4] of recently proposed UCF101 action dataset also provides person bounding boxes for 24 actions.

2.1.3 Multimodal Datasets

The advances in low-cost imagery devices has created new domains of research. For example, there is increasing interest in utilizing multiple cameras and other sensors for monitoring large public spaces such as train stations, shopping malls, and airports. Also, recent inventions in active depth sensing, e.g. launch of the Microsoft Kinect, has caused a revival of interest in 3D human motion tracking, pose estimation and action recognition in RGB+D data. These devices are more suitable for indoor environments where they offer many opportunities for automatic visual perception. The available multimodal data offer a rich presentation of the underlying motions due to multiple input sources (multiple views, RGB, depth, and skeleton). In this line, different multi-view datasets, including IXMAS [Weinland et al., 2006], MuHAVi [Singh et al., 2010] and the CASIA gait dataset [Yu et al., 20006], have emerged in the last decade. Datasets that are captured using depth or motion capture sensors include: MHAD [Ofli et al., 2013], MSR-DailyActivity3D [Wang et al.,

[4] http://crcv.ucf.edu/data/UCF101.php

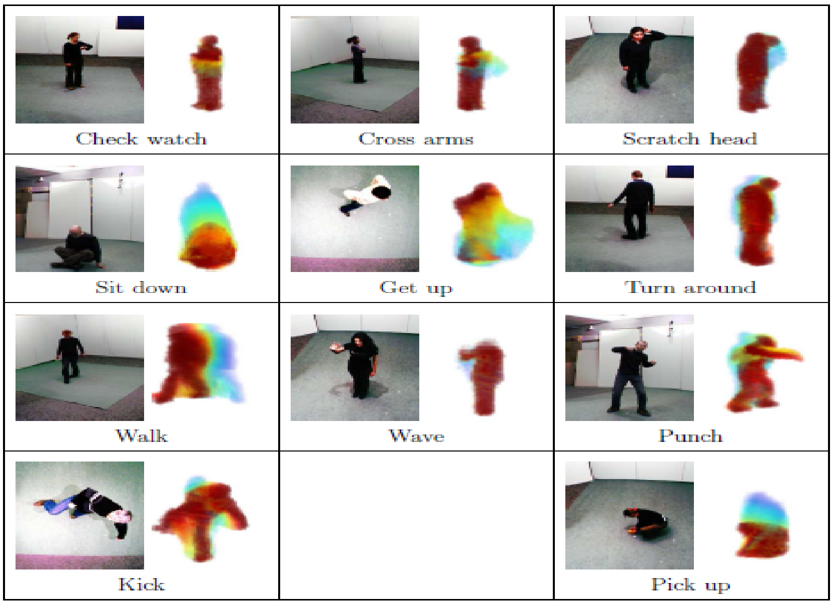


Figure 2.4: IXMAS actions and motion history volumes [Weinland et al., 2006]

2012], and TUM-Kitchen [Tenorth et al., 2009].

The importance of multimodal data in unconstrained environment has recently been adhered by many researchers [Hadfield and Bowden, 2013, Jhuang et al., 2013]. Jhuang et al. [2013] present the JHMDB dataset, which contains RGB videos and labeled skeletons of humans in a subset of HMDB. Hadfield and Bowden [2013] provided the Hollywood3D dataset, which includes both RGB and depth information for commercial 3D movies (by Sony Pictures). In this thesis, however, we will restrict our experiments to multimodal data in controlled 3D environments.

The Inria Xmas Motion Acquisition Sequences (IXMAS) [Weinland et al., 2006] is a multi-view action recognition dataset. The dataset was designed to investigate how to build spatio-temporal models of human actions that could support recognition of simple actions, independent of viewpoint and body shapes. It contains videos of 11 actions, each performed 3 times by 10 actors (See Fig. 2.4). The data were acquired in a lab using 5 standard Firewire cameras. Actors were given no instructions on how to perform an action, and they were free to choose their orientation and position. As ground truth, silhouettes (extracted by a background subtraction algorithm) and reconstructed volumes in MATLAB format are provided by the authors. The

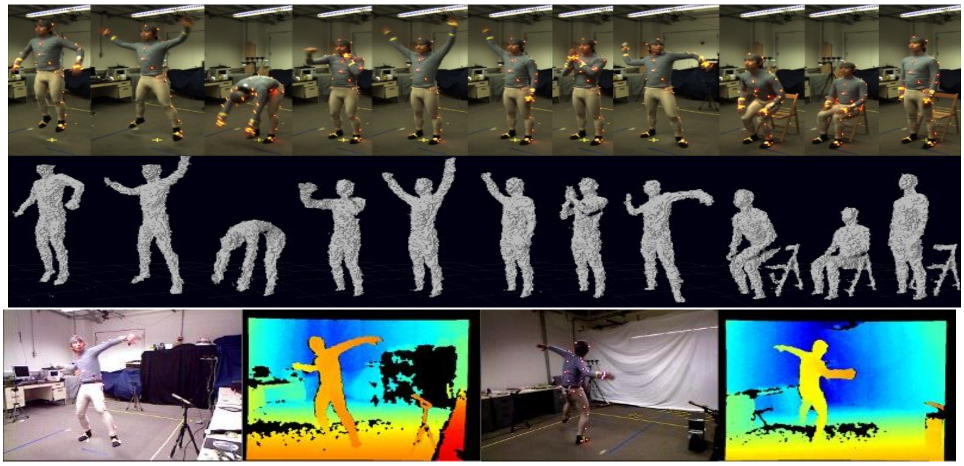


Figure 2.5: Berkeley-MHAD action dataset Ofli et al. [2013]

Multi-camera Human Action Video dataset (MuHavi) Singh et al. [2010] is another large collection of multi-view human action videos recorded through 8 non-synchronized CCTV cameras. There are 17 actions performed by 14 actors. The dataset is proposed to evaluate the robustness of pose-based methods with respect to change in viewpoint. Silhouettes of 14 primitive actions performed by 2 actors, captured from 2 different views, were manually annotated and made publicly available.

Berkeley-MHAD [Ofli et al., 2013] is a multimodal dataset consisting of sequences of 11 actions performed 5 times by each of 12 different subjects for a total of 660 action sequences. These activities are captured by 5 different sensory systems: an optical motion capture system, 2 Microsoft Kinect cameras, 4 multi-view stereo vision camera arrays, 6 wireless accelerometers and 4 microphones. The dataset provides opportunities to (a) evaluate action recognition algorithms for different modalities and (b) to develop fusion algorithms for multimodal action recognition (See Fig. 2.1.3 for examples).

2.2 Feature Representations

The most fundamental task in developing a human action recognition system is the extraction and representation of feature descriptors. An ideal feature representation is mainly characterized by its (a) robustness against (moderate) variations in background, viewpoint, and execution style of actions (b) richness and sufficiency towards classification of actions and (c) efficient extraction. While some representations explicitly take into account the temporal dimension, others extract features from each frame independently. In the latter case, the temporal context is embedded in the classifier itself, e.g. the Hidden Markov Models (HMM).

Existing feature descriptors used for human action recognition can be broadly divided into three categories: global, local, and part-based representations. A global representation describes the underlying motion as a whole. A global descriptor is obtained in a top-down fashion. First, a person is localized through background subtraction. Then global features around the regions of interest (ROI) are extracted. The global representations are not generalizable to unconstrained scenarios as they are more sensitive to viewpoint, noise, occlusion and background clutter. In recent years, local representations that describe visual observation as a collection of independent patches have gained popularity. Determining local representations is a bottom-up process that finds spatio-spatial interest points in a volume, determines some spatial or spatio-temporal features around those points, and then combines all the local features to form the final representation. While local representations are reasonably robust against small variations in viewpoint, position of people, noise, and partial occlusion, they depend on the *informedness* of the interest points. Many depth-map-based action representations are semi-local, since they need effective localization or sampling within large volumes. So their classification as global or local representations may seem arbitrary (See [Ye et al., 2013] for a survey on depth-based motion analysis). Finally, part-based representations are distinguished by their dependency on direct or indirect modeling of different body parts. In the following subsections, we give a brief review and examples of the different representations.

2.2.1 Global Representations

Most earlier techniques of action recognition, aimed at controlled scenarios, used global human representations since background subtraction and segmentation is relatively straightforward in such cases. In fact, global representations perform well in those situations. Below, we categorize them further and give some details.

Silhouettes and Silhouettes based Features

Many global representations extract silhouette of a person as a region of interest and subsequently construct features using the silhouettes. Bobick and Davis [2001] extracted silhouettes for each frame in a video and determined differences between consecutive frames. These difference images were exaggerated to form a single binary motion energy image (MEI) which indicated where motion occurred. They also proposed the motion history image (MHI), which is a real-valued image where intensity is a function of the recency of motion. Efros et al. [2003] construct a motion descriptor by separating the x and y components of optical flow vectors between consecutive frames. Wang et al. [2007] employed the radon transform (R transform) to extract silhouettes to achieve low computational complexity and geometrical invariance (translation and scale). Souvenir and Babbs [2008] incorporated time domain in this representation in order to determine the so-called R-surfaces. [Weinland and Boyer, 2008] presented an exemplar-based embedding approach where the template key frames are represented by person silhouettes.

Other variants consider silhouette's contour as a compact representation of the human pose [Cheema et al., 2011, 2012a, Chen et al., 2006, Baysal et al., 2010, Dedeoglu et al., 2006]. Cheema et al. [2011] and Dedeoglu et al. [2006] used features derived form the distance of points on a contour to its centroid. Baysal et al. [2010] considered manually marked line-pair edge segments on the smoothed contour as object representation. Chen et al. [2006] proposed the star-skeleton approach that connects gross extremities of a human contour to its centroid.

As a natural extension of the binary silhouette, Munoz-Salinas et al. [2008] proposed depth silhouettes in order to incorporate depth information for gesture recognition. Jalal et al. [2011] used R transform of depth silhouettes for action recognition in indoor environments. Ni et al. [2011] proposed 3DMHI, an extension of the motion history image to incorporate depth information. Some approaches, such as [Escalante et al., 2013, Cheema et al., 2013, Ofli et al., 2013], project the Kinect depth data to obtain gray-scale frames and employ methods based on motion energy image. A recent survey on human motion analysis using depth data is given in [Ye et al., 2013].

Space Time Volumes

The global video descriptors discussed above represent a sequence either as a single 2D vector (e.g. MHI and MEI) or as a collection of feature vectors based on features from each frame (e.g. [Dedeoglu et al., 2006, Cheema et al., 2011]). Another popular method is to build 3D spatio-temporal volumes (STV) over the action sequence [Blank et al., 2005, Yilmaz and Shah, 2005, Ke. et al., 2005]. Yilmaz and Shah [2005] treated a sequence of 2D contours as an STV object in (x, y, t) space. The action descriptors, called *action sketches*, were then computed by analyzing the differential geometric properties, such as maxima and minima, of STVs. Blank et al. [2005] first formed an STV by stacking silhouettes of a given sequenceand then extracted local space-time saliency and orientation features. Jiang and Martin [2008] proposed a global descriptor called *shape flow* that represents both the shape and movement of an object in a parsimonious manner. A shape flow is a 3D assembly of flow lines of object contours. Space-time volumes are also common in multi-camera imagery and depth analysis. Weinland et al. [2006] combined silhouettes from multiple view-points to form 3D hulls of humans. Then they generated motion history volumes (MHV) from the sequence of those 3D voxels, followed by the Fourier transform of cylindrical coordinates. Although these features are invariant to location, scale, and rotation, determining MHV requires camera calibration.

Grid-based and Template-based Global Representations

Some global representations apply spatial and/or temporal griding to achieve robustness against noise and partial occlusion. These approaches divide ROI into cells, extract local features in those cells, and combine them to form a global representation. For example, Thurau and Hlavac [2008] proposed a representation, called *pose primitives*, that is based on Histogram of Oriented Gradients (HOG) to better cope with articulated poses and cluttered backgrounds. They decoupled the background appearance from the foreground by means of non-negative matrix factorization. The local temporal context was incorporated by means of n-gram expressions. Ragheb et al. [2008] divided each space-time volume into sub-volumes (STSV) and computed their corresponding Fourier mean-power spectra as the feature vectors. Cheema et al. [2013] used motion histograms in a spatio-temporal grid.

2.2.2 Local Representations

Reliable localization of human and background subtraction is a challenging problem in unconstrained environments. Therefore, local representations that describe the visual observation as a collection of independent patches have become popular in this domain. Compared to global representations, local representations are fairly invariant to changes in viewpoint, appearance of people and partial occlusion. Most local features are computed in three steps: detecting space-time interest points, determining descriptors around those interest points, and Bag-of-Words clustering and quantization. In the following, we give an overview of these processes.

A. Space-Time Interest Points (STIP) Detection

Space-time interest points (STIP) refer to those locations in a volume where sudden changes in movement and appearance occur. A number of STIP detectors have been proposed in last few years [Harris and Stephens, 1988, Laptev and Lindeberg, 2003, Dollar et al., 2005, Willems et al., 2008]. These detectors differ in the way they employ saliency functions. The **Harris3D** detector [Laptev and Lindeberg, 2003] is an extension of the Harris corner detector [Harris and Stephens, 1988] to 3D. It computes a second-moment matrix at each spatio-temporal point in the video using independent spatial and temporal scales. The final interest points are the local maxima of an adaptation of the Harris 2D operator to 3D. Dollar et al. [2005] addressed the issue of the rarity of stable interest points found by the Harris3D detector and proposed the **Cuboid** detector, which employ the 2D spatial Gaussian smoothing kernel and a quadrature pair of 1-D Gabor filters along the time axis. Willems et al. [2008] showed that features can be localized both in the spatio-temporal domain and over both scales simultaneously when using the determinant of the Hessian as a saliency measure. Consequently they proposed the **Hessian** detector that efficiently determines dense scale-invariant spatio-temporal interest points. This efficiency is achieved by using an integral video structure. Wang et al. [2009] conclude that regular **dense sampling** of spatio-temporal interest points outperforms Harris3D, Cuboid, and Hessian detectors for recognizing human actions in realistic settings.

Some recent approaches track the interest points through several frames within a video sequence [Wang et al., 2011, Jiang et al., 2012]. The **dense trajectories** approach [Wang et al., 2011] is the most prominent method. This approach densely samples points from each frame and tracks them (up to a fixed number of frames) based on readily computed dense optical flow fields. Robustness to fast irregular motions as well as shot boundaries is achieved by global smoothness constraints. Wang et al. [2013] computed different features along dense trajectories and showed state-of-the-art performance on action recognition in the wild. Hadfield and Bowden [2013] considered interest point detection in realistic RGBD commercial videos. They proposed 4D (x,y,t,depth) extensions to Harris3D, Hessian3D, and other detectors. Ofli et al. [2013] considered depth-layered multi-channel(DLMC) videos. The represented depth videos as sequences of gray scale frames and employed the Harris3D detector to localize interest points.

B. Local Descriptors

Local descriptors are used to encode actual spatial and motion information within the sub-volume centered on interest points. Similar to detectors, most local descriptors for videos are extensions of image descriptors. Laptev et al. [2008] used histograms of oriented gradient (HOG) and histograms of oriented flow (HOF) descriptors. Willems et al. [2008] computed eSURF, an extended version of the SURF descriptor [Bay et al., 2008], where each cell in the sub-volume is characterized by the weighted sums of Haar wavelets along three axes. Scovanner et al. [2007] proposed an extension of the SIFT [Lowe, 2004] image descriptor to 3D. In this approach, spatio-temporal gradients are computed for each pixel in the cuboid. Kläser et al. [2008] computed HOG3D for a given 3D patch using integral videos. Their approach is very similar to 3D SIFT except that HOG3D bins the 3D gradients into regular polyhedrons. The evaluation of different local descriptors in [Wang et al., 2009] shows that HOG/HOF – a combination of image gradient and flow information – outperformed HOG3D, HOG, HOF, eSURF, and Cuboid features.

The descriptors that encode optical flow in unconstrained videos are prone to different artifacts, such as camera motion and background noise. Recent methods attempt to overcome this deficiency in order to exploit the discriminative information in the motion [Dalal et al., 2006, Kliper-Gross et al., 2012, Wang et al., 2011]. Kliper-Gross et al. [2012] presented **Motion Interchange Patterns**, an encoding that captures at every time point and at every image location both the preceding motion flow and the next motion element. They also used a suppression mechanism to decouple the shape from the motion, and to compensate for camera motion in a manner tailored for the encoding scheme. They employed the standard Bag-of-Words approach to achieve high recognition on HMDB and UCF50 data. Dalal et al. [2006] proposed motion boundary histogram (MBH) descriptors for human detection that are based on the gradients of the optical flow field. Spatial derivatives are computed for horizontal and vertical components of optical flow and, similar to HOG, the orientation information is quantized into histograms. MBH along dense trajectories has shown excellent results for action recognition [Wang et al., 2011, 2013].

Most depth-based features are determined in local patches. Li et al. [2010] proposed an action representation based on bag-of-3D-points from depth maps. The sparse points are selected through a simple, yet effective projection-based sampling scheme that relies on binary silhouettes and contours across different directions. [Vieira et al., 2012] represented a depth sequence in a 4D space-time grid. They used a saturation scheme to enhance the roles of the sparse cells that typically consist of points on the silhouettes or moving parts of the body. Some approaches, such as [Escalante et al., 2013, Cheema et al., 2013, Ofli et al., 2013], project the Kinect depth data to obtain 2D gray scale frames and employ feature extraction methods based on motion energy images.

C. Bag-of-Features (BoF) Representation Scheme

The bag-of-features method, motivated by the Bag-of-Words (BoW) approach in document classification, represents an image or a video as an orderless collection of features. Given a collection of features collected from the training data, the approach employs the following steps:

- **Building a Vocabulary**: Given a (large) set of N features from the training data, a clustering algorithm, e.g. k-means, k-mediods or the Gaussian mixture model is applied to extract k number of *words* (the cluster centroids). This set of words is referred to as vocabulary or the code book.

- **Quantization of Descriptors to Code Book**: For each data sample, a K-bin histogram is maintained where each bin corresponds to a unique word in the code book. In the hard quantization approach, the count of the most similar word is incremented for every individual (local) feature. The soft quantization approach considers the distance to all centroids and assigns relative weights to the corresponding bins.
- **Normalizing the Histograms**: Finally, the histograms are normalized e.g. using L_1 or L_2 norms.

The BoF thus transforms samples of different dimensions (number of frames, resolution) to a normalized vector of length K. Note that the BoF quantization is orderless and discards spatio-temporal localization information about local features. This orderless nature provides a great flexibility and robustness with respect to noise, drift, and partial occlusions. While the BoF (or any sparse coding scheme) is inevitable for local representations, it can also be used for global ones. For example, Chaaraoui et al. [2012] and Cheema et al. [2011] proposed approaches based on the bag-of-keyposes. In order to retain some spatio-temporal information in feature descriptors, most local features based approaches divide videos into spatio-temporal grids at different scales and apply BoF on each grid. The final representation is either based on concatenation of individual BoF vectors (e.g. [Cheema et al., 2013]) or aggregating the kernel matrices (e.g. [Wang et al., 2013]).

2.2.3 Part-based Representations

Part-based features rely on direct or indirect modeling of different body parts. Such representations have shown significant success in human action recognition in still images [Felzenszwalb et al., 2008, Bourdev and Malik, 2009, Yang et al., 2010]. In the presence of 3D depth and skeleton data for videos, action recognition by modeling the motion of different joints and body parts is intuitive. In fact, most prominent approaches to 3D action recognition are based on part-based modeling. For example, Wang et al. [2012] modeled depth appearance in proximity of 3D joints as local occupancy patterns (LOP). Different subsets of joints were combined to form the so-called *actionlets* descriptors.

Several researchers have recently shown the effectiveness of pose-based methods in human action recognition in simple RGB videos [Singh and Nevatia, 2011, Tran et al., 2011]. Tran et al. [2011] proposed an action representation described by a combination of human body-part movements corresponding to a particular action. They also proposed a computationally efficient algorithm capable of discriminating the key differences in movement of each body-part pertaining to a particular action. Yao et al. [2012] presented a system for coupling the closely related tasks of action recognition and pose estimation. Evaluation of their approach on TUM multi-view kitchen dataset [Tenorth et al., 2009] indicates the mutual gain by such coupling. However, Rohrbach et al. [2012] argues that current pose estimation approaches are not good enough to classify fine grained activities due to low inter-class variations. Recently, Jhuang et al. [2013] used bounding boxes and manually labeling of joints to action recognition in the wild. They advocated the idea that action recognition can be improved by focusing on person-specific areas (e.g. extracting MBHs only around person's contours). They also show that current pose estimation algorithms are not reliable for unconstrained videos. Sadanand and Corso [2012] proposed the use of action bank templates to determine video features. Their feature extraction is based on spotting different motion templates in the multiple-scale spatio-temporal cuboids. Action bank features outperform HOG/HOF around Harris 3D corner points for several realistic datasets.

2.3 Classification of Human Activities

After feature representation, the next step in human activity recognition is building a classification model to classify unlabeled query instances. Given the labeled training data $\{\mathbf{X}_{train}, \mathbf{y}_{train}\}$ and the query instance $\mathbf{x}_q$, a classifier is invoked to determine the most suitable label y, i.e. which maximizes $P(y|\mathbf{x}_q, \mathbf{X}_{train}, \mathbf{y}_{train})$. Various classification approaches handle this problem differently. We differentiate these approaches according to classical machine learning aspects in: instance based classification (e.g. kNN), discriminative methods (e.g. SVM), and generative methods (e.g. HMM) and and list them in the following subsections.

2.3.1 Instance-based Classification

Instance-based classification methods classify a given instance based on its similarity (in feature space) to other labeled instances. Often they do not build an explicit model during training and instead exploit geometrical properties of the data at classification time. Instance-based methods differ from each other with respect to how they represent neighborhood surfaces for classification. Nearest Neighbors (NN) classification is a simple instance based approach that assigns a query instance the label of the instance that is most similar to it. Often L_1 or L_2 norms are used to measure the similarity. kNN considers k nearest neighbors and uses a majority or weighted voting scheme to assign final class labels. Other methods such as Nearest Affine Hull (NAH), Nearest Convex Hull (NCH) and Nearest Hyperdisk (NHD) determine the label of a sample based on its distance to the geometrical surfaces spanned by each category. These approaches may be of great importance for large-scale online classification of human activities. We will give more details of these methods in Chapter 5 and point out their significance for large-scale activity recognition.

NN classification can be performed at frame level or at the video level. In [Efros et al., 2003], a nearest neighbor approach was employed on optical flow frames. Blank et al. [2005] also classified videos with silhouette-based global spatio-temporal volumes through NN. Weinland et al. [2006] used PCA to reduce the dimensionality of motion history volumes. They considered NN with the Mahalanobis distance between query instances and classes to take into the account the variance of each dimension. Bobick and Davis [2001] also used Mahalanobis distance for action recognition using MHI and MEI. Nearest Neighbor classification is also common among approaches that represent actions by a set of keyposes or templates. Carlsson and Sullivan [2001] recognized *forehand* and *backhand* tennis strokes in videos by computing an edge-based distance metric between candidate frames and manually chosen key frames. Thurau and Hlavac [2008] represented a video sequence as a normalized histogram of pose primitives. They used Kullback-Leiber(KL) divergence for histogram similarity and 1NN for action classification. Cheema et al. [2011] used 1NN at frame level and employed a discriminative weighted voting using a bag-of-keyposes approach. Cheema et al. [2012b] demonstrated state-of-the-art performance on gait recognition data by using NN classification for set key poses that represent a person' gait.

2.3.2 Generative Models

Generative models build a joint probability distribution $P(\mathbf{X}, \mathbf{Y})$ on the training data and determine $P(y|\mathbf{x}_q)$ by using Bayes'rule. Typical generative models are Naive Bayes and Hidden Markov Model (HMM). HMMs are popular state-space models which assume that (i) a transition to a hidden state is only possible from its previous state and (ii) an observation is only dependent on the current state. HMMs have been of great significance in classical literature on action recognition (mostly in controlled environments). Yamato et al. [1992], in one of the earliest work on action recognition, trained HMMs to model time-sequential images of different tennis strokes. Each action category is represented by an HMM and classification of a query sequence is made simply by determining the best matching HMM. Although the standard HMMs have been successfully employed for simple action recognition tasks (e.g. [Ivanov and Bobick, 2000, Yamato et al., 1992], they are not suitable for modeling complex activities that have large state- and observation-spaces. To this end, many variations of HMMs have been proposed. For example, Oliver et al. [2002] proposed layered hidden Markov models (LHMMs), where the bottom layer HMMs recognize atomic actions and the upper layer HMMs treat these atomic actions as observed states. Another example is [Ikizler-Cinbis and Forsyth, 2008] where authors constructed individual HMMs for 3D trajectories of different body parts such as legs and arms. Then the single action HMMs are joined together to form activity HMMs by linking the states that have similar emission probabilities.

While HMM and its variants are restricted by their model structure, a Dynamic Bayesian Network (DBN) – as an alternative for HMM – provides a more flexible structure by representing the hidden (and observed) states in terms of state variables, which may have complex inter-dependencies. Du et al. [2007] proposed Coupled Hierarchical Duration-State DBN (CHDS-DBN) which represents human motions in videos at two scales: the global activity state scale and the local activity state scale. Despite their success in tracking and classifying human actions in controlled scenarios, only a few generative models have been proposed for action recognition in the wild. Todorovic [2012] modeled complex activities by a generative model-graph, where nodes correspond to the primitives, and the graph's adjacency matrix encodes their affinities for probabilistic grouping into observable video features.

Generative models are very expressive as they maintain a joint probability distribution over the $(\mathbf{X}, \mathbf{Y})$ space, thus providing better models for the missing data or holes in the input space. Therefore, they are widely used in certain problems such as tracking. However for *classification* problems, such as human action recognition in segmented unconstrained videos, discriminative models (discussed in the next sub-section) are preferred for their simplicity and comparatively less parametrization.

2.3.3 Discriminative Classifiers

Discriminative classifiers focus on developing a model $\mathcal{M}$ for separating two or more classes. Unlike generative classifiers, they do not build a joint probability distribution of input and output. Instead, they classify an unlabeled instance $\mathbf{x}_q$ directly by using $P(y|\mathbf{x}_q, \mathcal{M})$. Unlike instance-based classifiers, they do not rely on explicit utilization of the labeled data (or class models) at classification time. Among the most popular discriminative methods are Support Vector Machines (SVM), Linear Discriminant Analysis(LDA), Conditional Random Fields(CRF), Random Forests(RF), and Artificial Neural Networks(ANN).

A Support Vector Machine (SVM) [Cortes and Vapnik, 1995] builds a separating hyperplane between two classes that has the largest distance to the nearest training data point of any class. Boser et al. [1992] suggested the *kernel trick*, a way to create nonlinear classifiers by replacing every dot product with a nonlinear kernel function. The kernel trick allows algorithms to fit models in (possibly high dimensional) transformed feature space. SVMs with different kernels are the most popular choice for human action recognition in complex scenarios [Sadanand and Corso, 2012, Wang et al., 2013]. Most of these methods extract local features, adopt bag-of-features representation and then apply SVM with a suitable kernel. For example, Kuehne et al. [2011] extracted HOG/HOF features around Harris corner points, represented them through BoF approach with 4,000 visual words, and used SVM with Gaussian kernels to give a baseline performance on HMDB dataset. Sadanand and Corso [2012] extracted action bank template features with dimensionality of nearly 15,000 and applied linear SVM to show improved results on HMDB and UCF50 datasets. Kliper-Gross et al. [2012] further improved these results by using dense low-level Motion Interchange Pattern features with BoF and Linear-SVM. Wang et al. [2013] extracted 5 different types of features (HOG, HOF, MBH along horizontal and vertical axis, and path) around dense motion trajectories on 6 different spatio-temporal divisions (grid schemes) – ending in 120 channels of features. Along each channel, they applied a BoF quantization with 4,000 words and used a multichannel SVM with χ^2 kernel. This representation and its extensions have shown state-of-the-art performance on large-scale unconstrained video datasets [5].

Conditional Random Fields (CRF) are discriminative state-space models that, unlike their generative counterpart HMMs, do not assume independence among observations in time. Instead, CRF models can take into account multiple overlapping observations on different time scales and model conditional probability of labels on the sequences of observations. Sminchisescu et al. [2006] presented an approach to recognize human motion in monocular video sequences, based on discriminative conditional random fields (CRFs) and maximum entropy Markov models (MEMMs). Their approach outperforms HMM. Natarajan and Nevatia [2008] proposed Shape, Flow, Duration-Conditional Random Field (SFD-CRF), which computes its observations potentials using shape similarity, and transition potentials using optical flow. Deep neural network based models such as convolutional neural networks (CNN) are also used for automatic feature construction for human actions. Unlike handcrafted (handpicked) features, these models can directly act on raw data and determine discriminative features which are then used in BoF scheme. For example, Le et al. [2011] learned hierarchical features from unconstrained videos using independent subspace analysis and convolutional networks. In [Shuiwang et al., 2013], a 3D CNN model is developed for feature construction in real world surveillance. It extracts features from both spatial and temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. BoF and SVM are then applied for classification.

Ensembles or Boosting of different classifiers is also a popular approach for discriminative classification. An ensemble classifier builds on a set of indi-

[5] http://crcv.ucf.edu/ICCV13-Action-Workshop/results.html

vidual diverse hypotheses (classifiers) where diversity is achieved by acting upon different data, applying different algorithms, or choosing different (hyper-)parameters. Popular ensemble classifiers are are Random Forests [Breiman, 2001] and Adaboost. Gall et al. [2011] proposed Hough forests for object detection, tracking, and action recognition. A Hough forest consists of a set of random trees that are trained to learn a mapping from labeled local cuboid features to their corresponding votes in a Hough space where the the Hough space encodes the hypotheses for action in scale(time)-space and class. The local voting mechanism leads to robustness towards occlusion and noise. However, most Hough based approaches require sufficient training data (labeled patches) to enable discriminative voting. Yu et al. [2012] proposed propagative Hough voting where the feature voting is performed using random projection trees (RPT), which leverages the low-dimension manifold structure to match feature points in the high-dimensional feature space. Wang et al. [2012] represented a 3D action as an actionlet ensemble which is a linear combination of the actionlet features. The discriminative weights of actionlets are learned via a multiple-kernel learning method.

Summary

We have discussed vision-based action recognition approaches and provided a brief overview of the domain. We explained different application areas and some related benchmark video datasets. Then we described well-known feature representations and their capacity to capture the dynamics of data of different natures. In particular, we see that the global features that take a top-down approach to model a frame or sequence are suitable to scenarios where person localization is simple and the amount of background noise is minimal. Whereas local representations, which are based on local features in salient spatio-temporal locations, are a natural choice for unconstrained scenario due to their robustness towards occlusion, scale, and noise. Finally, we discussed different classification methods in the context of action recognition. This chapter, thus, provided the big picture and an overview of action recognition in videos. For comprehensive surveys on different aspects, readers can look into recently published reviews and surveys [Aggarwal and Ryoo, 2011, Poppe, 2010, Ke et al., 2013, Chaquet et al., 2013, Jiang et al., 2013].

Chapter 3

Efficient Action Recognition by Learning Discriminative Key Poses

Recognizing human actions in a controlled environment has received significant attention by researchers in the last decade. Most early work focused on modeling human motion through global or local features. Recently, there has been growing interest in exploiting the pose information for action recognition as pose-based methods offer robustness against missing-frames, low-observational latency, and online classification [Weinland and Boyer, 2008, Thurau and Hlavac, 2008, Baysal et al., 2010]. In this chapter, we present an efficient approach to pose-based human action recognition in a controlled environment. Given a set of training videos, we first extract the contour distance signal (CDS), a scale invariant pose-based feature, from silhouettes. Then, we cluster the features in order to build a class-specific set of prototypical key poses. Based on their relative discriminative power for action recognition on the training data, we learn the weights that favor distinctive key poses. Finally, the classification of a query sequence is based on a simple and efficient weighted voting scheme that augments results with a confidence value which indicates recognition uncertainty. Our approach does not require temporal information and is applicable to action recognition from videos or still images. It is efficient and delivers real-time performance. In experimental evaluations, our approach shows high recognition accuracy for the single-view Weizmann dataset [Blank et al., 2005] as well as the multi-view MuHAVi dataset [Singh et al., 2010]. Consequently, a number of recent approaches have adopted or extended the ideas of CDS and discriminative weighting of key poses to action recognition in 3D data as well as in unconstrained videos [Chaaraoui et al., 2012, Liu et al., 2013, Climent-Prez et al., 2013, Zanfir et al., 2013].

3.1 Introduction

In Chapter 2, we categorized human activities, based on the complexity of motion dynamics, into three broad classes: *gestures* such as “smiling” or “rotating head”, *(primitive) actions* such as “walking” or “turning back”, and *(complex) activities* such as “cooking” or “playing cricket”. In this chapter, we focus on primitive actions that, when properly combined or sequenced (or put into context), could be used to explain more complex activities. In particular, we aim at recognizing actions that can be discriminated based on their pose. Most existing research on action recognition relies on temporal cues. Many methods directly use motion cues [Fathi and Mori, 2008, Cutler and Davis, 2000] or spatio-temporal features [Bobick and Davis, 2001, Blank et al., 2005, Roth et al., 2009, Ke. et al., 2005]. Other methods track interest points or local patches [Laptev and Lindeberg, 2003, Niebles et al., 2006] or use probabilistic models (e.g. *n*-grams or HMMs) to implicitly represent temporal contexts [Thurau and Hlavac, 2008, Martnez-Contreras et al., 2009, Wang and Suter, 2007].

Although motion information obviously plays an important role in action recognition, many human activities, such as “standing”, “running”, “reading a book”, or “playing football”, can be recognized from a single image or snapshot, assuming that the given pose is sufficiently distinctive or that enough context information is made available. Furthermore, in monocular videos where human localization and background removal can be efficiently achieved, pose information can be of great importance in building robust action recognition systems. Interestingly, only a few pose-based methods have been introduced so far that work equally well for action recognition from videos and still images. A common idea of these approaches is to represent and classify human poses for each image or frame in a sequence [Dedeoglu et al., 2006, Weinland and Boyer, 2008, Baysal et al., 2010, Thurau and Hlavac, 2008]. Examples of common pose representations include raw silhouettes [Weinland and Boyer, 2008], line pairs [Baysal et al., 2010], histogram of oriented gradient (HOG) descriptors [Thurau, 2007], and contour-HOG descriptors [Dedeoglu et al., 2006]. An action class is then normally represented as a histogram over a set of *key poses*, i.e. a representative pose of a complex action, or simply as a concatenation of pose representations.

We presented a novel, non-temporal, pose-based method for action recognition from videos and still images. In contrast to previous work, we apply a scale-invariant contour feature CDS for pose representation that can be efficiently computed from a silhouette image. For the representation of action classes, we employ the idea of key poses. However, in addition to previous work, we rate the class-specific key poses according to their discriminative power. For instance, key poses involved in a "turn back" action will include poses representing states of "walking", "standing", and "turning-head", among others. Since key poses such as "standing" may be shared among different actions (e.g. "walk" and "guard"), we apply statistical learning to determine the relative importance of key poses. Furthermore, the relative importance weights allow us to assign confidence values to classification results. As the proposed approach does not use any temporal information, it is suitable both for video and image-based action recognition. By benchmarking on single- as well as multi-view action datasets, we demonstrate that our approach favorably deals with variations in view or distance.

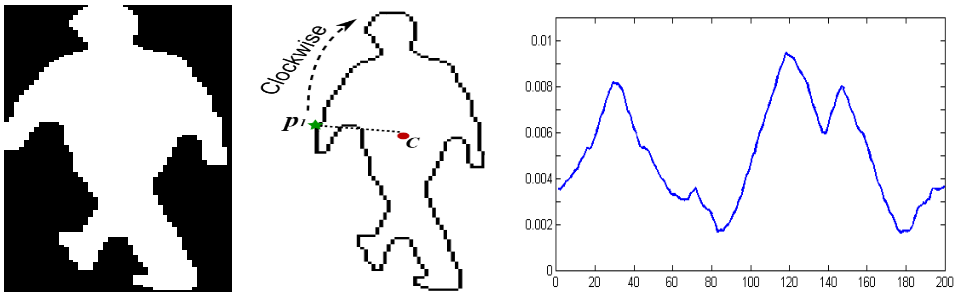
The technical contribution of our work is twofold: (i) a novel combination of a contour-based pose representation and non-temporal key pose based classification, and (ii) a novel weighting scheme for rating the relative importance of key poses. Also, unlike various other approaches [Weinland and Boyer, 2008, Ali et al., 2007], our approach does not require any subsampling, upsampling or trimming during training. Furthermore, we have no limitations with respect to the length of the considered video sequence, and the approach performs in real-time on any standard desktop computer or smartphone. This real-time capability is of crucial importance in intended applications which aim at pose-based recognition of actions in interactive environments.

The rest of this chapter is organized as follows: Section 3.2 reviews related work. Section 3.3 provides details on the underlying contour-based feature. Details on the leading of discriminative key poses are given in Section 3.4. Section 3.5 reports on our benchmark data, experiments, and results. Finally, Section 3.6 concludes the chapter and discusses the future applications.

3.2 Related Work

The idea of using key poses for action recognition has been applied successfully in previous work. In an early work, Carlsson and Sullivan [2001] used key-frame templates for action recognition. They recognized *forehand* and *backhand* tennis strokes in videos by computing an edge-based distance metric between candidate frames and manually chosen key frames. Recently, Kilner et al. [2009] used key poses to analyze 3D data in a multi-camera sports environment. However, their approach is not applicable if only one view is considered at a time. Thurau and Hlavac [2008] proposed a mutual information (MI) based weighting scheme for histograms of key poses. In contrast to their work, we directly adapt the weighting of each key pose and use a different weighting scheme. In [Weinland and Boyer, 2008], an exemplar-based embedding approach was presented which does not use any motion information and the key frames are determined by forward feature selection. The training data is then mapped to a distance space based on key frames. However, this process is computationally demanding, especially when applied without subsampling on large, multi-view and multi-actor datasets. Our approach differs in two ways. Firstly, we use cluster centroids as the representative key poses for each action class. Secondly, we model the inter-class and intra-class variations by efficiently learning weights for key poses.

Our approach is most similar to [Baysal et al., 2010] that determines discriminative key poses using k-mediods and a ranking scheme over their potential score towards discriminating actions. We, instead, propose the use of an intuitive, weighted voting scheme for classification. Also we advocate using a contour-based feature which is more informative and systematic than the manually marked *line-pair* edge segments considered in [Baysal et al., 2010]. Our feature extraction is based on method by Dedeoglu et al. [2006] who defined a distance signal over object contours. For action representation, however ,they use template histograms of key poses in a temporal context. We, on the other hand, avoid histograms or any other temporal model. This enables our action recognition approach to be applicable for both image and video datasets.



(a) (b) (c)

Figure 3.1: Extraction of the CDS feature: (a) original silhouette (b) the contour (c) CDS with $s = 200$.

3.3 Contour-based Pose Representation

Extraction of efficient and informative features is crucial for success of human activity recognition. Most pose-based approaches use global human representations that extract a binary silhouette of a person as a region of interest in each frame and subsequently construct features using those silhouettes [Carlsson and Sullivan, 2001, Niu and Abdel-Mottaleb, 2005, Martnez-Contreras et al., 2009, Wang and Suter, 2007, Weinland and Boyer, 2008]. Since our focus in this chapter is on learning discriminative key poses, we assume silhouettes images to be available, which is indeed the case for many well-known benchmark datasets. Given a human silhouette, we extract its contour and transform it into a *contour distance signal (CDS)* as in [Dedeoglu et al., 2006]. Details of this representation are described below.

For a given binary silhouette image H consisting of n pixels, we determine its center of mass $C = (x_c, y_c)$ where

$$x_c = \frac{\sum_{i=1}^{n} x_i}{n}, y_c = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3.1}$$

and n is the number of silhouette pixels.

Let $P = [p_1, p_2, ..., p_n]$ be the ordered set of contour points such that p_1 corresponds to the extreme left point on H and successive p_i are listed in a clockwise fashion (see Fig. 3.1). A distance vector $\mathbf{d} = [d_1, d_2, ..., d_n]$ is formed by calculating the Euclidean distance between p_i and C, i.e.

$$d_i = \|p_i - C\|\,, \qquad \forall i \in [1, 2, ..., n] \tag{3.2}$$

In order to provide robustness against varying image sizes and shapes, $\mathbf{d}$ is scaled to a vector $\widehat{D}$ of constant size s such that

$$\widehat{D[i]} = d\left\lceil \frac{i * n}{s} \right\rceil, \qquad \forall i \in [1, 2, ..., s] \tag{3.3}$$

where $\lceil \cdot \rceil$ is the ceiling function.

Finally, the scaled distance vector $\widehat{\mathbf{D}}$ is normalized to have a unit sum:

$$\overline{D[i]} = \frac{\widehat{D[i]}}{\sum_1^s \widehat{D[i]}} \tag{3.4}$$

In the rest of this chapter, $\overline{D}$ is referred to as the contour distance signal (CDS). Obviously, this contour-based feature is scale-invariant and can be efficiently extracted from silhouettes. Compared to the size of the original image, the size of the CDS is much smaller. For example, in the MuHAVi dataset, the resolution of the original silhouette images is 720 × 576 pixels, whereas the contours of the silhouettes consist of only a few hundred pixels. The CDS can be further scaled down if $s < n$. This implicit dimensionality reduction through transforming the silhouette to the CDS ultimately enables efficient learning and classification.

3.4 Learning Discriminative Key Poses

In this section, we describe how we determine discriminative key poses for different actions. Two major steps in the underlying process are (a) extraction of key poses and (b) learning the appropriate weights for those key poses. Our approach to key-pose extraction builds on [Baysal et al., 2010], where key poses are learned over a space of line-pair segments. A significant characteristic of our approach is its ability to *adapt to* and *exploit* the importance of key poses. Figure 3.2 summarizes the computational steps involved in the proposed framework.

Given a set of labeled video sequences or still images, silhouettes are extracted

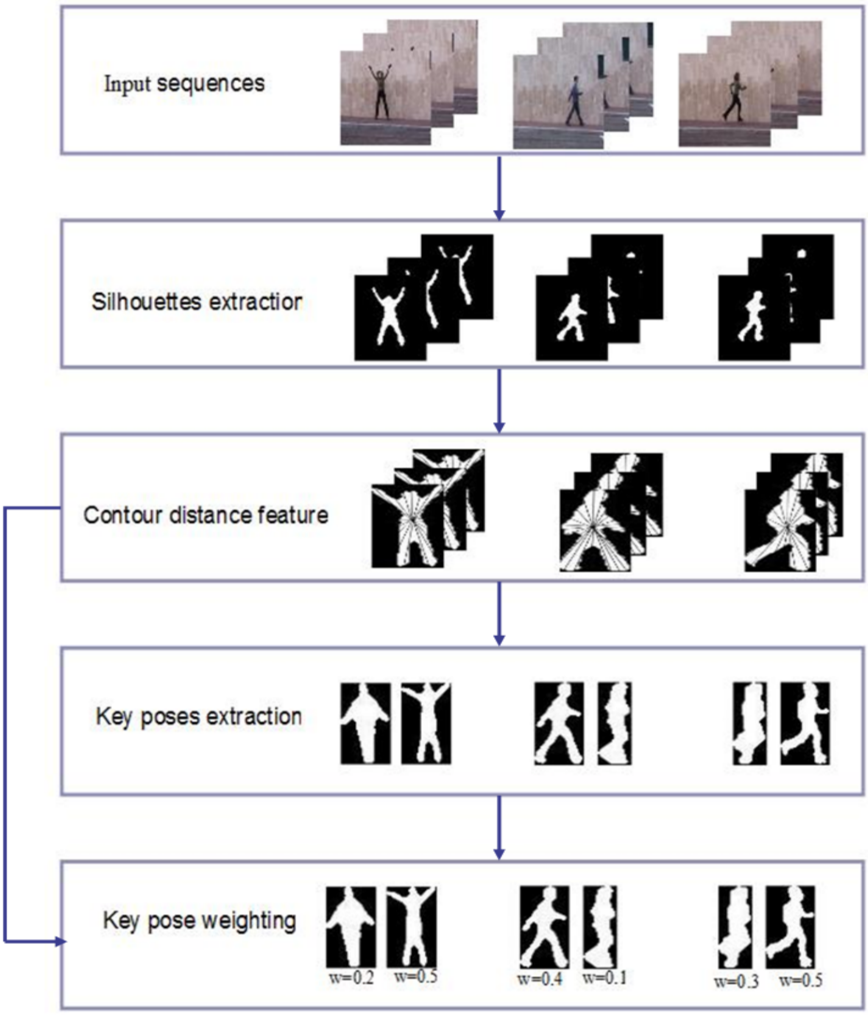


Figure 3.2: Overview of our discriminative key pose approach

for all frames through a background subtraction method, which can be done reliably in controlled videos. For several benchmark datasets, including those considered in this chapter, binary silhouettes are readily available, which permits us to focus on the problem of pose-based action recognition. Given an extracted silhouette, each input frame is mapped to the normalized contour distance signal $\overline{D}$ of size s (as described in the previous section). The granularity of the resulting feature may be controlled by setting the value of s – a free parameter of the distance transform.

Determining action-specific discriminative key poses from the available training data consists of two successive steps. In the first stage, representative key

Algorithm 1 Learning discriminative key poses

```
Input: Silhouettes for all input frames of all training videos
Output: Key poses and their weights
Let $k$ represents the number of clusters,
$A = \{a_1, a_2, ...a_r\}$ be the set of actions,
$p_{ij}$ denotes $j - th$ key pose of action $i$ and $w_{ij}$ be the its weight
 1: for all action $a \in A$ do
 2:     Cluster all frames into k groups using k-means
 3:     Take cluster centers as key poses thus ending up with $r \times k$ key poses
 4: end for
 5: for all actions $a \in A$ do
 6:     for all frames $f \in a$ do
 7:         Assign the key pose $p_{ij}$ to $f$ such that $\|f - p_{ij}\|$ is minimum
 8:     end for
 9: end for
10: Let $n_{ij}$ and $n_{ij}^{'}$ respectively denotes number of within-class assignments
    and number of out-of-class assignments to $p_{ij}$
11: $w_{ij} := \frac{n_{ij}}{n_{ij} + n_{ij}^{'}} \forall i, j$
```

poses are determined for each action by clustering all frames belonging to the corresponding class. In the second stage, weights are assigned to these key poses according to their ability to discriminate among different actions in the training data. Algorithm 1 formally summarizes the procedure.

Lines 1 – 4 corresponds to key-pose extraction stage. We employ $k - means$ clustering with Euclidean distances to calculate key poses for each action. Notice that we determine key poses for each action category separately. In our early experiments, we observed that using global key poses undermines classification performance. This could be explained by the very nature of the clustering algorithms as they would retrieve more centroids from the classes involving high pose variation (e.g. “kick“) as compared to the classes involving low pose variation (e.g. “walk“). As our model is strictly non-temporal, we do not create any ordered histogram of key poses (KPs). Thus, key poses represent a *set* of different possible states of an action. For example, key poses for the action “kick” may correspond to spatial states such as “standing”, “arm adjustment”, or “pulling the leg”. Figure 3.3 gives an example of 8 key poses extracted from a video of the action KickRight in the MuHAVi

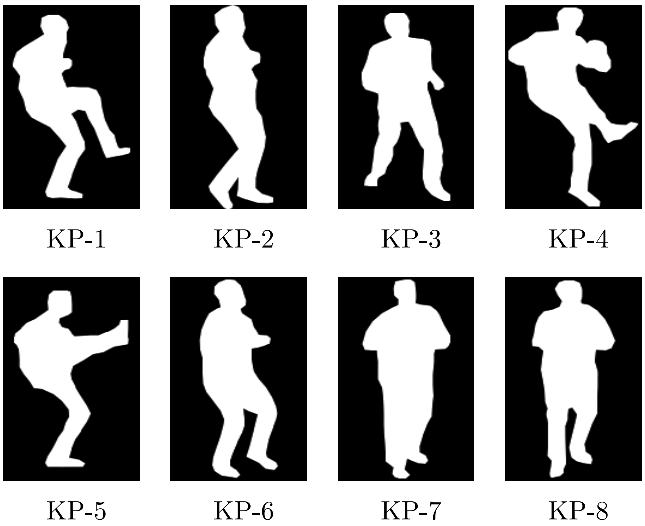


Figure 3.3: Different key poses for action KickRight in MuHAVi dataset

dataset. Notice that KP-2 and KP-6 through KP-8 do not seem to represent distinctive states of the action. Instead, they look more related to actions such as “walk”, ”punch”, or “guard”. Still, they are automatically extracted since they apparently represent significant parts of the action sequence.

The problem of common or ambiguous key poses is resolved by adopting a simple and intuitive mechanism of assigning rewards and penalties to key poses (Lines 5 – 11 of Algorithm 1). This procedure computes relative importance weights of key poses for different actions. For each frame f in the training data, the closest key pose p_{ij} is determined by taking into account all key poses of every action. Each time a key pose p_{ij} is favored by some frame, the actual label of that frame is compared with the action class i of the favored key pose. Thus for each key pose p_{ij}, two values n_{ij} and $n_{ij}^{'}$ are stored, where the former denotes the number of correct classifications to the key pose and the latter denotes number of false assignments to the key pose. In this way, those key poses which frequently match to the frames from other classes, will have lower weights (Line 11). On the other hand, key poses which appear only within one class will get higher weights. From the perspective of key poses: if a key pose corresponds to frames from different action classes, it will have some false assignments which would decrease its weight. From the perspective of action classes: key poses which are common only *within* the action class and are discriminative with respect to other action classes will have higher weights.

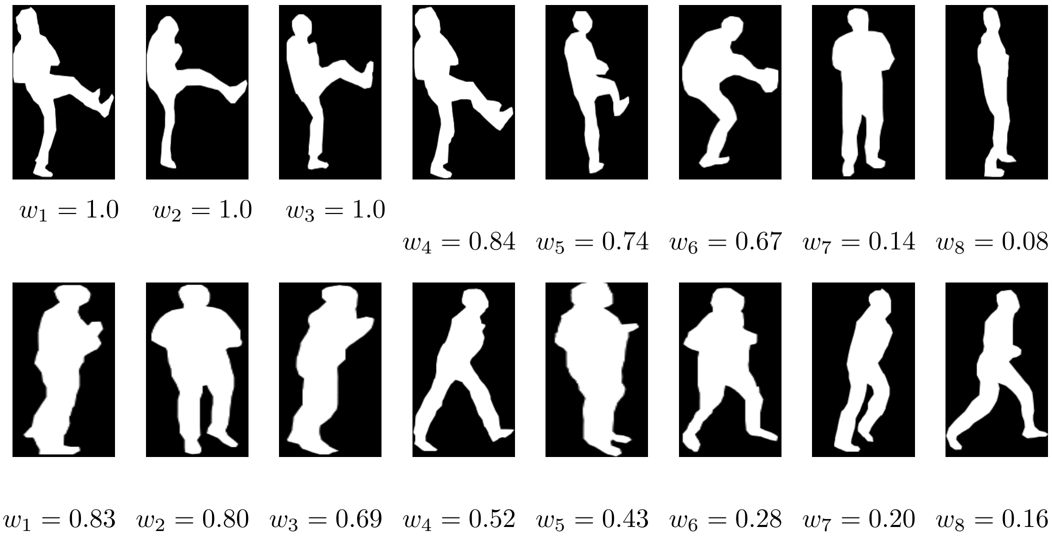


Figure 3.4: The 8 high-ranking key poses and their weights for the two *overlapping* actions ***KickRight*****(first-row)** and *GuardToKick*(second-row) in the MuHAVi dataset. The most distinctive and representative key poses have higher weights. Key poses corresponding to overlapping states have different relative importance e.g. w_7 of KickRight and w_2 of GuardToKick.

This mechanism also allows for automatically eliminating effects of overlapping actions. For instance, KickRight and GuardToKick in the MuHAVi datasets are two such actions for they share many common states, such as "standing straight" or "standing in a punching position". Figure 3.4 shows the 8 top ranking key poses and their weights as determined by our approach. Notice that (a) the larger weights are assigned to more discriminative key poses (b) the poses corresponding to overlapping states (such as "standing in punching position" depicted by the 7th key pose of KickRight and the 2nd key pose of GuardToKick) have very different importance for the two actions. This indicates the ability of our approach to learn the relative importance of key poses.

In the application phase, in order to classify a given frame sequence, we first extract its contour feature. Then we determine the classes and the weights of the closest key poses, with respect to Euclidean distance, for each query frame. Based on these weights, we apply a simple weighted voting scheme. Weights are accumulated for all related key poses, and the label of the action class which has highest sum of weights is chosen. Notice that more discriminative poses dominate this process. In contrast to approaches such as [Thurau and

Hlavac, 2008, Baysal et al., 2010], this allows all query frames and all key poses to participate in the classification process. Due to a compact and non-temporal feature representation and a moderate number of key poses, we thus achieve real-time classification.

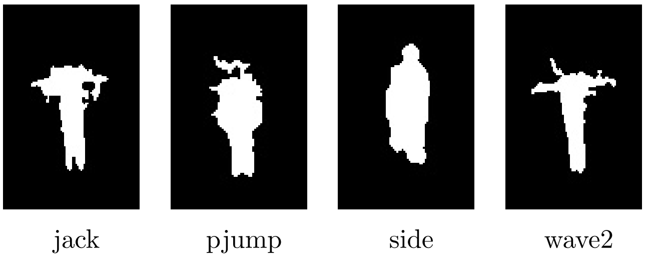
3.5 Experiments

To evaluate the robustness and effectiveness of our approach, we performed experiments on two well-known controlled datasets, namely the Weizmann collection [Blank et al., 2005] and the MuHAVi set [Singh et al., 2010]. In comparison to single-view Weizmann data, MuHAVi is a versatile multi-view action dataset with more primitive action classes. All experiments presented in this section were carried out on a standard notebook computer using MATLAB 7. The average processing rate was measured to be **56 frames per second**, indicating the real-time applicability of the approach. In the following, we elaborate on the two datasets, our experimental results, and their significance in comparison to state-of-the-art methods.

3.5.1 Results on Weizmann Dataset

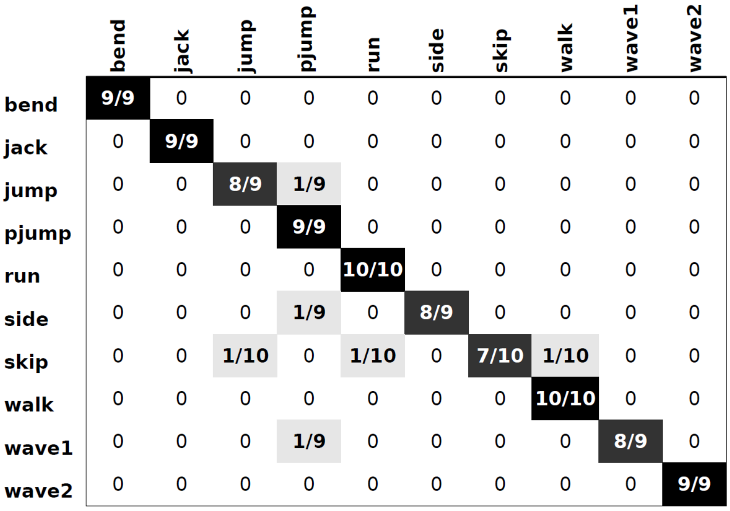
The Weizmann data [Blank et al., 2005] is a popular single-view action dataset which contains 93 video samples for 10 different actions performed by 9 actors. These actions include: walking, running, jumping, galloping sideways (side), bending, one-hand waving (wave1), two-hands waving (wave2), jumping in place (pjump), jumping jack, and skipping. A common tradition is to consider only 9 actions by eliminating the samples of the action "skip" or otherwise excluding some noisy samples. Here, we consider readily available silhouettes for the all the 93 videos in Weizmann dataset. It is worth noting that many of these silhouettes are very noisy (e.g. see Figure 3.5). We use them *as they are* without any preprocessing.

To judge the performance of our approach on this dataset, we applied leave-one-out cross validation. We used 200 points on the contour. The best performance was achieved for 40 key poses per action. Figure 3.6 shows the resulting confusing matrix. Notice that among the 6 misclassified instances, 3 belong to a single class: "skip“.



jack pjump side wave2

Figure 3.5: Examples of some noisy silhouettes in Weizmann dataset



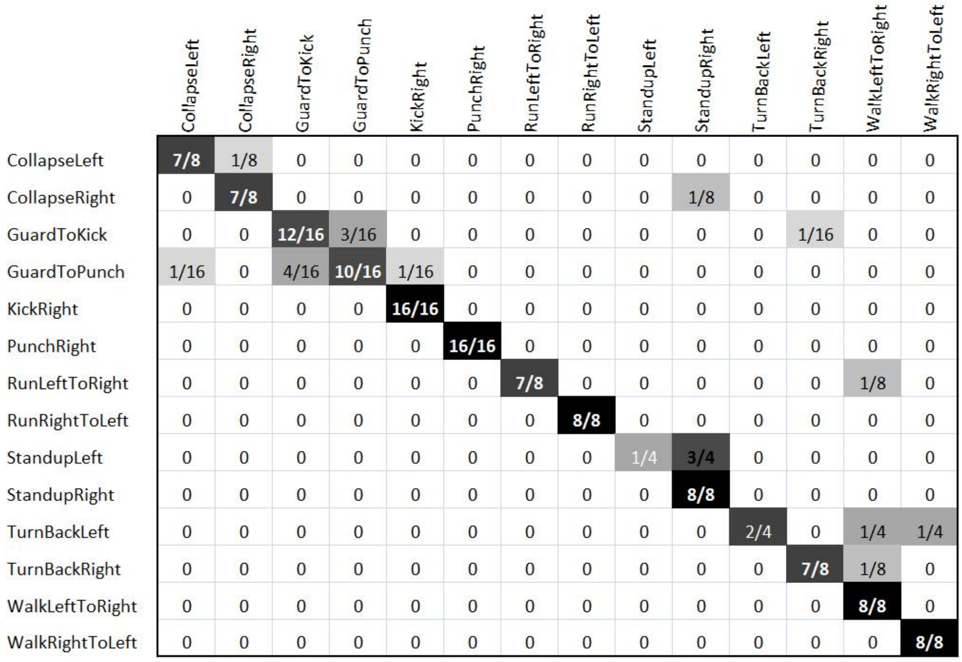
	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	9/9	0	0	0	0	0	0	0	0	0
jack	0	9/9	0	0	0	0	0	0	0	0
jump	0	0	8/9	1/9	0	0	0	0	0	0
pjump	0	0	0	9/9	0	0	0	0	0	0
run	0	0	0	0	10/10	0	0	0	0	0
side	0	0	0	1/9	0	8/9	0	0	0	0
skip	0	0	1/10	0	1/10	0	7/10	1/10	0	0
walk	0	0	0	0	0	0	0	10/10	0	0
wave1	0	0	0	1/9	0	0	0	0	8/9	0
wave2	0	0	0	0	0	0	0	0	0	9/9

Figure 3.6: Confusion matrix for Weizmann dataset

Table 3.1 compares our approach to other methods. While there are a number of motion-based action recognition approaches which report accuracies between 90% and 100%, we compare our approach to the existing non-temporal (key pose based) methods. Only [Weinland and Boyer, 2008] achieved significantly higher accuracy than the proposed method. However, recall that, in contrast to their method, our approach does not require any subsampling of the data. Moreover their approach is based on forward selection of key poses, which is computationally expensive for large and versatile action datasets. Thurau [2007] reported accuracies of 86.6% and 94.4% by using non-temporal unigrams and 2-frame temporal bigrams, respectively. In terms of methodology, the work of Baysal et al. [2010] is most close to our approach. It follows that by using contour-based pose features, we can achieve significantly higher accuracy. Moreover, our approach is very efficient for its feature extraction, dimensionality reduction, and the similarity measure.

Approach	Act.	Seq.	Acc.(%)
[Baysal et al., 2010]	9	81	92.6
[Thurau, 2007]	10	90	86.6
[Weinland and Boyer, 2008]	10	90	100
Our approach	9	81	**97.5**
	10	93	**93.5**

Table 3.1: Comparison of our approach with other non-temporal approaches on Weizmann dataset

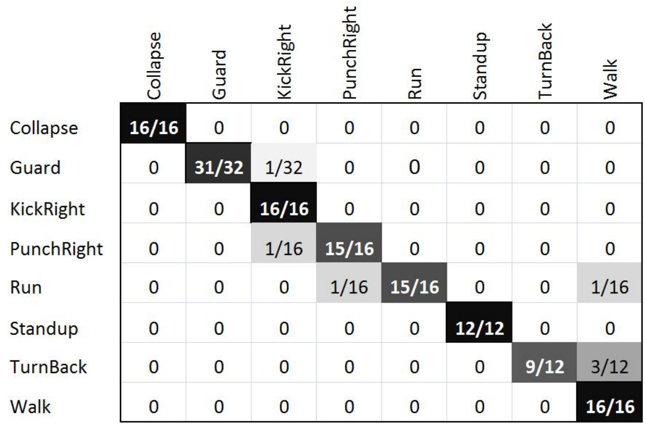


	CollapseLeft	CollapseRight	GuardToKick	GuardToPunch	KickRight	PunchRight	RunLeftToRight	RunRightToLeft	StandupLeft	StandupRight	TurnBackLeft	TurnBackRight	WalkLeftToRight	WalkRightToLeft
CollapseLeft	7/8	1/8	0	0	0	0	0	0	0	0	0	0	0	0
CollapseRight	0	7/8	0	0	0	0	0	0	0	1/8	0	0	0	0
GuardToKick	0	0	12/16	3/16	0	0	0	0	0	0	0	1/16	0	0
GuardToPunch	1/16	0	4/16	10/16	1/16	0	0	0	0	0	0	0	0	0
KickRight	0	0	0	0	16/16	0	0	0	0	0	0	0	0	0
PunchRight	0	0	0	0	0	16/16	0	0	0	0	0	0	0	0
RunLeftToRight	0	0	0	0	0	0	7/8	0	0	0	0	0	1/8	0
RunRightToLeft	0	0	0	0	0	0	0	8/8	0	0	0	0	0	0
StandupLeft	0	0	0	0	0	0	0	0	1/4	3/4	0	0	0	0
StandupRight	0	0	0	0	0	0	0	0	0	8/8	0	0	0	0
TurnBackLeft	0	0	0	0	0	0	0	0	0	0	2/4	0	1/4	1/4
TurnBackRight	0	0	0	0	0	0	0	0	0	0	0	7/8	1/8	0
WalkLeftToRight	0	0	0	0	0	0	0	0	0	0	0	0	8/8	0
WalkRightToLeft	0	0	0	0	0	0	0	0	0	0	0	0	0	8/8

Figure 3.7: Confusion matrix for MuHAVi-14

3.5.2 Results on MuHAVi Dataset

MuHAVi [Singh et al., 2010] is another multi-camera and multi-action controlled video dataset. It consists of videos of 17 activities performed multiple times by 14 actors. The action sequences are captured by 8 different CCTV cameras, each with an angular difference of 45°. Silhouettes of 14 primitive actions performed by 2 actors (A1 and A4) captured from 2 views (45° and 90°) were manually annotated and made publicly available. This dataset (also known as MuHAVi-MAS) contains 136 annotated silhouette sequences and provides a test bed for benchmarking silhouette-based action recognition approaches. In the following discussion, we refer to this data as *MuHAVi-14*. The contained primitive actions can be further grouped into 8 action classes.



	Collapse	Guard	KickRight	PunchRight	Run	Standup	TurnBack	Walk
Collapse	16/16	0	0	0	0	0	0	0
Guard	0	31/32	1/32	0	0	0	0	0
KickRight	0	0	16/16	0	0	0	0	0
PunchRight	0	0	1/16	15/16	0	0	0	0
Run	0	0	0	1/16	15/16	0	0	1/16
Standup	0	0	0	0	0	12/12	0	0
TurnBack	0	0	0	0	0	0	9/12	3/12
Walk	0	0	0	0	0	0	0	16/16

Figure 3.8: Confusion matrix for MuHAVi-8

For example, “WalkLeftToRight” and “WalkRightToLeft” may be merged into “Walk”. We refer to this merged dataset as *MuHAVi-8*.

In order to validate our approach w.r.t. the multi-view, multi-actor nature of the dataset, we performed different experiments, as suggested in [Singh et al., 2010].

Leave-one-out Cross Validation

In this test, we iteratively trained the classifier on all instances except one, and tested it on the left-out instance. Finally, the average accuracy was calculated over all 136 silhouettes. By using $k = 60$, we achieved an accuracy of up to 86.03% and 95.58% for MuHAVi-14 and MuHAVi-8, respectively. See Figs. 3.7 and 3.8 for the resulting confusion matrices.

Notice that our approach is able to distinguish between actions involving similar poses in different temporal order. For instance, “Collapse” and “Standup” as well as actions involving many overlapping poses in the same order (e.g. “GuardToKick” and “KickRight”) can be distinguished. Action-wise comparisons to the temporal baseline approach [Singh et al., 2010] are listed in Tables 3.2 and 3.3.

Identical Training and Test Cameras, Novel Test Actor

In this experiment, we trained our classifier on all instances related to one actor, tested on the data of the other actor, and calculated average classification

	Accuracy (%)	
Action	**Baseline**[Singh et al., 2010]	**Our Approach**
CollapseLeft	50.0	87.5
CollapseRight	62.5	87.5
GuardToKick	81.2	75.0
GuardToPunch	62.5	62.5
KickRight	93.7	100.0
PunchRight	100.0	100.0
RunLeftToRight	87.5	87.5
RunRightToLeft	87.5	100.0
StandUpLeft	0.0	25.0
StandUpRight	100.0	100.0
TurnBackLeft	100.0	50.0
TurnBackRight	87.5	87.5
WalkLeftToRight	100.0	100.0
WalkRightToLeft	87.5	100.0
	82.35	**86.03**

Table 3.2: Action-wise comparison of our approach with the baseline on MuHAVi-14

	Accuracy (%)	
Action	**Baseline**[Singh et al., 2010]	**Our Approach**
Collapse	100.0	100.0
Guard	100.0	96.9
KickRight	93.7	100.0
PunchRight	100.0	93.7
Run	93.7	93.7
StandUp	100.0	100.0
TurnBack	91.7	75.0
Walk	100.0	100.0
	97.80	95.58

Table 3.3: Action-wise comparison of our approach with the baseline on MuHAVi-8

rates. A comparison with the baseline is given in Tables 3.4 and 3.5. Note that the baseline evaluation in [Singh et al., 2010] was based on training on Actor-1 and testing on Actor-4. However, we alternatively considered both actors for training and testing.

Again, we observed a significant improvement in accuracy for both MuHAVi-14 and MuHAVi-8 collections. An increase of about 12% in accuracy, for MuHAVi-14, shows the relative robustness of the proposed approach towards

	Accuracy (%)	
Action	**Baseline**[Singh et al., 2010]	**Our Approach**
CollapseLeft	75.0	87.5
CollapseRight	50.0	87.5
GuardToKick	0.0	68.7
GuardToPunch	0.0	25.0
KickRight	87.5	81.3
PunchRight	100.0	68.7
RunLeftToRight	100.0	75.0
RunRightToLeft	75.0	100.0
StandUpLeft	0.0	50.0
StandUpRight	100.0	75.0
TurnBackLeft	50.0	75.0
TurnBackRight	50.0	75.0
WalkLeftToRight	100.0	100.0
WalkRightToLeft	100.0	75.0
	61.76	**73.53**

Table 3.4: Novel actor validation on MuHAVi-14

	Accuracy (%)	
Action	**Baseline**[Singh et al., 2010]	**Our Approach**
Collapse	75.0	100.0
Guard	50.5	75.0
KickRight	87.5	81.25
PunchRight	100.0	62.5
Run	87.5	75.0
StandUp	83.3	100.0
TurnBack	50.0	83.3
Walk	100.0	100.0
	76.47	**83.08**

Table 3.5: Novel actor validation on MuHAVi-8

individual characteristics of actors. Although human silhouettes differ in test and training data, the novel combination of scale-invariant features with discriminative key-pose learning exhibits improved performance.

Identical Training and Test Actors, Novel Test Camera

In this scenario, we aimed to determine the robustness of the algorithm towards changes in viewpoint. Here, we trained our classifier on all instances captured by one camera and tested on data captured by the other camera. We alternatively considered both camera-views for training and testing. Our

	Accuracy (%)	
Action	**Baseline**[Singh et al., 2010]	**Our Approach**
CollapseLeft	0.0	87.5
CollapseRight	50.0	37.5
GuardToKick	0.0	50.0
GuardToPunch	25.0	0.0
KickRight	87.5	100.0
PunchRight	75.0	62.5
RunLeftToRight	0.0	50.0
RunRightToLeft	0.0	50.0
StandUpLeft	50.5	25.0
StandUpRight	100.0	62.5
TurnBackLeft	100.0	25.0
TurnBackRight	75.0	62.5
WalkLeftToRight	0.0	0.0
WalkRightToLeft	75.0	62.0
	42.6	**50.0**

Table 3.6: Novel view validation on MuHAVi-14

	Accuracy (%)	
Action	**Baseline**[Singh et al., 2010]	**Our Approach**
Collapse	62.5	56.3
Guard	18.7	40.6
KickRight	87.5	87.5
PunchRight	75.0	56.2
Run	0.0	37.5
StandUp	83.3	91.6
TurnBack	83.3	83.3
Walk	37.5	37.5
	50.0	**57.4**

Table 3.7: Novel view validation on MuHAVi-8

average results for the two cases are compared with the baseline in Tables 3.6 and 3.7.

Although our approach outperformed the baseline method, the accuracy results demonstrate the challenging nature of this problem. We notice, in particular, a lower performance of our pose-based approach for actions where the novel pose involves high self-occlusion (e.g. GuardToPunch and PunchRight). Given that our approach relies on shape information only, a large change in viewpoint (45° in this case) can obviously affect its performance. We expect that, if the size and variety of training data were increased (e.g. by adding more actors

Approach	**Weizmann**	**MuHAVi-14**	**MuHAVi-8**
Discriminative Weights	93.5	86.03	95.58
Majority Voting	86	81.6	91.9

Table 3.8: Significance of our discriminative weighting scheme

or views to the training set), our simple yet effective approach would perform even better.

3.6 Conclusion

In this chapter, we presented a novel, simple yet effective approach to pose-based action recognition in constrained videos and silhouette images. We could show that by employing a contour-based pose representation and an efficient weighting scheme that favors distinctive key poses, a very high recognition accuracy could be achieved on standard benchmark data, even though the presented approach does not incorporate any temporal information or implicit modeling of the underlying sequence of key poses. The contribution of the proposed CDS key poses is apparent from experimental results on the Weizmann dataset (Table 3.1), where our approach outperforms [Baysal et al., 2010] and [Thurau, 2007], who use mutual information based weighting, but employ line-pairs or HOG-based feature representations. In order to show the significance of learning discriminative weights, we compared our algorithm to a naive majority-based approach that represents each action as a set of CDS key poses without weights (lines 1–4 in Algorithm 1) and uses majority voting to classify a given sequence. The (highest) results by leave-one-out cross validation given in Table 3.8 show the significance of our discriminative weighting scheme.

While we are confident that the addition of temporal information might further increase accuracy, the high recognition rates for a strictly pose-based approach are an interesting result. Although our approach already outperforms a recent baseline in more difficult *novel-view* and *novel-actor* scenarios, it may be further improved by enlarging the set of training data. The proposed approach is directly applicable to controlled scenarios such as surveillance and human-computer interaction. However, it will be interesting to see the benefits of discriminative key poses in complex scenarios.

To this end, Liu et al. [2013] recently used a similar weighting scheme to determine a set of discriminative key poses for action recognition in more realistic scenarios (e.g. a subset of HMDB videos that contain single-person actions). They used features that encode contour, orientation, and intensity information. Compared to the one-shot weight learning in our approach, they used Adaboost to iteratively determine the difficult (discriminative) key poses. In contrast to our hard assignment of key poses to query frames, they used soft assignment and achieved state-of-the-art performance on a number of activity recognition datasets. Other recent approaches that relate-to or extend our approach include: [Chaaraoui et al., 2012], [Zanfir et al., 2013] and [Climent-Prez et al., 2013]. For example, Climent-Prez et al. [2013] adopted the CDS representation to 3D skeletal data and applied a similar weighting scheme to determine discriminative joints for action recognition in MSR-Action3D data[Li et al., 2010].

Summary

Pose-based action recognition has generated a growing interest for its efficiency and robustness towards missing data, observational latency, and online classification. This chapter described a novel, efficient, pose-based approach. The proposed approach represents a human pose as a contour distance signal (CDS) feature that takes into account the distances of the contour points from the center of mass. The CDS is a compact representation, and it can be efficiently extracted from silhouettes. Each action category is represented by a set of $(keypose, weight)$ pairs where the key poses are determined by k-means clustering for each class. The weights are learned by a novel mutual information-based scheme that favors distinctive key poses (only present within specific action classes) and moderates ambiguous key poses (common among different actions). A query sequence is classified by identifying the nearest neighbor key poses and their classes for all of its frames and accumulating the corresponding weights per class. In this way, we achieve real-time classification with high recognition accuracy for single- as well as multi-view controlled action datasets. Recently, many researchers have directly used or extended the ideas discussed in this chapter to action recognition in more challenging scenarios (unconstrained videos and 3D data). Consequently, they have presented more sophisticated systems of learning discriminative key poses ([Climent-Prez et al., 2013], [Zanfir et al., 2013], [Liu et al., 2013]) and exploiting contour-based information ([Chaaraoui et al., 2012], [Liu et al., 2013], [Climent-Prez et al., 2013]).

Chapter 4

Applicability of Keypose based Learning to Large Scale Gait Recognition

The approach proposed in the previous chapter builds on three concepts: (a) a contour based representation of silhouette images (b) a key pose based representation of action categories and (c) a weighting scheme that identifies most discriminative key poses based on inter-class and intra-class variation. While, our approach [Cheema et al., 2011] and related work [Chaaraoui et al., 2012, Liu et al., 2013, Climent-Prez et al., 2013, Zanfir et al., 2013] show promising results on different action datasets, we are interested in extending the applicability of key pose based classification to a related large-scale problem – gait recognition. Gait recognition, which aims to recognize people by the way they walk, historically benefited from developments in action recognition and often serves as a test bed for evaluating the scalability of the methods used in action recognition. Note that while action recognition assumes and maximizes inter-class variation, gait recognition faces low inter-class variation since different classes (person labels) are deviations of a single action, "walk". Our experimental evaluation on a large gait dataset shows that the high inter-class ambiguities at large scale may undermine the discriminative weighting scheme. Still the contour-based features combined with key poses and a simple majority voting scheme outperforms most other approaches that model individual gait patterns as sequences of temporal templates either by determining gait cycles or by aggregating spatio-temporal information into a 2D signature [Cheema et al., 2012b].

4.1 Introduction

Most existing benchmarks for pose-based action recognition in controlled (surveillance) environments are restricted to a few primitive actions such as “run“, “kick“, “punch“, etc. These primitive actions can be considered as different sequences or combinations obtained from a set of basic gestures or poses. While different actions may share some poses, they are usually characterized by their very discriminative poses or execution order. In such cases, a (discriminative) key pose based framework may perform well – as seen in the previous chapter. In this chapter, we investigate applicability of the contour distance signal (CDS) and key pose based recognition to a large-scale activity recognition scenario. Specifically, we focus on gait recognition that can be considered as a fine-grained action recognition problem with low inter- and intra-class variation.

Gait recognition research can be divided into two categories: model-based approaches that consider the motion of joints and model-free approaches that rely on holistic features of shapes (silhouettes) or motions of human bodies as a whole. Model-based methods are usually robust to changes in view and scale. Nonetheless, they are sensitive to image quality and incur high computational cost in RGB videos. Model-free approaches are insensitive to the quality of silhouettes and can be computed efficiently. Most of the current approaches are model-free. Our keypose based approach, when applied to gait recognition, can be considered model-free. Note that the idea of extending human action recognition to gait recognition is by no means novel. Often gait recognition approaches benefit from developments in action recognition and vice versa. For example, gait energy images (GEI) [Han and Bhanu, 2006] and gait history images (GHI) [Jianyi and Nanning, 2007] are built on the ideas of motion energy images (MEI) and motion history images (MEI) presented for action recognition in [Bobick and Davis, 2001]. However, to our knowledge, we are the first to propose a strictly non-temporal pose-based approach to large-scale gait recognition [Cheema et al., 2012b].

Our experimental results show that, though the discriminative weighting does not scale to such problems, a naive combination of CDS shape (model-free) representation, key pose based category representation and majority voting based classification can achieve state-of-the-art performance on benchmark gait data. On the one hand, this approach, by its very nature, does not compromise useful information concerning spatio-temporal variations. On the other hand, it does not require any temporal alignment or dynamic time warping for gait classification. Furthermore, intrinsic dimensionality reduction is accomplished by using the CDS compact feature. These characteristics make the proposed approach highly accurate and capable of real-time computation.

4.2 Related Work

Over the last decade, a number of vision-based techniques for gait recognition have been proposed [Wang et al., 2003b, Han and Bhanu, 2006, Zhang et al., 2009, Chen et al., 2009, Goffredo et al., 2010, Kusakunniran et al., 2011]. Most recent approaches are model-free, i.e. they rely on holistic shape or motion features and do not try to estimate joint locations. For example, gait energy images (GEI) [Han and Bhanu, 2006] and gait history images (GHI) [Jianyi and Nanning, 2007] convert spatio-temporal information of a walk sequence into a single 2D image. Chen et al. [2009] proposed frame difference energy images (FDEI) to preserve kinetic and static information in each frame, even when the silhouettes are incomplete.

Many model-free approaches employ Procrustes shape analysis (PSA) to obtain affine invariant descriptors known as Procrustes mean shape (PMS) [Wang et al., 2003a]. Zhang et al. [2009] used a computationally expensive combination of shape context (SC) and PMS to address gait recognition. Kusakunniran et al. [2011] proposed the so-called pairwise shape configuration (PSC) which embeds local shape information, opposed to the holistic nature of PMS. The local spatial information in PSC is embedded by automatically determining head and feet position. Other approaches represent gait by extracting multiple frames from gait cycles. For example, Collins et al. [2002] chose four frames that correspond to peaks and valleys of a gait cycle. In [Wang et al., 2003b], dynamic time warping is used for comparing average gait cycles based on contour distance features. In [Chen et al., 2011], a two-level dynamic Bayesian network (layered time series model (LTSM)) is proposed which models each cluster of *temporally* adjacent frames as logistic dynamic texture. Recently, Iosifidis et al. [2012] employed linear discriminant analysis (LDA) and fuzzy vector quantization (FVQ) on raw silhouettes for person identification.

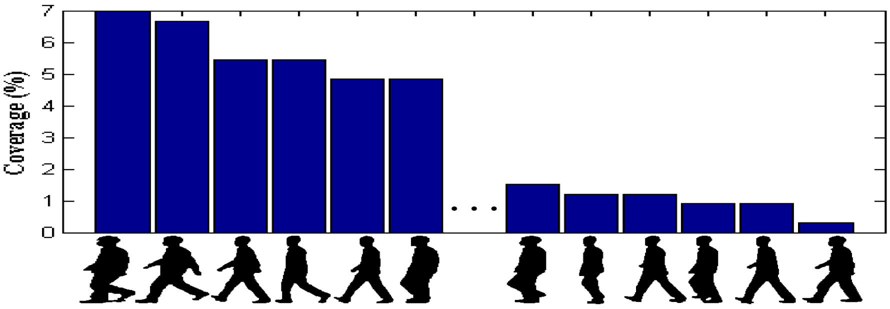
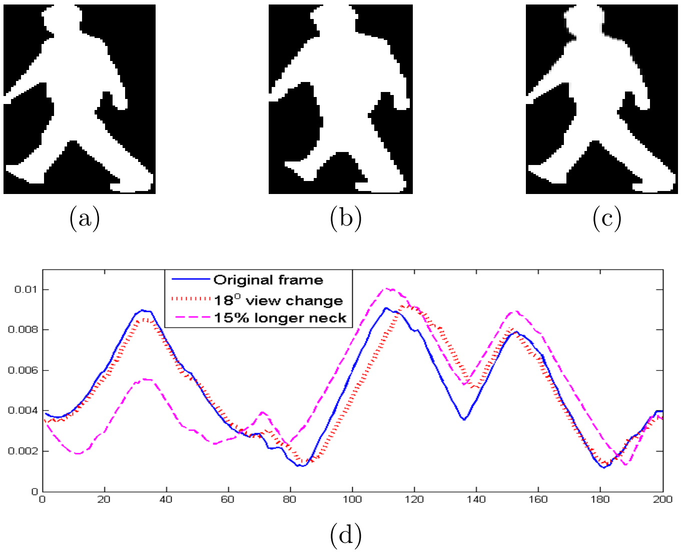


Figure 4.1: 6 top and 6 bottom key poses of an individual's walk and their coverage

Reviewing state-of-the-art methods, we notice that: (a) almost all approaches depend on reliable estimation of gait cycles, which are not feasible in many scenarios, e.g. in case of interrupted or partially visible walk sequences, (b) although model-free approaches are relatively efficient, most of them still incur high computational cost due to frame by frame alignments, (c) most existing approaches (e.g. GEI, GHI, PMS) tend to compress information of the walking behavior of an individual into a single template, which may neglect useful information about spatio-temporal variations.

4.3 Learning (Distributed) Key Poses for Gait Recognition

Here, for the sake of completeness, we briefly describe our key pose based approach. Given training sequences of silhouettes for C classes, we first extract CDS features for each frame of every sample. These feature vectors are grouped together for each class without preserving any order of the gait cycles or sequences. K-means clustering with Euclidean distance is then employed to determine k cluster centers for each class. For a sufficient value of k, these cluster centroids represent common poses as well as less frequent poses of an individual's walk. This can be seen in Fig. 4.1, which shows the coverage of training data by different key poses for and individualś walk sequences. To classify a given query video (sequence of CDS), we determine the nearest key poses for all frames and apply majority voting to compute the class label. We notice that in the case of gait recognition, majority voting outperform several action recognition weighting schemes [Thurau and Hlavac, 2008, Baysal et al.,



(a) (b) (c)

(d)

Figure 4.2: Invariance of CDS feature against small changes in viewpoint and body shape: (a) Original frame (b) with an 18° change in viewpoint (c) with a 15% longer neck (d) comparison of CDS's of (a), (b), (c).

2010, Cheema et al., 2011] by 20% to 30%.

In contrast to existing gait recognition methods, we represent a gait pattern as a collection of non-temporal key poses distributed over a whole walk sequence. Also, we do not embed any temporal context into the features or into the classifier. Our representation is *distributed* in the sense that an individual's gait is modeled as a *set* of key poses that are determined without explicitly approximating the gait cycles but still cover different gait phases. In contrast to other approaches, the underlying gait representation is not constrained by any structural limitations imposed by gait cycles or aggregate templates.

A major advantage of our approach is the use of CDS feature, which is scale invariant and can be efficiently computed. Compared to the size of the original image, the size of the contour is much smaller. This implicit dimensionality reduction through transforming a silhouette to CDS consequently enables efficient learning and classification. Furthermore, CDS is invariant to a realistic (small) change in viewpoint. This is apparent from Fig. 4.2, where we compare the CDS of a silhouette with respect to a small change in viewpoint and a small change in body configuration. This invariance is achieved because: (a) extreme points including p_1 remain extreme after a small change in viewpoint and (b) relative positions of points on contour remain (almost) intact.

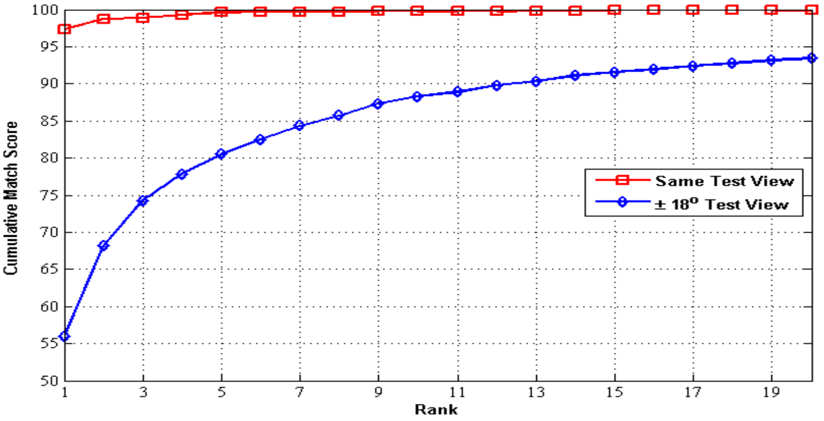


Figure 4.3: CMS curves for person identification with the same unchanged viewpoint and with 18° change

4.4 Experimental Results

We performed an experimental evaluation on the CASIA-B dataset which is one of the largest and most used benchmarks for gait recognition [Yu et al., 20006]. CASIA-B contains walk sequences of 124 subjects captured from 11 different viewpoints, ranging from 0°(front view) to 180°(back view). As with most competing methods (see Table 4.1), we consider the "normal" videos that do not include samples where the individual wears a coat or carries a bag. There are 6 videos for each person under each different viewing angle and silhouettes are provided. Sequences which contain very noisy silhouettes (e.g. 3 connected components of almost equal size) are discarded[6]. After filtering this way, the data of 106 subjects were considered for the evaluation.

Following the standard practice, for each person, the first 4 sequences were chosen for training and the remaining 2 for testing. For classification of a given test sequence, each of its frames was compared to key poses of candidate classes, and a decision was made by majority voting. We evaluated our approach for the following two cases: (i) the viewpoint remains the same for training and testing data (ii) there is a 18° difference in the viewpoints of the two sets. Results are reported for settings $k = 48$ and $s = 200$.

In case of an unchanged view, we achieved a very high average recognition accuracy of **97.3%**, which is among the highest scores obtained on this dataset.

[6] Many of the silhouettes are very noisy. In fact, 14 subjects have more than 30% incomplete frames [Chen et al., 2009]

(Test, Train)	PMS-PSA	GEI-MDA	Model-based	PSC-PSA	Our
$(0°, 0°)$	–	–	–	–	98.6
$(0°, 18°)$	–	–	–	–	46.2
$(18°, 0°)$	–	–	–	–	22.2
$(18°, 18°)$	–	–	–	–	98.6
$(18°, 36°)$	–	–	–	–	71.2
$(36°, 18°)$	45.0	39.5	–	65.6	68.9
$(36°, 36°)$	76.7	98.1	72.1	96.0	99.1
$(36°, 54°)$	23.3	33.2	64.6	77.6	65.6
$(54°, 36°)$	21.1	27.7	56.5	78.1	64.2
$(54°, 54°)$	75.4	98.0	79.5	97.7	97.1
$(54°, 72°)$	22.0	21.5	65.1	74.2	58.0
$(72°, 54°)$	18.3	16.3	72.1	80.1	51.4
$(72°, 72°)$	77.2	98.3	85.0	97.7	97.7
$(72°, 90°)$	38.7	46.5	64.0	74.2	63.7
$(90°, 72°)$	36.8	50.4	72.6	84.5	63.2
$(90°, 90°)$	77.4	99.2	86.5	97.7	94.4
$(90°, 108°)$	46.9	73.7	69.2	85.3	70.3
$(108°, 90°)$	45.0	65.5	64.0	70.0	55.2
$(108°, 108°)$	72.4	98.8	82.3	96.0	93.9
$(108°, 126°)$	33.1	28.8	72.8	75.9	78.3
$(126°, 108°)$	45.0	31.7	67.6	78.2	76.4
$(126°, 126°)$	79.0	98.1	81.1	95.8	98.1
$(126°, 144°)$	46.0	36.2	–	78.0	74.1
$(144°, 126°)$	44.0	37.6	–	70.3	72.2
$(144°, 144°)$	78.0	98.7	–	96.5	98.6
$(144°, 162°)$	16.0	4.4	–	33.9	28.3
$(162°, 144°)$	–	–	–	–	48.1
$(162°, 162°)$	–	–	–	–	97.7
$(162°, 180°)$	–	–	–	–	17.9
$(180°, 162°)$	–	–	–	–	24.1
$(180°, 180°)$	–	–	–	–	97.7
Avg*	45.0	57.2	72.2	81.1	**74.7**

Table 4.1: Comparison of our approach with other representative approaches. *To make a fair comparison, the average was calculated for $Test \in [36°, 144°]$.

For a small change in viewing angle, we achieved a very reasonable average recognition rate of **56%** over all viewpoints and **63.6%** on commonly reported lateral viewpoints, i.e. between 36° and 144°. Figure 4.3 plots the Cumulative Match Score (CMS) for the two cases. The CMS is computed by finding the rank of the query among an ordered list of predicted classes. Notice, in particular, that under small view changes, the probability of having the correct subject among the 12 top-ranking subjects is above 90%.

Table 4.1 shows detailed results of our approach for different viewpoints and

Approach	**same-view**	**changed-view**
PMS-PSA [Wang et al., 2003a]	76.6	29.2
GEI-MDA [Han and Bhanu, 2006]	98.5	36.6
GEI-HMM [Chen et al., 2009]	83.2	–
GHI-HMM [Chen et al., 2009]	62.1	–
FDEI-HMM [Chen et al., 2009]	93.9	–
Model-based self calibration [Goffredo et al., 2010]	81.1	65.5
PCA-MDA [Liu et al., 2010]	97.7	–
PSC-PSA [Kusakunniran et al., 2011]	96.7	73.2
Wavelet-LTSM [Chen et al., 2011]	95.7	–
FVQ-LDA [Iosifidis et al., 2012]	93.3	–
Our approach	**97.3**	**63.6**

Table 4.2: Comparison of our approach with other representative approaches.

compare them to other multi-view approaches. Table 4.2 presents a further comparison of our approach with state-of-the-art approaches for the two cases (no change and a small change in viewpoint). Only pair-wise shape configuration [Kusakunniran et al., 2011] shows a matching performance. However note that [Kusakunniran et al., 2011] is a semi model based approach which depends on localizing the left and right foot. Consequently, it can not be applied to (near) frontal or (near) back views (See second-last column of Table 4.1). Notice also that the GEI-based approach [Han and Bhanu, 2006], though relatively accurate for unchanged-views, is not robust against small changes in viewing angle. Our approach of learning key poses based on the contour distance signal, on the other hand, shows a high and consistent performance. All experiments were carried out on a standard laptop using MATLAB 7, and we achieved nearly real-time classification performance, i.e. 12 frames per second, using a single core.

4.5 Conclusion

We applied an efficient, key pose based approach to large-scale gait-based human identification. We could show that, using a contour-based pose representation and key pose learning, very high recognition accuracy can be achieved on standard benchmark data, even though the presented approach does not incorporate any temporal information or implicit modeling of the underlying sequence of key poses. Although, static key pose based approaches have recently been applied successfully to activity recognition, hardly any such approach was applied on gait recognition. This chapter thus establishes the effectiveness of non-temporal pose-based methods for large-scale activity recognition problems such as gait recognition. Also, from the perspective of our key pose based classification presented in the previous chapter, we could (a) identify the limitation of the discriminative weighting scheme to some scenarios, and (b) express the strength of CDS representation and key pose based learning to model human activities at a different level of granularity.

Our CDS features, like any other shape-based descriptor [Han and Bhanu, 2006, Zhang et al., 2009, Iosifidis et al., 2012], may suffer from medium to large changes in a person's appearance e.g. due to carrying a bag. Although this is not within the core focus of our research, it will be interesting to see in future how such pose-based approaches can perform in conjunction with domain adaptation methods i.e. by learning transformations from a normal case to different conditions such as clothing and carrying objects. Another interesting idea is building models of motion in consecutive frames and then applying key pose based learning. Such approaches have successfully been applied to action recognition (e.g. Bigrams [Thurau and Hlavac, 2008] and Moving Poses [Zanfir et al., 2013]).

Summary

Gait recognition can be considered a fine-grained activity recognition problem which is receiving increasing attention from computer vision researchers for its applicability in areas such as visual surveillance, access control, and smart interfaces. Most existing research attempts to model individual gait patterns as sequences of temporal templates either by determining gait cycles or by aggregating spatio-temporal information into a 2D signature. In this chapter, we extended the application of the contour distance signal and key pose based classification to large-scale gait recognition. We achieved high recognition accuracy and real-time classification on a multi-view gait dataset with over 100 subjects [Cheema et al., 2012b]. In short, we have established the effectiveness of an efficient combination of the contour-based features and key pose based learning for human activity analysis in a controlled environment.

Chapter 5

Activity Recognition in High Dimension Low Sample Size Unconstrained Data

Human activity recognition in large unconstrained databases such as YouTube and Flicker is a challenging task due to the presence of cluttered backgrounds, poor illumination conditions, camera motion, different viewpoints, occlusions, poor quality of the medium, and the evolution of the data. These challenges impede most approaches designed for activity analysis in controlled environments. Building efficient applications (such as content-based retrieval systems) for such realistic data calls for robust algorithms that can efficiently organize, analyze, classify, and retrieve this data. While the problem is largely unsolved, approaches that are based on extracting spatio-temporal features around interest points or motion trajectories have shown promising results in terms of recognition accuracy [Laptev et al., 2008, Kliper-Gross et al., 2012, Wang et al., 2013]. The dimensionality of these spatio-temporal features often ranges in the tens of thousands, whereas the number of classes and the number of labeled instances per class usually ranges between ten and a hundred. Such high dimension, low sample size (HDLSS) data are prone to neighborliness – the lack of a proper neighborhood among the instances in a very high dimensional space [Donoho and Tanner, 2005, Ahn et al., 2007]. Asymptotic studies show a tendency for HDLSS data to lie at the vertices of a regular simplex [Hall et al., 2005, Donoho and Tanner, 2005].

Existing methods of recognizing human activities in the wild, however, overlook the underlying distribution of the data and employ off-the-shelf classifiers such as Nearest Neighbor(NN) and Support Vector Machine(SVM). Such naive application may cause over-fitting, e.g. in case of SVMs that are observed to use nearly all the training data as a support of decision function; or it may result in under-fitting, e.g. in case of NNs that assume that the nearby points have the same label (high inductive bias). Moreover, for online HDLSS settings, they may compromise on accuracy (e.g. kNN) or efficiency (e.g. SVM). In this chapter, we address these issues and through extensive experimentation, we affirm the lack of neighborhoods within HDLSS data that undermines most existing classifiers. Consequently, we propose a QR factorization approach to Nearest Affine Hull (NAH) classification which remedies the HDLSS dilemma and noticeably reduces the time and memory requirements of existing methods. We show that the resulting non-parametric models provide smooth decision surfaces and yield efficient and accurate solutions in multiclass HDLSS scenarios. On a number of established benchmark datasets, the proposed NAH-lsq classifier outperforms other instance-based methods and shows competitive or superior performance compared to SVMs. In addition, for online settings, NAH-lsq is faster than online SVMs, as SVMs would need complete retraining.

5.1 Introduction

Modern computer vision and pattern recognition tasks deal with large amounts of data of arguably moderate dimensionality. High dimension, low sample size (HDLSS) data therefore constitute a special case which, however, is becoming increasingly common in practical settings. Consider, for example, the problem of recognizing activities in unconstrained web videos. The dimensionality of spatio-temporal features for videos often ranges in the tens of thousands whereas number of classes and the number of labeled instances per class usually ranges between ten and hundred. Classification of such multiclass data poses several challenges. For example, the curse of dimensionality as it is commonly known leads to a scenario where the neighborhood among the instances in a very high dimensional space tends to be uniform [Donoho and Tanner, 2005, Ahn et al., 2007].

The scarcity and sparsity of labeled training data results in simplicial class regions in the feature space where every data instance is a vertex of the convex hull of the dataset. Consequently, the principles underlying popular classification methods *viz* (a) approximation of class regions, (b) discrimination across

HDLSS	high dimension low sample size
LDA	linear discriminant analysis
kNN	k nearest neighbor
MNP	minimum norm point
NAH	nearest affine hull
NCH	nearest convex hull
NHD	nearest hyperdisk
SVM-OAA	one-against-all SVM
SVM-OAO	one-against-one SVM
SVM	support vector machine
SVM-SGD	SVM with stochastic gradient descent
SVD	singular value decomposition

Table 5.1: Acronyms used throughout this chapter

different class regions, and (c) low-rank approximations of the data, suffer from artifacts of high dimensionality. It is therefore important to understand and analyze the elemental distribution and geometry of the data in multiclass HDLSS scenarios. Most existing approaches in computer vision, and in particular in human activity recognition, do not pay attention to these issues. The popular trend is to use an off-the-shelf classifier such as Support Vector Machines, Linear Discriminant Analysis, or $k-$ Nearest Neighbors.

This chapter investigates the geometry of subspaces spanned by high dimensional feature descriptors related to human activities and our experimental results affirm the neighbourlessness of these data. Accordingly, we show that Affine Hulls, being loose approximations of class regions, tend to facilitate better or competitive classification in multi-class HDLSS. While the applicability of the existing SVD-based NAH approach [Cevikalp et al., 2008] to HDLSS classification is constrained by a compromise between time and memory, we propose an efficient NAH approach by adopting a least squares method based on QR factorization. We compare the empirical performance of the proposed NAH-lsq classifiers with that of the other hull-based methods classifiers with that of other hull-based methods, namely Nearest Hyperdisk, Nearest Convex Hull, and NAH-svd, as well as with more traditional approaches such as Minimum Norm Point, kNN, LDA, and SVM which represent different classes of algorithms. For instance, kNN classifiers are based on the principle of local proximity; LDAs focus on low rank approximations, and SVMs are based on the principle of maximum margin separation.

Our results show that the proposed NAH-lsq classifiers are competitive with SVMs in terms of accuracy and efficiency and far superior to all other classifiers in our tests. The decision surfaces of NAHs and one-against-all SVMs are among the most smooth surfaces – offering better generalization. Note that unlike most other methods (e.g. SVM and kNN), NAHs are inherently non-parametric, and like other lazy classifiers (e.g. kNN and NCH), they *ideally* require no training. We also show the efficiency of NAH-lsq to be comparable to one-against-one SVMs and that it is far superior to other instance-based approaches, including NAH-svd.

The empirical evaluation also reveals that optimal classification of HDLSS data is achieved when using almost all the training data as support of the decision surfaces. Consequently, we show that NAH-lsq is well suited for online learning where the fast SVM-based methods, e.g. LASVM [Bordes et al., 2005] and SVM-SGD [Bottou, 2010], suffer from expensive retraining. Furthermore, the non-parametric nature of NAH-lsq, that facilitates efficient model fitting without any cross validation, offers a clear advantage in online settings. In short, our work in this chapter provides important empirical insights into the complexity of the multiclass HDLSS classification problem (e.g. neighbourlessness and over-fitting) and the simplicity of the solution (e.g. due to NAH-lsq).

The rest of the chapter is organized as follows: Section 5.2 discusses related work on HDLSS classification. In Section 5.3, we describe different representation-based (geometric) classifiers. NAH-lsq, an efficient approach to Nearest Affine Hull for HDLSS classification, is presented in Section 5.4. Section 5.5 provides details as to our benchmark datasets and feature extraction methodology while Section 5.6 reports our results. Finally, Section 5.7 discusses the results and future directions.

5.2 Related Work

Analysis of HDLSS data has been an active area of theoretical and applied research throughout the last decade. Asymptotic studies reveal a tendency for high dimensional data to lie at the vertices of a regular simplex [Hall et al., 2005, Donoho and Tanner, 2005]. Hall et al. [2005] proved that for two sets X and Y in $\mathcal{R}^d$ where $d >> |X| + |Y|$ and no k points lie in a $k-2$ dimensional hyperplane, it is always possible to find a hyperplane that separates X and Y. Donoho and Tanner [2005] show that the projection of a simplex from very large n dimensions to a lower $d = \rho n$ dimensional polytope does not reduce the number of corresponding $l-$dimensional faces for $l \leq \lfloor \rho d \rfloor$. Even the property of *k-neighborliness* holds for a certain range of values of k. A polytope is called k-neighborly if every subset of k vertices forms a $(k-1)$-face [Gruenbaum, 2003]. [Ahn et al., 2007] proved that such a geometric representation of the data holds under mild conditions such as non-independent samples.

Other researchers have quantified the extent or degree of ultrametricity in a dataset [Rammal et al., 1985, Murtagh, 2009]. Murtagh [2009] argued that ultrametricity becomes pervasive as dimensionality and spatial sparsity increase and used this property for model-based clustering. Klement et al. [2008] proved that, for $d \to \infty$, random and non-random scenarios are not distinguished by any metric-based measure, i.e. distances become approximately equal. They also showed that the soft-margin approach does not improve the generalization performance of SVMs on HDLSS data. Zhang and Lin [2011] compared the performance of several conventional classifiers on simulated and two-class data and reported that SVM and Distance Weighted Discriminant techniques achieve relatively better performance than Mean Difference and the Naive Bayes classifiers. Recent work in [Bolivar-Cime and Marron, 2013] evaluates different binary discrimination methods in the context of HDLSS Gaussian data and points out that these methods are asymptotically equivalent except for Naive Bayes, which may have a different asymptotic behavior as d tends to infinity.

A conventional approach to high dimensional classification consists in fitting models that maximally separate the class regions; common examples are SVM and LDA. Another common approach is to build a geometric model of each class that approximates the region covered by it. Such approximations include affine hulls [Vincent and Bengio, 2001], convex hulls and polytopes [Nalbantov et al., 2007, Sekitani and Yamamoto, 1993], bounding hyperspheres [Tax and Duin, 2004] and bounding hyperdisks [Cevikalp et al., 2008]. Unlike margin-based classifiers, these region- or volume-based classifiers are instance based in the sense that they do not require an explicit formulation of a decision boundary between classes. All these models represent a class as a bounded region in a corresponding subspace, except for the affine hull-based approach which covers the whole affine subspace spanned by given data points. For high dimensional data, these models are preferred over conventional instance-based models such as kNN, since kNN implicitly assumes a dense sampling which requires training sets that are exponentially large in the dimensionality of the underlying feature space. Cevikalp et al. [2008] suggested the use of NHD[7] as a compromise between too loose a structure of affine hulls and too tight a structure of convex hulls. Recently, large margin classifiers based on NAH, NCH, and NHD have been studied further in [Cevikalp and Triggs, 2009, Cevikalp et al., 2010, Cevikalp and Triggs, 2013]. Moreover, affine hull based modeling is elegantly applied in [Hu et al., 2012] in order to approximate unseen appearances in the context of image set classification.

Human activity recognition in unconstrained videos and still images poses a practical problem that underlines the importance of investigating the performance of several approaches in real-world HDLSS data classification. The experimental evaluations in this chapter were performed on recent challenging benchmark datasets of unconstrained videos [Kuehne et al., 2011, Reddy and Shah, 2013], still images [Ikizler-Cinbis et al., 2009], and depth and skeletal data [Ofli et al., 2013] on which most prior works applied SVMs with linear or Gaussian kernels. An increasingly popular trend in human activity classification is to use multiple feature descriptors such as motion cues, pose- and scene-related information [Yao et al., 2011b, Reddy and Shah, 2013, Wang et al., 2013]. These methods employ either early fusion of feature descriptors or late fusion of ensemble classifiers. Regular SVM or multiple-kernel-based SVM classifiers are used accordingly. Obviously, these techniques can achieve higher performance as compared to settings where a single feature descriptor is used. Since the analysis of ensemble classifiers or multiple features is not the focal point of this chapter, we restrict our practical experiments to recent single descriptors which are known to exhibit good performance on these datasets.

[7] Refer to Table 5.1 for acronyms

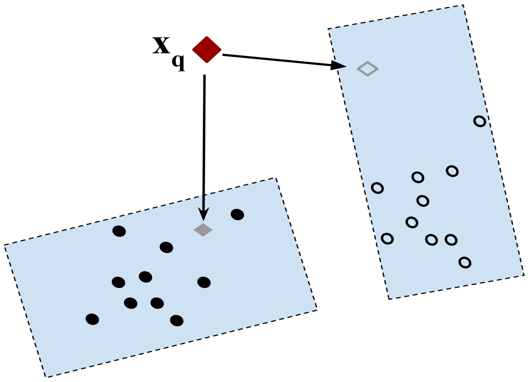


Figure 5.1: Visualization of the idea of nearest affine hull classification. The class assignment for a query point x_q is based on the minimal distance to its projections on affine subspaces.

5.3 Representation-based Classification

Essentially every approach to classification aims at discriminating between different classes either by determining appropriate decision functions in the feature space (e.g. SVMs or Decision Trees) or by relying on local or global instance-based representations of the classes (e.g. kNN or NCH). SVMs are often used *de facto* without paying attention to the geometry or distribution of class regions. This becomes very important in high dimensional classification problems. In this section, we review representative geometric classification methods in the context of high dimensional data.

5.3.1 Nearest Affine Hull Classification

The *affine hull* of a set of data is the smallest affine subspace that contains all the samples. Given training samples $\mathbf{x}_{ci} \in \mathcal{R}^d$ where $c \in \{1, 2, ..., C\}$ and $i \in \{1, 2, ..., N_c\}$ are class and instance indices, their affine hull is defined as

$$\mathbf{\Phi}_c^{aff} = \left\{ \mathbf{x} = \sum_{i=1}^{N_c} \alpha_i \mathbf{x}_{ci} \;\middle|\; \sum_i \alpha_i = 1 \right\}. \tag{5.1}$$

The affine hull provides a loose approximation of the class region in that it ignores exact locations of the training data, but models each class as an affine subspace. Consequently, it is least affected by artifacts that arise from assuming neighborliness in high dimensional spaces. The distance $d(\mathbf{x}_q, \mathbf{\Phi}_c^{aff})$ from a query point $\mathbf{x}_q$ to an affine hull is the norm of the displacement from the

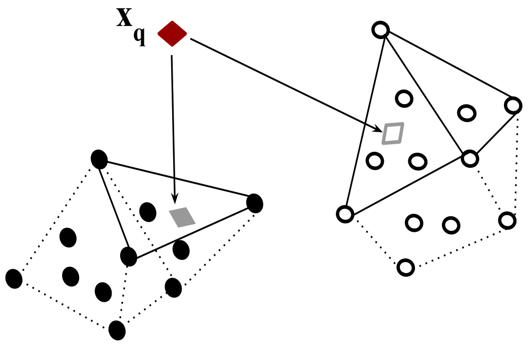


Figure 5.2: A visualization of nearest convex hull classification. The decision is based on minimal distance to the projection on convex hull facets.

closest point $\mathbf{x}_c^*$ on the hull. Equivalently, $d(\mathbf{x}_q, \mathbf{\Phi}_c^{aff})$ can be expressed as the orthogonal projection of the normal to the subspace. Figure 5.1 visualizes the concept of nearest affine hull classification.

Surprisingly, there are only few reports on affine hull based classifiers for high dimensional data. The work in [Cevikalp et al., 2008] proposed an offline training procedure using SVD and projections. There approach proceedes as follows: Let $\mathbf{X}_c$ denote a data matrix whose columns correspond to training examples from class c. The orthogonal projection $\mathbf{P}_c$ onto the spanning subspace can be determined by SVD of the centered matrix $\mathbf{X}_c^m = \mathbf{X}_c - \boldsymbol{\mu}_c$, where $\boldsymbol{\mu}_c$ is the centroid of the class c. In particular, $\mathbf{P}_c = \mathbf{U}\mathbf{U}^T$ where the matrix $\mathbf{U}$ contains the left singular vectors of $\mathbf{X}_c^m$, i.e. $\mathbf{X}_c^m = \mathbf{U}^T\mathbf{\Sigma}\mathbf{V}$. Given a query instance $\mathbf{x}_q$, its distance $d(\mathbf{x}, \Phi_c^{aff})$ from the affine hull of $\mathbf{X}_c$ can be directly computed as the orthogonal projection of $\mathbf{x}_q$ normal to the subspace, i.e.

$$d(\mathbf{x}_q, \mathbf{\Phi}_c^{aff}) = \big\|(\mathbf{I} - \mathbf{P}_c)(\mathbf{x}_q - \boldsymbol{\mu}_c)\big\|. \tag{5.2}$$

5.3.2 Nearest Convex Hull Classification

The *convex hull* of a given set of points is the minimal convex set that encloses them. Given training samples $\mathbf{x}_{ci} \in \mathcal{R}^d$, their convex hull is defined as

$$\mathbf{\Phi}_c^{conv} = \left\{ \mathbf{x} = \sum_i \alpha_i \mathbf{x}_{ci} \,\middle|\, \sum_i \alpha_i = 1,\ \alpha_i \geq 0 \right\} \tag{5.3}$$

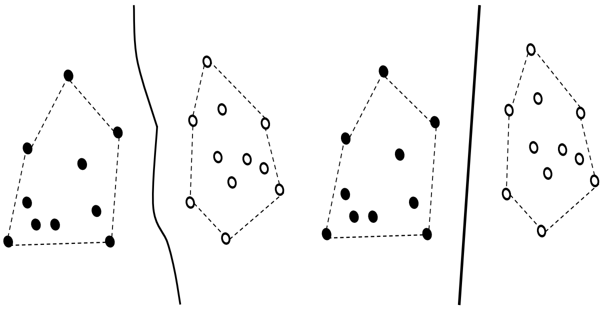


Figure 5.3: A 2D example of decision boundaries of (a) NCH (b) SVM

Compared to the affine hull, the convex hull provides a tight approximation of the class region because of the non-negativity constraints $\alpha_i \geq 0$ on the coefficients. The distance $d(\mathbf{x}, \Phi_c^{conv})$ between a query instance $\mathbf{x}_q$ and the convex hull of class c is calculated as the norm of the displacement of $\mathbf{x}_q$ to the closest point $\mathbf{x}_c^*$ on the convex hull. Unlike for affine hulls, where class regions are unbound subspace and the exact boundaries of the regions are not important, convex hull classification projects data onto class-specific regions within a subspace. In fact, the convex hull of $\mathbf{X}_c$ lies within the affine hull of $\mathbf{X}_c$. To determine the closest point on the convex hull requires solving the following constrained quadratic program

$$\begin{aligned} &\min_{\boldsymbol{\alpha}_c} \left\| \mathbf{x}_q - \mathbf{X}_c \boldsymbol{\alpha}_c \right\|^2 \\ &\text{s.t.} \sum_{i=1}^{N_c} \alpha_{ci} = 1 \\ &\qquad \alpha_{ci} \geq 0. \end{aligned} \tag{5.4}$$

The optimal solution $\boldsymbol{\alpha}_c^*$ of this problem provides the mixture coefficients to obtain $\mathbf{x}_c^* = \mathbf{X}_c \boldsymbol{\alpha}_c$. Thus the distance between the query and the class region becomes

$$d(\mathbf{x}_q, \boldsymbol{\Phi}_c^{conv}) = \left\| \mathbf{x}_q - \mathbf{X}_c \boldsymbol{\alpha}_c^* \right\|. \tag{5.5}$$

Note that every time a query instance arrives, a nearest point needs to be determined by solving the above problem. Compared to other instance-based approaches such as kNN or NAH, this approach is rather slow and may not be feasible for high dimensional data. Instead, SVMs work according to a similar principle and determine maximal margin hyperplanes in parametric forms in order to facilitate efficient classification. If a query instance x_q is considered as a class of its own, the problem of finding a maximum margin between this class and any other class c is equivalent to finding the closest points on the convex hulls $\mathbf{\Phi}_c^{conv}$ and $\mathbf{\Phi}_{x_q}^{conv}$ [Bennett and Bredensteiner, 2000]. In this sense, NCH classification is equivalent to SVM classification and the piecewise linear decision boundary of NCH contains the query-vs-class boundary of an SVM as a facet (see Fig. 5.3).

5.3.3 Nearest Hyperdisk Classification

A *hyperdisk* as proposed in [Cevikalp et al., 2008] can be understood as a compromise between the tight convex hull model of a class and the looser affine hull representation. Given data from a class c, its hyperdisk is defined as the intersection of the minimum enclosing hypersphere and the affine hull of the data; formally this amounts to

$$\mathbf{\Phi}_c^{disk} = \left\{ \mathbf{x} = \sum_i \alpha_i \mathbf{x}_{ci} \;\middle|\; \sum_i \alpha_i = 1,\ \|\mathbf{x} - \mathbf{s}_c\|^2 \leq r_c^2 \right\} \tag{5.6}$$

where $\mathbf{s}_c = \sum_{i=1}^{N_c} \alpha_i \mathbf{x}_{ci}$ is the center and r_c the radius of the bounding hypersphere. In order to determine the radius of the bounding hypersphere, the following constrained quadratic programming problem needs to be solved

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \sum_{i,j} \alpha_i \alpha_j \langle \mathbf{x}_{ci}, \mathbf{x}_{cj} \rangle - \sum_i \alpha_i \langle \mathbf{x}_{ci}, \mathbf{x}_{ci} \rangle \\ &\text{s.t.} \sum_{i=1}^{N_c} \alpha_i = 1 \\ &\quad 0 \leq \alpha_i \leq \gamma \end{aligned} \tag{5.7}$$

where α_i are Lagrange multipliers and $\gamma \in [0,1]$ is used to exclude distant points (outliers). The radius of the hypersphere becomes $r_c = \|\mathbf{x}_{ci} - \mathbf{s}_c\|$ for any $\mathbf{x}_{ci}$, with $0 \leq \alpha_i \leq \gamma$.

The distance between a query instance $\mathbf{x}_q$ and a hyperdisk $\mathbf{\Phi}_c^{disk}$ of class c is determined by two terms: (i) the norm of the displacement from $\mathbf{x}_q$ to its

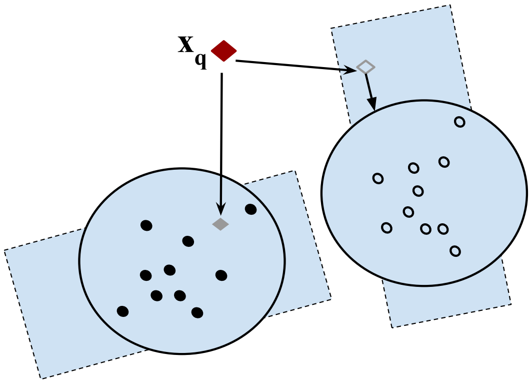


Figure 5.4: Nearest hyperdisk classification. The decision is based on projection on affine hull and distance from hypersphere

projection $\mathbf{x}_q^{aff}$ on the affine hull (5.2) and (ii) the distance of $\mathbf{x}_q^{aff}$ to the boundary of the hypersphere, that is

$$d(\mathbf{x}_q, \mathbf{\Phi}_c^{disk}) = \sqrt{\left\|\mathbf{x}_q - \mathbf{x}_q^{aff}\right\| + \max\left(\left\|\mathbf{x}_q^{aff} - \mathbf{s}_c\right\| - r_c, 0\right)^2} \tag{5.8}$$

The distance used for NHD classification couples a neighbourlessness affine hull distance and neighborhood-based hypersphere distance (see Fig. 5.4). Interestingly, as can be seen in our results (Section 5.6), the involvement of the neighborhood term (hypersphere distance) does not result in a gain of classification accuracy, instead we notice degraded performances in many cases.

In terms of efficiency, NHD outperforms NCH since the model parameters, i.e. those related to hypersphere and affine hull, can be computed beforehand or lately by the least squares method. Obviously, NHD is slower than NAH methods where only the affine hull is needed. The complexity of the underlying one-class model is again a compromise between NAH and NCH.

5.3.4 Minimum Norm Point Classification

A classical view of the problem of determining the closest point on a convex hull is the minimum norm point (MNP) problem. In his seminal work [Wolfe, 1976], Wolfe provided several basic results and proposed an iterative algorithm for the problem of finding the minimal Euclidean distance between a query point and the convex hull of a given set of points. Given a query point $\mathbf{x}_q$ and a set $\mathbf{X}$, the smallest norm point $\hat{\mathbf{x}} \in \mathbf{\Phi}^{conv}(X)$ can be determined by transforming all the data such that $\mathbf{x}_q$ becomes the origin, i.e. using $\bar{\mathbf{X}} = \{\mathbf{x} - \mathbf{x}_q, \forall \mathbf{x} \in \mathbf{X}\}$. Let $\mathbf{H}(\mathbf{p}, \alpha) = \{\mathbf{x} \mid \langle \mathbf{p}, \mathbf{x} \rangle = \alpha\}$ and $\mathbf{H}^+(\mathbf{p}, \alpha) = \{\mathbf{x} \mid \langle \mathbf{p}, \mathbf{x} \rangle \geq \alpha\}$ respectively denote the hyperplane and the half space due to a point $\mathbf{p}$ and a real number α. Then, according to [Wolfe, 1976]

Theorem 1 $\hat{\mathbf{x}} \in \mathbf{\Phi}_c^{conv}(\bar{\mathbf{X}})$ *is a minimum norm point iff* $\mathbf{X} \subseteq \mathbf{H}^+(\hat{\mathbf{x}}, \|\hat{\mathbf{x}}\|)$ *or equivalently* $\|\hat{\mathbf{x}}\| \leq \langle \hat{\mathbf{x}}, \mathbf{x} \rangle \forall \mathbf{x} \in \mathbf{X}$

The Wolfe algorithm for finding the minimum norm point forms a simplex of a subset of several affinely independent points $\mathbf{Q}$, starting with $\mathbf{Q} = \emptyset$. It solves a system of linear equations and finds a minimum norm point on the affine hull containing the simplex. If the point lies inside the relative interior of the simplex, it extends $\mathbf{Q}$ with another point of $\bar{\mathbf{X}}$ and forms a higher dimensional simplex. Otherwise, it drops a point from $\mathbf{Q}$ and forms a simplex of a lower dimension. The algorithm stops once $\hat{\mathbf{x}} = \mathbf{0}$ or the hyperplane $\mathbf{H}(\hat{\mathbf{x}}, \|\hat{\mathbf{x}}\|)$ separates $\bar{\mathbf{X}}$ from the origin.

However, for our investigation in this chapter, we applied an even more efficient recursive algorithm proposed in [Sekitani and Yamamoto, 1993] that does not require solving linear equations. Instead, it iterates over vertices and facets of the data polytope without repetition. Although the worst case complexity of both these MNP algorithms is $O(N^d)$, they usually converge quickly.

5.4 NAH-lsq: An Efficient Nearest Affine Hull Classifier

Most instance-based representations, such as Voronoi Diagrams, Convex Hull, Bounding Hyperdisk, and Bounding Hypersphere, rely on the notion of neighborhoods within the training data. Affine Hulls, however, give a loose approximation of class regions as they model each class as an affine subspace. Therefore NAH classification may prove a promising instance-based classifier in HDLSS settings where the notion of a neighborhood breaks down [Donoho and Tanner, 2005, Hall et al., 2005, Murtagh, 2009].

The existing SVD-based method to NAH classification discussed in Section 5.3.1 is space- and time- consuming. It requires storing a $d \times d$ matrix $\mathbf{P}_c$ for every class, which is unreasonable when d is large. For example, fitting an NAH-svd model for the HMDB activity dataset would require more than 80 GB of memory. An alternative solution is to store the $N_c \times d$ matrix $\mathbf{U}_c$ during training and to compute $\mathbf{P}_c$ during classification. This is still very demanding since the computing of $\mathbf{P}_c = \mathbf{U}_c\mathbf{U}_c^T$ requires efforts on the order of $O(d^2 N_c)$.

We propose an efficient least squares approach to NAH that exploits the QR factorization. Our approach builds on the observation that HDLSS training matrices $\mathbf{X}_c$ are of full rank since high dimensional data are vertices of a simplex and hence linearly independent [Hall et al., 2005, Donoho and Tanner, 2005]. We therefore propose to compute NAH classification in an entirely lazy fashion by finding a point $\mathbf{x}^*$ in the affine hull that is closest to $\mathbf{x}_q$. This requires solving

$$\begin{aligned} &\min_{\boldsymbol{\alpha}_c} \|\mathbf{x}_q - \mathbf{X}_c\boldsymbol{\alpha}_c\|^2 \\ &s.t. \sum_i \alpha_{ci} = 1,\ i = 1, 2, ..., N_c \end{aligned} \quad . \tag{5.9}$$

This problem does not involve inequality constraints and can therefore be cast as a simple least squares problem

$$\begin{bmatrix} \mathbf{X}_c \\ \mathbf{1} \end{bmatrix} [\boldsymbol{\alpha}] = \begin{bmatrix} \mathbf{x}_q \\ 1 \end{bmatrix} \tag{5.10}$$

where $\mathbf{1}$ is a row vector of all 1s of dimension N_c. Then, the distance from the query point is

$$d(\mathbf{x}_q, \boldsymbol{\Phi}_c^{aff}) = \|\mathbf{x}_q - \mathbf{x}^*\| \tag{5.11}$$

Again, in HDLSS settings, data matrices are (nearly) of full rank so that a stable solution of the system in (5.10) can be computed efficiently using the QR factorization. Compared to the computational complexity of NAH-svd $O(d^2 N_c)$, our approach of NAH-lsq involves significantly reduced computational effort, i.e. $O(dN_c^2)$. Note that for the datasets considered in this chapter, $d \approx N_c^2$.

Dataset	Type	#samples	#classes	dim. of features
Ikizler	Image	2,458	5	13,312
MHAD-skl	Mocap	660	11	2,400
MHAD-mhg	Depth	660	11	2,000
HMDB	Video	6,766	51	14,965
UCF50	Video	6,681	50	14,965

Table 5.2: Human activity recognition benchmark datasets

5.5 Datasets and Features

In this work, we consider 4 well-known datasets containing unconstrained videos, images or Kinect depth data (Table 5.2). Note that a popular trend in human activity recognition in the wild is to use multiple feature descriptors such as motion cues, pose, and scene context information. These methods use either early fusion of feature descriptors or late fusion of ensemble classifiers. Since the analysis of ensemble classifiers or multiple features is beyond the scope of this chapter, we restrict our practical experiments to recent single descriptors which are known to show good performance on these datasets. Below we give details of the datasets and how we performed the feature extraction on each dataset.

5.5.1 HMDB and UCF50 Video Datasets

HMDB [Kuehne et al., 2011] is one of the largest and most versatile datasets for action recognition in videos. It contains 6,766 video sequences of 51 action categories such as facial action, body movement, and human interaction. The UCF50 data [Reddy and Shah, 2013] consists of 6,681 real-world videos retrieved from YouTube that shoe of 50 action categories. For all 50 categories, the videos are split into 25 groups. For both of these datasets, we used action bank templates [Sadanand and Corso, 2012] as features for classification since they have shown very good performance in combination with linear SVM classification. Action bank feature extraction is based on spotting several motion templates in the multiscale spatio-temporal cuboids – resulting in a $14,965$ dimensional feature.

5.5.2 Ikizler Image Dataset

The Ikizler action dataset [Ikizler-Cinbis et al., 2009] contains 2458 still images downloaded from the internet. The images show five different human actions: dancing, playing golf, sitting, running, and walking. The dataset represents a hard challenge as it requires coping with a wide range of pose variations ranging from actions like dancing to sitting. We operated on the processed version of the dataset with cropped images of aligned human postures with respect to head position. Still, many of the training examples provided suffer from severe occlusions and invisible body parts. We used VLFeat [Vedaldi and Fulkerson, 2008] to extract local SIFT descriptors over multiple scales which were used to build a code book of size $1,024$. Each image was then divided into a three-level spatial pyramid and quantization was carried on each grid component – resulting in $13,312$ dimensional feature vector.

5.5.3 Berkeley MHAD Datasets

The recently introduced Berkeley Multimodal Human Action Database (MHAD) [Ofli et al., 2013] consists of 660 sequences of 11 actions performed repeatedly by 12 people and captured by multiple sensors. In our experiments, we used two modalities: skeleton information from a motion capture system and depth information from a Kinect sensor. In each case, we divided the video into N_s overlapping temporal segments or windows. In addition, we adapted the popular Bag-of-Features (BoF) approach to quantize all frames within a temporal window. To this end, we built a vocabulary of N_w skeletal or visual words using k-means. Finally, every action sequence was represented as a vector of length $K = N_s \times N_w$.

Mocap Data

The MHAD-Mocap data was acquired by tracking the relative $3D$ positions of 43 LED markers placed on different body parts and joints. Consequently, we represented each activity as a sequence of a skeletal vector of length 129. First, we sampled $100,000$ skeletal vectors from all the data and built a vocabulary of $N_w = 60$ skeletal words. We set $N_s = 40$, and as a result represented each action sequence by a vector of length $2,400$.

Kinect Depth Data

The Kinect-based depth videos divided into 8 disjoint Depth-Layered Multi-Channel (DLMC) are provided by the owners of the MHAD dataset. We used only channel C-3 DLMC videos since almost all the subjects and their movements lie within this depth range. Our feature representation for this modality is based on motion histograms features similar to [Escalante et al., 2013]. Given a sequence of gray scale depth images $I = I_1, I_2, ...I_n$, a set of motion energy images $D = \{D_1, D_2, ...D_{(n-1)}\}$ is obtained where $D_i = I_{(i+1)} - I_i$. Each difference image is divided into a grid with equal number of cells and average motion energy is estimated for each cell. The 2D grid of motion energies is transformed into a vector of length N_b, where N_b is the number of cells in the grid. We considered 20×20 sized cells, so $N_b = 768$. For BoW representation, we sampled $100,000$ feature vectors from all the data and built a vocabulary of $N_w = 100$ words. Finally we set $N_s = 20$ to represent each action sequence by a vector of length $2,000$.

5.6 Experimental Results and Discussion

We evaluate the NAH-lsq and other geometric classification methods (discussed in Section 5.3) and compare their performance to traditional approaches such SVM, kNN, and LDA. For the two parametric methods kNN and SVM, we report the best results over choices of hyper-parameters. For linear SVMs, the optimal penalty parameter was determined within the range between 10^{-5} and 10^5. For kNN classifiers, the parameter k was varied in the range from 1 to N_c. All methods discussed in Sections 5.3 and 5.4 were implemented in Python. For other methods, we used implementations and wrappers provided by the Python-based machine learning library Scikit-learn[8]. All experiments were carried out on a PC with 16GB RAM using a single core.

For each dataset, we adopted the popular cross validation scheme and report average results over all iterations. For HMDB, we used the original three train-test splits [Kuehne et al., 2011], where in each case for each action, 70 videos were used for training and another 30 for testing. For UCF50, we applied a 5-fold Leave-Five-Groups-Out approach, and for MHAD datasets we use the

[8] http://scikit-learn.org

	Ikizler	MHAD-skl	MHAD-mhg	HMDB	UCF50
kNN	42.89	75.45	67.72	14.81	36.20
LDA	53.61	65.6	80.30	22.37	15.93
MNP	52.5	75.76	79.24	3.51	9.87
NCH	52.0	76.96	81.51	23.59	49.99
NHD-svd	53.98	77.12	79.84	5.86	9.87
NHD-lsq	53.72	77.12	79.54	5.80	9.73
NAH-svd	55.25	77.12	81.81	27.02	55.88
NAH-lsq	54.94	**77.12**	81.67	**27.08**	55.94
SVM	**56.05**	75.45	**84.54**	25.23	**57.58**

Table 5.3: Accuracy (%) of different classifiers on high dimensional activity recognition

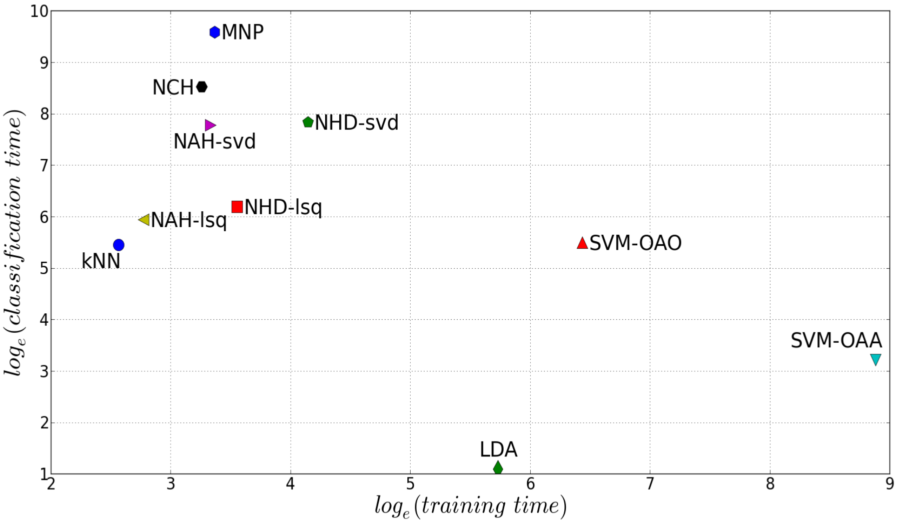
Leave-One-Actor-Out scheme. For the Ikizler dataset, we randomly sampled 100 images from each class for training, while the rest were used for testing.

5.6.1 Recognition Accuracy

Table 5.3 compares recognition accuracies obtained from the different classifiers tested. While NHD exhibits good performance on some datasets, it does poorly when the number of classes is high. The piecewise boundaries of NCH also do not generalize well enough to compete with SVMs. All instance-based methods that bound the class regions (kNN, NCH, MNP, and NHD) suffer from a loss of performance in one way or another. On the other hand, NAH methods, which represent classes as an (unbounded) affine subspaces, are least affected by artifacts due to high dimensional neighbourlessness. Notice NAH-lsq achieves an accuracy similar to NAH-svd, indicating the stability of rather efficient QR factorization on HDLSS matrices. In short, non-parametric NAH-based classifiers outperform other instance-based methods in all cases and show better results than the optimal SVM in some cases.

5.6.2 Efficiency

Instance based methods usually do not require any training and model fitting is deferred to the classification phase. Other classifiers such as LDA, SVM, and Decision Trees / Random Forests explicitly learn a model from the training data in an offline manner. Consequently, these approaches are efficient during classification. In those cases, the classification time also depends on the complexity of the decision surface. For example for C number of classes,



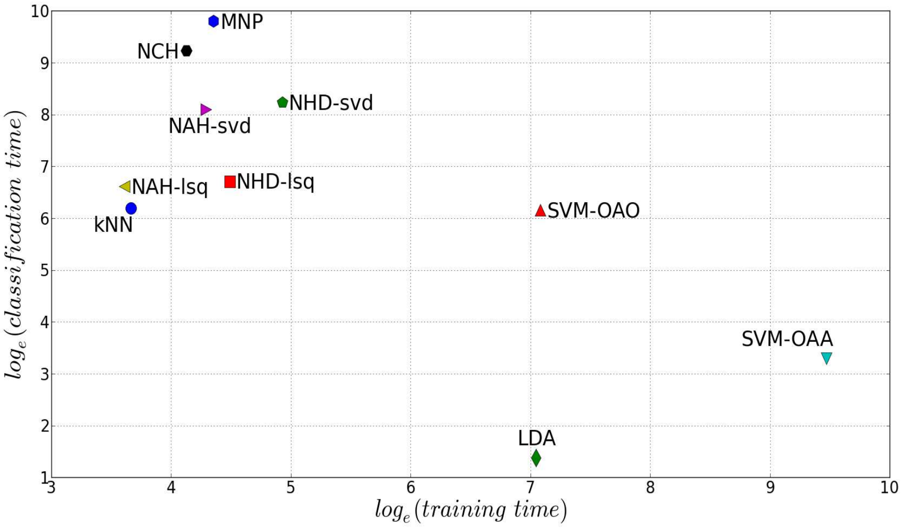


Figure 5.5: Overall training and test times on HMDB (above) and UCF50(below) datasets

the decision surface of SVM-OAO will have $C(C-1)/2$ different d-dimensional hyperplanes, whereas SVM-OAA will have only C hyperplanes. However, determining those C hyperplanes in SVM-OAA is often more time consuming than fitting an SVM-OAO since the complexity of SVM training is between $O(N^2)$ and $O(N^3)$ and the number of input examples N can be significantly large for an SVM-OAA.

Figure 5.5 plots overall logarithmic training and testing times (in CPU seconds)

	Local	Low rank	Param.	Dec. Surf.	Tr. Time	Te. Time	Acc.
kNN	✓	×	✓	●●●●●	●○○○○	●○○○○	●○○○○
LDA	×	✓	×	●●●○○	●○○○○	●○○○○	●●○○○
MNP	✓	×	×	●●●●○	●○○○○	●●●●●	●●●○○
NCH	✓	×	×	●●●●○	●○○○○	●●●●○	●●●●○
NHD-svd	✓	×	×	●●●○○	●●●○○	●●●●○	●●●○○
NHD-lsq	✓	×	×	●●●○○	●○○○○	●●●○○	●●●○○
NAH-svd	×	×	×	●○○○○	●●○○○	●●●●○	●●●●●
NAH-lsq	×	×	×	●○○○○	●○○○○	●●●○○	●●●●●
SVM-OAO	×	×	✓	●●●○○	●●●○○	●●●○○	●●●●●
SVM-OAA	×	×	✓	●○○○○	●●●●●	●○○○○	●●●●●

Table 5.4: A comparison of different classifiers on HDLSS activity data

for all methods on various datasets. The non-zero training time of instance-based methods NAH-lsq, NAH-svd, NHD, kNN, MNP and NCH reflects an overhead due to preprocessing (indexing and storage) of the training data. Notice that the proposed least squares approach NAH-lsq is more efficient than any other hull-based method and is competitive with kNN classification. In particular, NAH-lsq gains significantly over NAH-svd. It also shows a competitive classification time when compared to SVM-OAO. Table 5.4 summarizes several theoretical, structural, and empirical aspects of a number of classifiers. The last four columns are based directly on the complexity of decision surfaces, training and testing time, and accuracy on the 5 datasets.

5.6.3 Applicability in Online Learning

The empirical results discussed in the previous sections clearly indicate the effectiveness of NAH-lsq in high dimensional classification. The only competitive method is the classical SVM. The theoretical foundations, geometrical simplicity, and computational efficiency (model fitting) of NAH as compared to the more complex SVM illustrates the power of simple linear models in high dimensional data processing. In practice, we observed that almost all the training samples form the support of the decision surfaces. Table 5.5 shows the percentage (to the nearest integer) of training data of a class that is used to determine individual piecewise boundaries (point-hull and class-class) and overall decision surfaces. For example, in the case of the HMDB dataset, a single binary decision surface of SVM-OAO was observed to require, on average, 34% of the data of each of the two corresponding classes as support vectors and, for a given class, almost every instance became a support to one or more

Dataset	NCH	SVM		NAH
	point-hull	class-class	class-all	all
HMDB	12%	34%	98%	100%
UCF50	9%	26%	90%	100%

Table 5.5: Proportion of the training examples of a class used as support of decision surfaces

binary decision surfaces.

On the one hand, these observations confirm the lack of structure within the class regions. On the other hand, they justify the use of affine subspaces in classification. They also hint at an advantage of using NAH-lsq over SVM in online settings, since online SVMs would require (nearly) complete retraining when, at a previous time step, most data had been selected as support vectors. Let N be the number of instances, S be the number of support vectors, and $R \leq S$ be the number of support vectors such that $0 \leq |\alpha_i| \leq C$: if $N \equiv S \equiv R$, then according to [Bordes et al., 2005], the training time of online SVM (even by only adding a single example) is the same as that for a regular SVM. We empirically evaluated the performance of SVM combined with stochastic gradient descent optimization scheme, as suggested in [Bottou, 2010]. Below, we briefly describe SVM-SGD, an online SVM approach based on stochastic gradient descent.

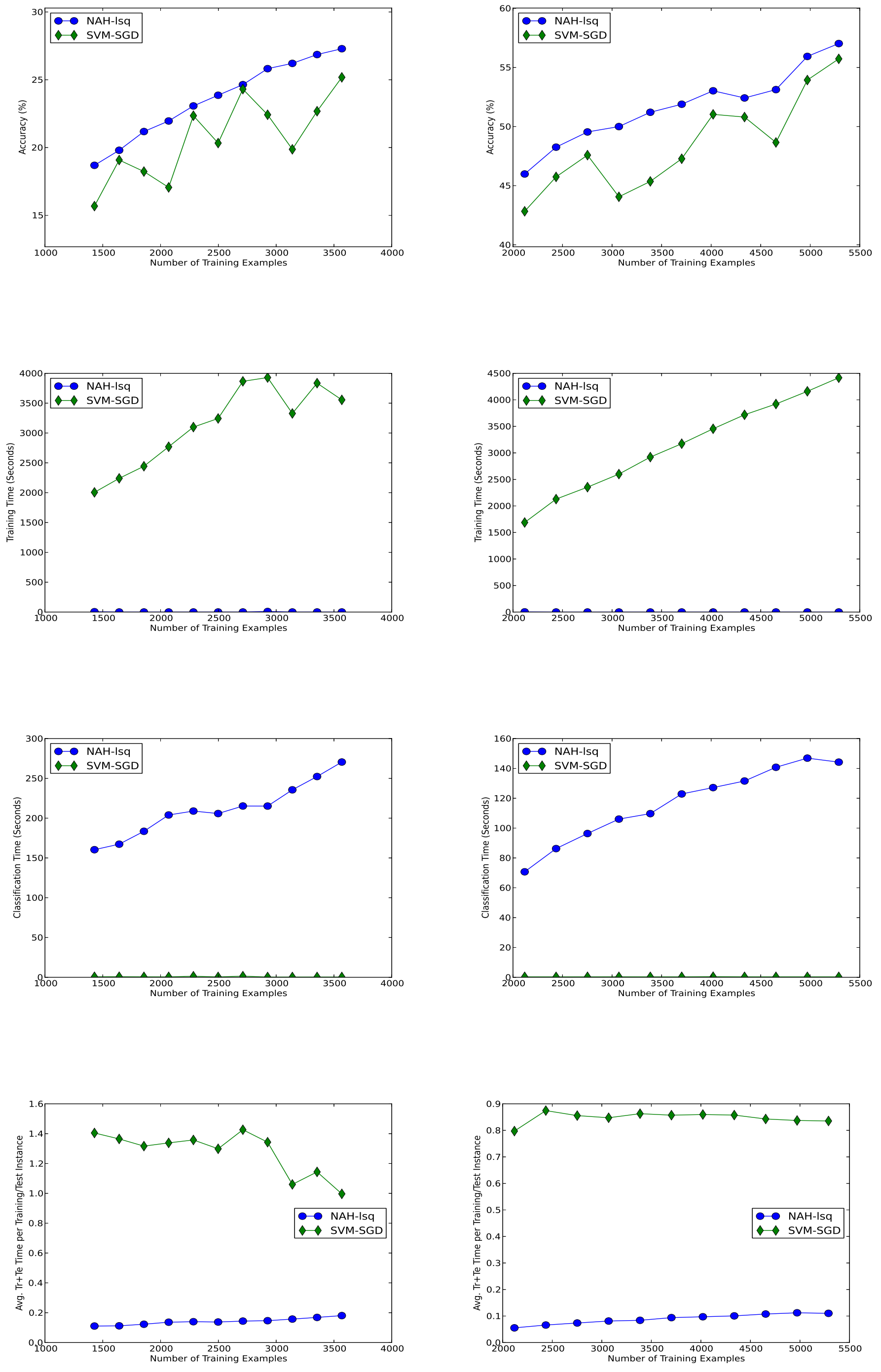
Given a training set $\left\{(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathcal{R}^d\right\}$, several supervised classification approaches aim to determine a decision function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ by minimizing an error function of the form:

$$E(\mathbf{w}, b) = \sum_{i=1}^{n} L(y_i, f(x_i)) + \alpha R(\mathbf{w}) \tag{5.12}$$

where L and R are loss and penalty functions, respectively. In the case of SVMs, the error function is composed of the Hinge loss function and the $L2$ norm of $\mathbf{w}$. A stochastic gradient descent algorithm is an iterative procedure to determine the optimal $\mathbf{w}$ by applying the following update rule:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \left\{ \alpha \frac{\partial R(\mathbf{w})}{\partial \mathbf{w}} + \frac{\partial L(\mathbf{w}^T x_i + b, y_i)}{\partial \mathbf{w}} \right\} \tag{5.13}$$

where η is the learning rate.



HMDB UCF50

Figure 5.6: Comparison of NAH and SVM-SGD in online settings on and HMDB (left) and UCF50 (right) dataset in terms of (first row) accuracy, (second row) overall training time, (third row) overall classification time, and (last row) average model fitting time per instance

Stochastic gradient descent based optimization, in combination with popular classifiers such as SVMs and CRFs, has gained popularity recently for its applicability to large-scale online classification problems [Bottou, 2010, Zhu et al., 2009]. We compared the performance of the NAH-lsq and one-against-all SVM-SGD in online settings for two large datasets: HMDB and UCF50. For HMDB, we used the split-1 [Kuehne et al., 2011] for training and testing. For UCF50, we chose videos belonging to all groups from $g6$ to $g25$ for training and the rest for testing. We trained each classifier initially on 40% of the training data and subsequently added equal amounts of the remaining data in 10 episodes. While NAH-lsq has to explicitly fit a new model on arrival of new data, SVM-SGD takes a *warm start* from the current optimal $\mathbf{w}$ and iterates until convergence or until a maximum number of iterations have been executed. The optimal hyper-parameters of SGD were determined offline through cross validation.

The results are presented in Fig. 5.6. In each case, the first row shows that the performance of NAH-lsq consistently improves with an increase in training data while the performance of SVM-SGD may occasionally suffer, e.g. when the distribution of new data points differs significantly from the one for which the previous solution $\mathbf{w}$ was estimated. The second and third rows plot training and test time respectively after each episode. While NAH-lsq spent most of their time on lazy classification, adding new data almost always caused a complete retraining of SVM too. Usually, in online settings, the model estimation costs are not distinguished for training and classification phases. The fourth row in Fig. 5.6 plots combined average time, per instance, for model fitting in training or classification phases. It can be seen that the NAH-lsq emerges to outperform SVM-SGD in terms of efficiency.

5.7 Conclusions and Future Directions

In this chapter, we investigated an important but overlooked aspect of human activity recognition in large unconstrained databases. In particular, we examined the underlying geometry of class regions for HDLSS classification of real-world human activities by employing several geometric classifiers. For such scenarios, we empirically affirm the lack of neighborhoodness in HDLSS activity data. On the one hand, neighborliness negatively affects the performance of neighborhood-based classification methods such as kNN, NCH, or NHD; on the other, it causes over fitting artifacts for SVMs. Our results show that representing each class as an affine subspace spanned by its members remedies this situation. We propose NAH-lsq, a least square- and QR factorization-based approach that significantly reduces time and memory requirements of the existing NAH methods. Consequently, NAH-lsq appears to be a suitable choice for multiclass HDLSS activity recognition as it provides parameter-free models, yields smooth decision surfaces, achieves high accuracy; and is efficient. On several challenging datasets, we found the NAH-lsq classifier to show competitive or superior performance than the the widely used SVMs. Moreover, despite its lazy classification approach, NAH-lsq appears to be a faster approach when compared to SVMs in online settings (a major application area). To conclude, we (a) provided an empirical insight into the complexity of the multiclass HDLSS activity recognition problem and (b) proposed a simple yet powerful solution.

Although we have observed competitive performance of NAH-lsq as compared to SVM regardless of the choice of svm kernel, it will be interesting to see in future how kernelization can improve the performance of NAHs. In fact, it is not clear whether kernelization helps SVM-based classification of HDLSS data either. To this end, our preliminary experiments on the UCF50 and HMDB datasets show that for a given feature descriptor, linear SVMs achieve an accuracy similar to that of non-linear SVMs. Other possible directions of future work are modeling joint affine subspaces by using multiple feature descriptors and clustering data points represented as affine subspaces [Hu et al., 2012, Lee and Schulman, 2013].

Summary

Currently, most research on human activity recognition in unconstrained data focuses on developing ever more complex features and naively choosing off-the-shelf classifiers. This may not be suitable when efficiency and model flexibility are of concern. To this end, we investigated the popular discriminative and not-so-popular geometrical classifiers in the context of high dimension, low sample size (HDLSS) human activity recognition and proposed an efficient least squares and QR factorization-based approach to Nearest Affine Hull classification. Through extensive experimentation on 5 benchmark datasets, we showed that the proposed parameter-free NAH-lsq achieves recognition accuracy as high as SVM and NAH-SVD and is much faster. It also turns out that NAH-lsq is most effective for on-line classification of large-scale data where SVMs would need expensive retraining. In short, this chapter (a) discussed issues that, to our knowledge, have not yet been studied by the human action recognition community (b) presented the NAH-lsq method that is distinguished by its high recognition accuracy, parameter-free modeling, efficiency and applicability to intended computer vision applications.

Chapter 6

Human Activity Recognition in 3D by Separating Style and Content

Recent advances in 3D sensor technology, such as the invention of Microsoft Kinect sensors, has boosted low-cost imagery capture in the form of different modalities. Accordingly, there is a boost in research on human activity recognition in emerging environments. On the one hand, the underlying 3D pose and motion information has facilitated reliable action recognition (e.g. [Wang et al., 2012, Ofli et al., 2013]). On the other hand, it invites researchers to think of out-of-the-box applications, e.g. physiotherapy exercises, interactive gaming, and home security. Most existing approaches, however, have focused only on a single aspect – recognition of actions.

This chapter introduces our approach to a new direction of research in human activity recognition. It builds on studies in psychophysics [Cutting and Kozlowski, 1977, Thoroughhman and Shadmehr, 1999] which suggest that people tend to perform different actions in their own style. Specifically, we handle the novel issue of recognizing human actions and the underlying execution styles (actors) in 3D videos using motion dynamics only. We propose a hierarchical approach that is based on conventional action recognition and asymmetrical bilinear factorization. In particular, we apply bilinear decomposition on the tensorial representation of action videos to characterize styles of performing different actions. The proposed approach is solely based on the dynamics of the underlying action. Our model is evaluated on the Inria-IXMAS and the Berkeley-MHAD action datasets using different modalities based on optical motion capture, Kinect depth videos, and 3D motion history volumes. The proposed approach achieves high recognition accuracy in comparison to alternate methods, i.e. Nearest Neighbor classification and symmetric bilinear modeling. Our approach is not only directly applicable to interactive 3D environments and surveillance systems, but can also work as a baseline for future research towards multifactor activity analysis in unconstrained videos.

6.1 Introduction

Most existing research on human activity analysis has focused on the very single aspect i.e. recognition of human actions. However, people tend to perform different actions and activities such as walking, kicking and cooking in their own personal style. Prominent studies in psychophysics and biomechanics have shown that individuals build specific internal models for different movements and they can be recognized solely from their motion dynamics [Cutting and Kozlowski, 1977, Thoroughhman and Shadmehr, 1999]. In this line, corresponding research on vision-based gait recognition has shown great success in the last decade [Wang et al., 2010].

A significant development in recent years is the availability of 3D data through low-cost image capturing. Compared to projected data in monocular videos, 3D videos are becoming increasingly popular for their robustness against (self-) occlusion and rich pose and motion information. Such information may be of great impact for non-conventional activity analysis, e.g. when modeling both inter-class and intra-class variations.

In this chapter, we are interested in determining *if it is possible to recognize both the actions and the actors in 3D sequences using motion dynamics only.* This problem is related to the well-known issue of *separating style from content* in areas such as handwriting or face recognition. We treat observed actions as resulting from a generative process with two factors, namely actor (style) and action (content). We use bilinear factorization to model underlying phenomena since bilinear models immediately lend themselves towards two-factor classification and since they can be efficiently determined through singular value decomposition (SVD).

Conventional symmetric bilinear models assume independence between content and style factors (e.g. face and illumination). This is not the case in motion-based action recognition since both actions and execution styles are based on the variation of the same cue, i.e. motion. Due to challenges posed by the high level of articulation of human bodies, a conventional symmetric

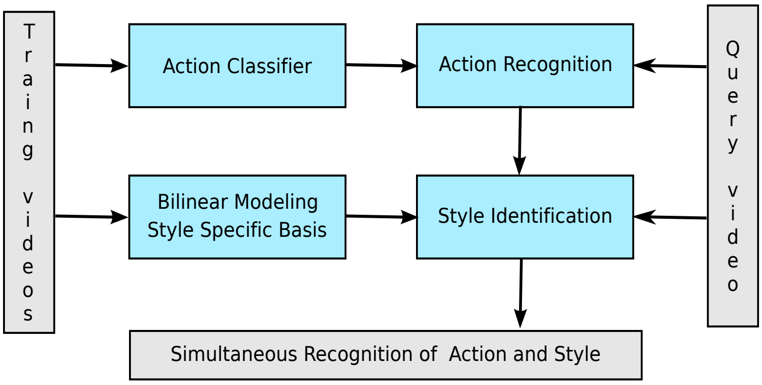


Figure 6.1: Block diagram of our hierarchical bilinear approach

bilinear model would not suffice to separate content and style in human activity videos. We therefore use a two-step approach to classify a given test video (query video). In the first step, we apply a classical action classification to predict the underlying action of the query video. In the second step, we use this prediction to generate a style-specific basis for the query video using an asymmetric bilinear model. Finally, we compare this basis with the style-specific basis learned from training data in order to identify the most likely style.

Figure 6.1 gives an overview of our approach. For experimental evaluation, we consider two multi-actor, multi-action datasets namely the Inria Xmas Motion Acquisition Sequences (IXMAS) [Weinland et al., 2006] and the Berkeley Multimodal Human Action Database (MHAD) [Ofli et al., 2013]. We show that, compared to naive nearest neighbor classification and symmetric bilinear modeling, the proposed hierarchical model significantly improves results for different motion cues. Consequently our approach extends motion-based person identification to multiple common actions and shows that the identification is not limited to walking or running actions. To the best of our knowledge, our approach is the first at such an attempt in the context of human action recognition.

The remainder of this chapter is organized as follows: Section 6.2 briefly discusses related research on separating style and content. Section 6.3 reviews the basics of bilinear models. Section 6.4 presents how we deal with action recognition by using nearest neighbor classifiers and asymmetric bilinear models. Section 6.5 reports details on our benchmark data, experiments, and results.

Finally, Section 6.6 concludes the work.

6.2 Related Work

Separating style and content has been of great interest for the recognition of speech, handwriting, and faces since the idea was pioneered by Tanenbaum and Freeman [2000]. There, the authors employed bilinear modeling and showed promising results on classical problems such as handwritten character, face, or pose recognition. Chaung and Bregler [2005] used bilinear factorization to separate emotional styles from speech content in order to create expressive facial animations. Shin et al. [2008] proposed an efficient approach to "illumination-robust" face recognition, based on symmetric bilinear modeling, by separating an identity factor and an illumination factor.

Despite a great deal of research on human action recognition [Laptev et al., 2008, Marszalek et al., 2009, Wang et al., 2011, Reddy and Shah, 2013], hardly any efforts have yet been made to separate style from content. Most of the existing work in this direction deals with person identification for a single action [Elgammal and Lee, 2004, Cuzzolin, 2006, Perera et al., 2009]. Elgammal and Lee [2004] applied a non-linear model for separating poses from walking patterns of individuals; Cuzzolin [2006] used bilinear separation models for different gait gestures.

The approach presented in [Yam et al., 2002] was the first to consider styles of running in recognizing individuals. Perera et al. [2009] employed multifactor tensor decomposition to identify different styles of the dancing action using motion capture data. Recently, Iosifidis et al. [2011] trained person-specific activity classifiers to improve recognition of different human actions. The issue of varying styles for human activities has been discussed by Taralova et al. [2011], who presented a source-constrained clustering approach to accommodate different sources (e.g. actors). However, their focus is on clustering from *known* sources and not on identifying the sources.

6.3 Bilinear Models

In this section, we review basic concepts of bilinear models for separating style from content; our terminology is similar to that used by Tanenbaum and Freeman [2000]. A bilinear model is a generative model where each K dimensional observation $\mathbf{y}$ in a style $s \in [1, 2, ..., S]$ and content class $c \in [1, 2, ..., C]$ is given in the form:

$$y_k^{sc} = \sum_{i=1}^{I} \sum_{j=1}^{J} w_{ijk} a_i^s b_j^c \quad , k \in [1, 2, ..., K] \tag{6.1}$$

where $\mathbf{a}^s$ and $\mathbf{b}^c$ are I and J dimensional coefficient vectors representing style s and content c and the entries w_{ijk} govern the interaction between the two underlying factors. Let $\mathbf{W}_k$ represent k^{th} matrix of dimension $I \times J$ then Eq. (6.1) becomes:

$$y_k^{sc} = \mathbf{a}^{sT} \mathbf{W}_k \mathbf{b}^c \tag{6.2}$$

The matrices $\mathbf{W}_k$ define bilinear mapping from content and style space to the K dimensional observation space. The model in Eq. (6.1) and Eq. (6.2) is called the *symmetric bilinear model.*

While the symmetric model assumes the independence of the interaction terms w_{ijk} w.r.t style and content classes, the *asymmetric bilinear model* lets these terms vary with one of the factors (by convention with style) and thus allows for more flexibility. For instance, with a style-specific basis $\mathbf{a}_{jk}^s = \sum_i a_i^s w_{ijk}$, Eq. (6.1) becomes:

$$y_k^{sc} = \sum_{j=1}^{J} \mathbf{a}_{jk}^s b_j^c \tag{6.3}$$

Equivalently in matrix notation we write:

$$\mathbf{y}^{sc} = \mathbf{A}^s \mathbf{b}^c \tag{6.4}$$

such that $\mathbf{A}^s$ denotes $K \times J$ matrix with entries $\mathbf{a}_{jk}^s$. Here, $\mathbf{A}^s$ represents a style-specific map from the content space to the observation space.

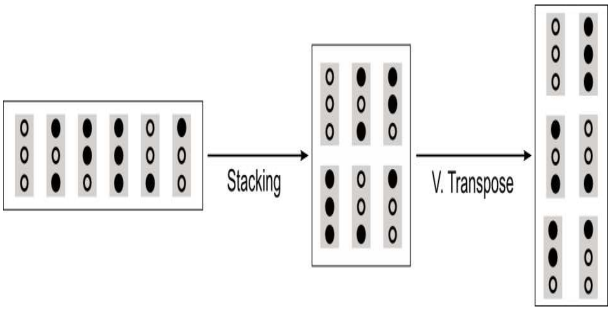


Figure 6.2: Stacked representation of a tensor and its vector transpose (VT)

6.3.1 Training an Asymmetric Bilinear Model

Let $\mathbf{y}(t)$ denote the t^{th} training sample $(t = 1, \ldots, T)$ and let $\boldsymbol{\chi}_{sc}(t)$ be a characteristic function such that $\boldsymbol{\chi}_{sc}(t) = 1$ if $\mathbf{y}(t)$ has style s and content c and 0 otherwise. Then, the sum of squared errors $\mathbf{E}$ for the asymmetric model over all training data is given by

$$\mathbf{E} = \sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{c=1}^{C} \boldsymbol{\chi}_{sc}(t) \left\| y(t) - \mathbf{A}^s \mathbf{b}^c \right\|^2. \tag{6.5}$$

Fitting an asymmetric model aims at finding solutions for $\mathbf{A}^s$ and $\mathbf{b}^c$ that minimize $\mathbf{E}$. If a given sample of training data consists of nearly equal numbers of observations for each style and content (as in the case of this chapter), a closed form procedure can be adopted from using SVD.

Let $\overline{\mathbf{y}}^{sc}$ denote the *mean* of all observations in style s and content c. The training set can be thought of as a 3^{rd} order tensor $\overline{\mathbf{Y}}_{S\times K\times C}$. For making efficient use of matrix algebra, $\overline{\mathbf{Y}}$ is represented as a stacked matrix with dimensions $(SK) \times C$ such that each of C columns contains S parts of $K \times 1$ vectors. Further, the *vector transpose* operation VT is defined for stacked matrices as follows: the vector transpose of an $(AK) \times B$ matrix $\mathbf{Q}$ is a $(BK) \times A$ matrix $\mathbf{Q}^{VT}$ such that the (l, m) entry of $\mathbf{Q}$ becomes the $(mK + mod(l, K), l)$ entry of $\mathbf{Q}^{VT}$ (See Fig. 6.2).

For training data in stacked matrix form, the asymmetric model can then be expressed as $\overline{\mathbf{Y}} = \mathbf{AB}$, such that $\mathbf{A} = \left[\mathbf{A}^1 \ldots \mathbf{A}^S\right]'$ is a $(SK) \times J$ matrix of style-specific basis and $\mathbf{B} = \left[\mathbf{b}^1 \ldots \mathbf{b}^C\right]$ is a $J \times C$ is matrix of content parameters. A least squares optimal solution is obtained by computing the SVD of $\overline{\mathbf{Y}}$ such that $\overline{\mathbf{Y}} = \mathbf{U\Sigma V}^T$. The style-specific basis matrix $\mathbf{A}$ is obtained from the first J columns of US and the content parameter matrix $\mathbf{B}$ is defined by the first J rows of V^T.

6.3.2 Training a Symmetric Bilinear Model

The sum of squared errors for the symmetric model in Eq. (6.2) is

$$\mathbf{E} = \sum_{t=1}^{T}\sum_{s=1}^{S}\sum_{c=1}^{C}\sum_{k=1}^{K} \boldsymbol{\chi}_{sc}(t) \left\| y_k(t) - \mathbf{a}^{s^T}\mathbf{W}_k\mathbf{b}^c \right\|^2. \tag{6.6}$$

To solve this optimization problem, asymmetric modeling through SVD is iterated by alternatively switching the roles of content and style and an expectation maximization (EM) approach is used to simultaneously update parameters of style and content. This process is based on the following relationship where the the symmetric model is given as

$$\overline{\mathbf{Y}} = \left(\mathbf{W}^{VT}\mathbf{A}\right)^{VT}\mathbf{B} \tag{6.7}$$

or equivalently

$$\overline{\mathbf{Y}}^{VT} = \left(\mathbf{WB}\right)^{VT}\mathbf{A} \tag{6.8}$$

where $\mathbf{W}$, $\mathbf{A}$, and $\mathbf{B}$ are $(IK) \times J$, $I \times S$, and $J \times C$ matrices, respectively.

An EM algorithm is used to iteratively update estimates of $\mathbf{A}$ and $\mathbf{B}$ (See Algorithm 2). The procedure starts by initializing $\mathbf{B}$. From the orthogonality of $\mathbf{B}$ and Eq. (6.7), we derive $(\overline{\mathbf{Y}}\mathbf{B}^T)^{VT} = \mathbf{W}^{VT}\mathbf{A}$. Now, the SVD of $(\overline{\mathbf{Y}}\mathbf{B}^T)^{VT} = \mathbf{U\Sigma V}^T$ is computed and the estimate for $\mathbf{A}$ is updated to be the first I rows of $\mathbf{V}^T$. Since $\mathbf{A}$ is orthogonal, Eq. (6.8) yields $(\overline{\mathbf{Y}}^{VT}\mathbf{A}^T)^{VT} = \mathbf{WB}$. This estimate of $\mathbf{A}$ is used for the SVD of $(\overline{\mathbf{Y}}^{VT}\mathbf{A}^T)^{VT} = \mathbf{U\Sigma V}^T$ and $\mathbf{B}$ is updated to be the first J rows of $\mathbf{V}^T$. This completes one iteration of the EM procedure. Upon convergence, the basis vectors are computed as

$$\mathbf{W} = \left(\left(\overline{\mathbf{Y}}\mathbf{B}^T\right)^{VT}\mathbf{A}^T\right)^{VT}. \tag{6.9}$$

Algorithm 2 Fitting a symmetric bilinear model

```
Initialize $\mathbf{B}$ using asymmetric assumption
while Not converged do
    From orthogonality of $\mathbf{B}$ and (6.7), we have $(\overline{\mathbf{Y}}\mathbf{B}^T)^{VT} = \mathbf{W}^{VT}\mathbf{A}$ and SVD of $(\overline{\mathbf{Y}}\mathbf{B}^T)^{VT} = \mathbf{U\Sigma V}^T$
    Set $\mathbf{A}$ equal to the first $I$ rows of $\mathbf{V}^T$
    From orthoganility of $\mathbf{A}$ and (6.8), we have $(\overline{\mathbf{Y}}^{VT}\mathbf{A}^T)^{VT} = \mathbf{WB}$ and SVD of $(\overline{\mathbf{Y}}^{VT}\mathbf{A}^T)^{VT} = \mathbf{U\Sigma V}^T$
    Set $\mathbf{B}$ equal to the first $J$ rows of $\mathbf{V}^T$.
end while
Set $\mathbf{W} = \left((\overline{\mathbf{Y}}\mathbf{B}^T)^{VT}\mathbf{A}^T\right)^{VT}$
```

6.4 Our Approach

Asymmetric bilinear models do not enforce independence among the factors and therefore allow more flexibility if one of the factors is known. On the other hand, symmetric models do not assume any dependency or prior knowledge as to one of the factors and simultaneously update content and style parameters. In the literature, symmetric bilinear models have been successfully applied to several domains, such as separating emotional speech styles from facial expressions [Chaung and Bregler, 2005], jointly modeling body shapes and gait motion [Elgammal and Lee, 2004], and to separate identity factor from illumination factor for robust face recognition [Shin et al., 2008]. Since the style and content factors, e.g. face and illumination, are obviously independent in those cases, symmetric modeling achieves good results. In contrast, for the problem of recognition of human action and the execution style, the content and the style are based on the same generative process – human motion. Consequently, we observed a low performance by symmetric models for our task (Section 6.5).

This motivated us to develop a two-stage procedure where one factor class is identified in each stage and the estimation from the first stage informs the classification in the second stage. We empirically evaluated the individual *discriminativeness* of the two factors by implementing separate single-label classification using NN, i.e. by considering the problem as being either action-

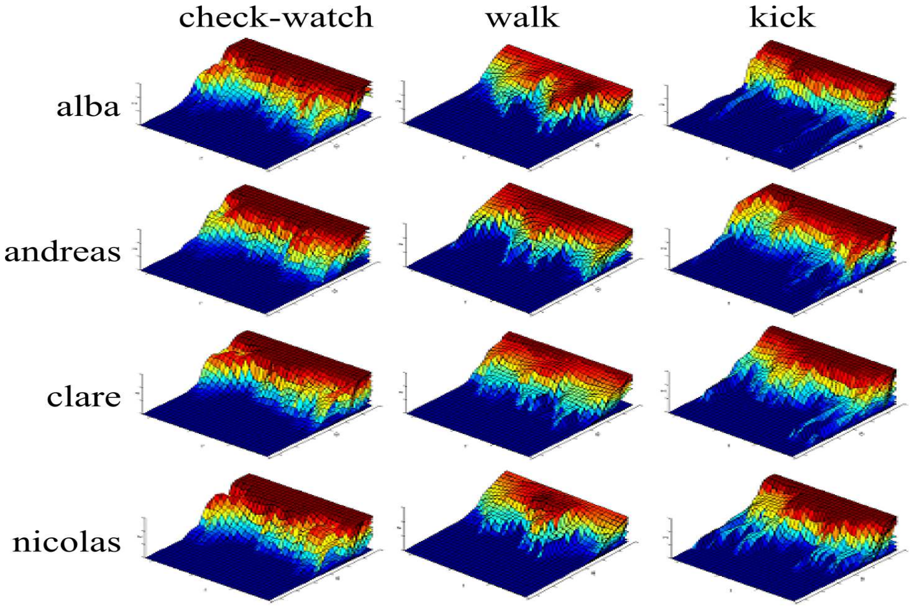


Figure 6.3: 3D color plot in cylindrical coordinates (r, θ, z) of average Motion History Volumes of 3 actions each performed 3 times by 4 different actors in IXMAS dataset

or actor-recognition. We observed high accuracy for action classification as compared to actor classification. This is also obvious from Fig. 6.3, which shows 3D color plots in cylindrical coordinates of average Motion History Volumes of some of the actions and actors in the IXMAS action data. It is clear from the figure that discrimination is more evident among actions (columns) than among styles (rows).

We, therefore apply an action recognition module in the first stage of our system and use its output as the input (predicted content parameters) to the second module, which is based on asymmetric bilinear models with actor-specific bases. The second module classifies the style of the query observation by using the learned basis and the predicted content class. Here, we model the training data as follows

$$\overline{\mathbf{Y}}_{train} = \mathbf{A}_{(SK)\times J}\mathbf{B}_{J\times C} \tag{6.10}$$

where $\mathbf{A} = \left[\mathbf{A}^1 ... \mathbf{A}^S\right]'$ and $\mathbf{B} = \left[\mathbf{b}^1 ... \mathbf{b}^c\right]$ are obtained as described in Section 6.3.1.

During classification, the action-recognition module predicts an action class $\tilde{c}$

for the $K \times 1$ query observation $\tilde{\mathbf{y}}$. In the second step, we use $\tilde{\mathbf{y}} = \tilde{\mathbf{A}}\mathbf{b}^{\tilde{c}}$, where $\mathbf{b}^{\tilde{c}}$ is $\tilde{c}$-th column of $\mathbf{B}$ (see Eq. 6.10) to determine $\tilde{\mathbf{A}}$, i.e. the style of $\tilde{\mathbf{y}}$. To this end, we compute

$$\tilde{\mathbf{A}} = \tilde{\mathbf{y}} \times (\mathbf{b}^{\tilde{c}})^{\dagger} \tag{6.11}$$

where $(\mathbf{b}^{\tilde{c}})^{\dagger}$ is the Moore-Penrose pseudo inverse of $\mathbf{b}^{\tilde{c}}$.

Finally, we compare the $(1K) \times J$ style matrix $\tilde{\mathbf{A}}$ for each of the S chunks of $\mathbf{A}$ and select $\tilde{s}$ such that $\text{argmin}_s \, |\tilde{\mathbf{A}} - \mathbf{A}^s|$; this procedure yields an optimal label-pair $(\tilde{c}, \tilde{s})$ for $\tilde{y}$.

6.5 Data and Experiments

To evaluate our approach, we consider two 3D datasets: Inria Xmas Motion Acquisition Sequences (IXMAS) [Weinland et al., 2006] and Berkeley Multimodal Human Action Database (MHAD) [Ofli et al., 2013]. IXMAS is a popular multi-actor and multi-view dataset, for which features based on 3D motion history volumes have shown good results for action recognition. Whereas the recently proposed MHAD consists of data captured from different sensors, including optical motion capture (mocap) system, Kinect depth sensors, multiview stereo cameras, wearable accelerometers and microphones. In both of the datasets, each subject performs every action multiple times (runs). Multiple executions of the same action in the same environment allow us to focus on the question of whether humans have unique styles for executing actions? For both datasets, we provide an empirical insight by evaluating the individual *discriminativeness* of the two factors by implementing single-label classification using NN approaches, i.e. by considering the problem as being either action or actor recognition. We observe (as expected from Fig. 6.3) high accuracy for action classification as compared to actor classification in all cases.

For the multi-label classification problem, i.e. action-actor recognition, we compare our approach with pure symmetric modeling and with NN classification. In these settings, a query instance with the actual label (c, s) and the predicted label $(\tilde{c}, \tilde{s})$ is considered a true positive only if $(c, s) = (\tilde{c}, \tilde{s})$. Accordingly, all accuracies are based on the correct classification of the action-actor pairs. Recall that while the symmetric bilinear modeling approach predicts the label $(\tilde{c}, \tilde{s})$ in a single step, our hierarchical approach first determines $\tilde{c}$

and then uses it to predict $\tilde{s}$. The NN adaptation to action-actor classification is also achieved in two steps. In the first step, an action label $\tilde{c}$ is predicted by NN-based action classification on all training data $\mathbf{Y}_{train}$. Subsequently, an actor label $\tilde{s}$ is predicted by NN based actor classification on the action-specific subset of the training data, i.e. $\mathbf{Y}^{\tilde{c}}_{train}$.

We also examined another possible implementation of multi-label NN by considering the number of classes equal to the number of contents times the number of styles. However, the results were similar. The role of the number of training samples per action-actor pair in the recognition task is explained in Section 6.5.3. All our experimental results are based on *leave-one-run-out* cross validation. In each iteration of this scheme, samples from all but one run are selected for training and the rest are used for testing. For each experiment, we report the average accuracy over all runs.

6.5.1 IXMAS Dataset

IXMAS is a popular multiview action recognition dataset. It consists of videos of 11 actions performed 3 times by 10 different actors, i.e. 330 video sequences. These videos are acquired by using 5 cameras. The actors were free to choose their location and orientation for each run. Weinland et al. [2006] generated 3D motion history volumes from those videos and used a Fourier transform of cylindrical coordinates to get locations, scale, and rotation invariant features (Fig. 6.4). PCA was then applied to reduce feature space dimensionality. We used the same motion history volume features with the dimensionality equal to 329.

Figure 6.5 shows the confusion matrices for the individual factors by NN classification. As expected, a high average accuracy for action classification (88.79%) was observed as compared to actor classification (58.48%). Notice that actions such as *sitdown*, *getup*, *turnaround*, and *walk* that involve full body movements were more distinguishable than the other actions, such as *checkwatch* and *scratchhead*.

For action-actor classification, we achieved around **31.21%** and **58.67%** accuracies for symmetric bilinear modeling and nearest neighbor approaches, respectively. On average, symmetric bilinear models took 60 iterations to con-

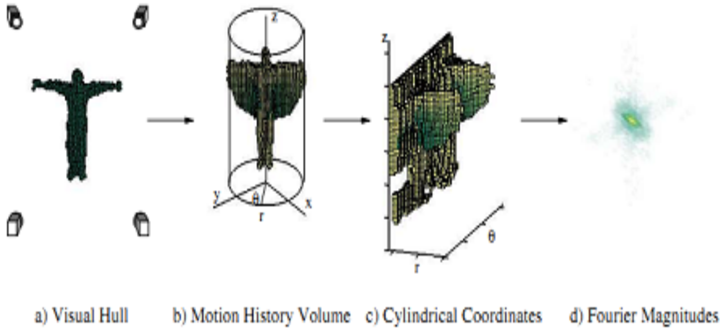
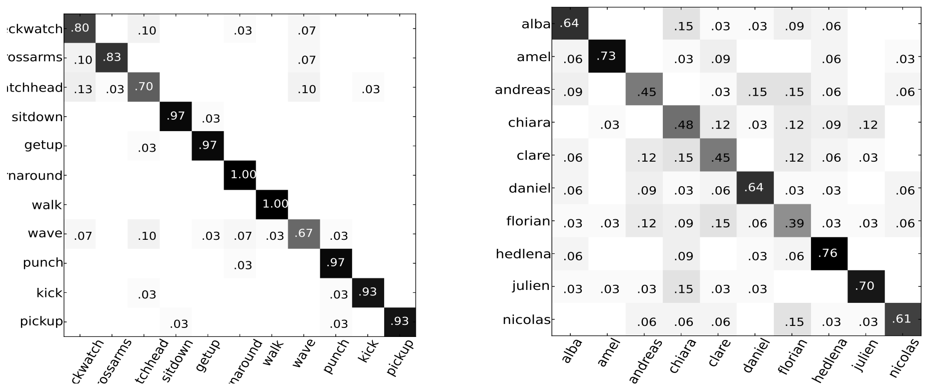


Figure 6.4: Feature extraction using motion history volumes [Weinland et al., 2006].



	ckwatch	rossarms	tchhead	sitdown	getup	naround	walk	wave	punch	kick	pickup
ckwatch	.80		.10			.03		.07			
rossarms	.10	.83						.07			
itchhead	.13	.03	.70					.10		.03	
sitdown				.97	.03						
getup			.03		.97						
naround						1.00					
walk							1.00				
wave	.07		.10		.03	.07	.03	.67	.03		
punch						.03			.97		
kick			.03						.03	.93	
pickup				.03					.03		.93

	alba	amel	andreas	chiara	clare	daniel	florian	hedlena	julien	nicolas
alba	.64			.15	.03	.03	.09	.06		
amel	.06	.73		.03	.09			.06		.03
andreas	.09		.45		.03	.15	.15	.06		.06
chiara		.03		.48	.12	.03	.12	.09	.12	
clare	.06		.12	.15	.45		.12	.06	.03	
daniel	.06		.09	.03	.06	.64	.03	.03		.06
florian	.03	.03	.12	.09	.15	.06	.39	.03	.03	.06
hedlena	.06			.09		.03	.06	.76		
julien	.03	.03	.03	.15	.03	.03			.70	
nicolas			.06	.06	.06		.15	.03	.03	.61

Figure 6.5: IXMAS: confusion matrices for (a) action recognition (b) actor recognition

verge to optimal style and content parameters. Our proposed approach yields a significantly higher accuracy of **88.79%**. Figure 6.6(a–c) plots accuracies for identification of each (action, actor) pair for the three methods. Figure 6.6(d) compares the accuracies of the three approaches *across* different actions. The columns represent style or person-specific uniqueness for different actions with standard error. It shows that our approach consistently outperforms other methods. Notice, in particular, the poor performance of nearest neighbor and symmetric bilinear approaches to style recognition for actions that involve full body movements, e.g. *sitdown*, *getup*, *walk* and *turnaround*. In contrast, the highest style separability is obtained for all actions by our approach, which indicates its robustness towards the activities involving articulations at different

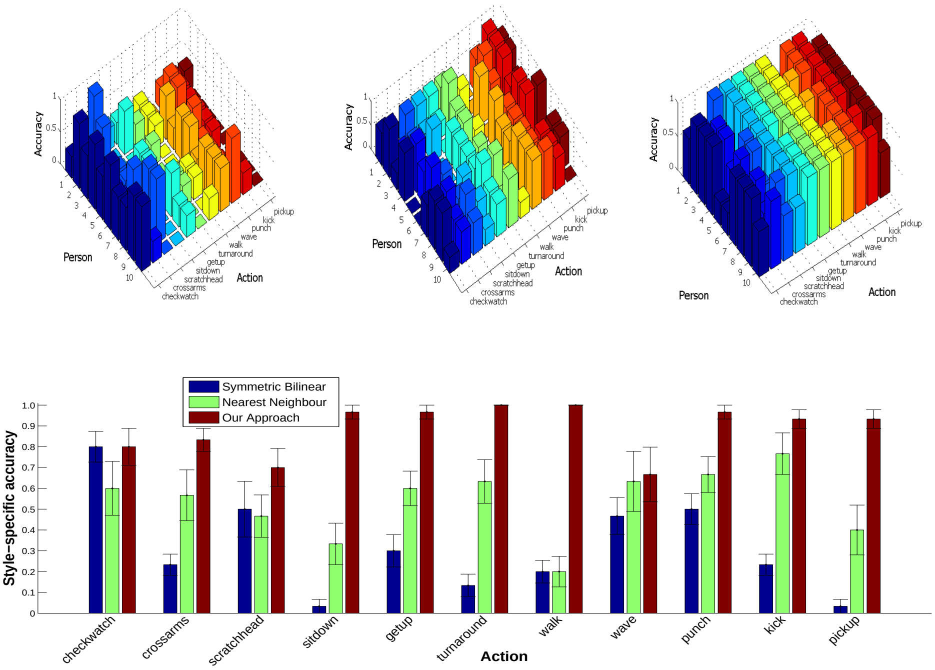


Figure 6.6: Results on IXMAS: (a–c) accuracy for actor-action recognition by symmetric bilinear model, NN, and our approach respectively (d) Comparison of style recognition per action

scales.

6.5.2 Berkeley-MHAD Dataset

Berkeley-MHAD [Ofli et al., 2013] is a multimodal dataset consisting of sequences of 11 actions performed 5 times by 12 different subjects for a total of 660 action sequences. These activities are captured in different modalities including mocap, audio, body acceleration, color, and depth data. In our experiments, we used two modalities: skeleton information from motion capture system and depth information from Kinect sensors. The bag-of-features extraction using temporal segmentation is the same as discussed in the previous chapter (Section 5.5.3). Next, we discuss the results for our action-actor recognition problem.

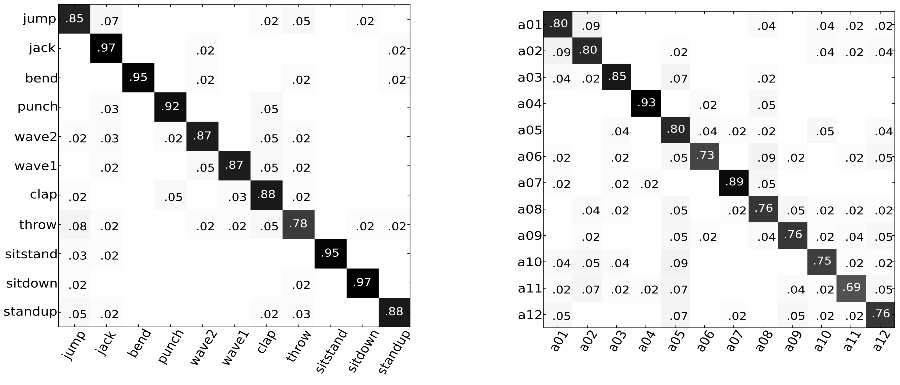


Figure 6.7: MHAD-Mocap: confusion matrices for (a) action recognition (b) actor recognition

Mocap data

Figure 6.7 shows the confusion matrices for the single-factor action and actor recognition by nearest neighbor approach. A high action classification rate (89.85%) was observed compared to actor classification (79.39%). Interestingly, it turns out that the identification of the style (independent of automatic action recognition) in skeletal motion data is convincingly plausible.

Experimental results of multi-factor classification show **33.94%**, **68.79%** and **87.83%** accuracies for symmetric bilinear modeling, nearest neighbor and our approach respectively. Figure 6.8(a–c) plots accuracies for identification of each (action, actor) pair for the three methods. Figure 6.8(d) compares the accuracies of the three approaches *across* different actions. Clearly, our approach achieves high recognition accuracy for all actions except the action *throw*. This is mainly due to the miss-classification of *throw* in the first stage (See Fig. 6.7 (a)).

Kinect depth data

Figure 6.9 shows the resulting confusion matrices for the single-factor action and actor recognition using NN. For the two cases, we achieved 77.73% and 57.73% accuracy, respectively. While some actions such as *jack*, *bend*, *wave*

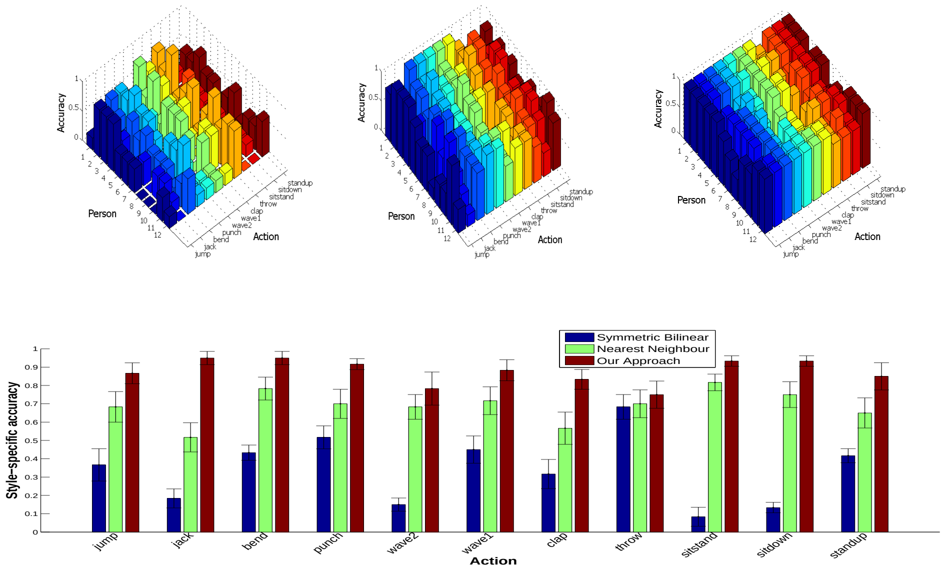


Figure 6.8: Results on MHAD-Mocap: (a–c) accuracy for actor-action recognition by symmetric bilinear model, NN, and our approach respectively (d) Comparison of style recognition per action

and *sitstand* show high recognition rate, the others, such as *throw* and *jump* appear to be less discernible. Similarly, actors such as *a09* and *a01* were more consistent with respect to their style as compared to the other actors.

Figure 6.10 provides a detailed comparison for action-actor classification and shows that our approach (**73.03%**) significantly outperforms symmetric bilinear modeling (**30%)** and nearest neighbor classification (**47.73%**). By comparing Fig. 6.8 with Fig. 6.10, one can notice that (i) for most of the actions (except *jump*, *clap* and *throw*), our approach is capable of almost correctly identifying human actions and actors from both the mocap and the depth sequences and (ii) although higher recognition rates were achieved on mocap data, the results on depth data are still significant. We expect that even higher recognition accuracy can be achieved on depth imagery using information from other DLMC channels (See [Ofli et al., 2013]).

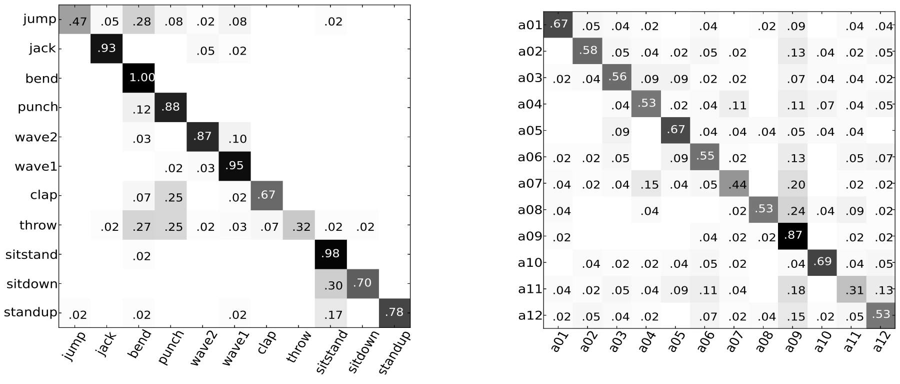


Figure 6.9: MHAD-Depth: confusion matrices for (a) action recognition (b) actor recognition

6.5.3 The Role of the Number of Training Examples per Action-actor Pair

Apparently, there are no theoretical and technical constraints on our approach to use multiple instances per action-actor pair. However, we empirically evaluated the extent to which the number of training instances influences the performance of the multi-factor classification. To this end, we considered MHAD datasets, which contain 5 instances for every combination of action and actor (due to 5 repetitions of each action by each person). Again, we adopted the *leave-one-run-out* scheme such that, in each iteration, we first selected all samples of a run as query instances, i.e. one test instance per action-actor pair. From the remaining data, we (randomly) formed 4 training subsets $S_i, 1 \leq i \leq 4$, such that each S_i consisted of exactly i unique samples per action-actor pair. For each i, the average accuracy was determined across all the runs.

Figure 6.11 plots average accuracies of single-label action or actor recognition while Fig. 6.12 plots multifactor action-actor recognition for different values of i. The results show that (a) providing more data for training improves the classification performance as, for instance, it allows a robust estimation of the underlying style-specific basis in our hierarchical approach, (b) however, the rate of this gain decreases in our experimental data, e.g. increasing the number of training instances per action-actor pair from 2 to 3 or 4 from the

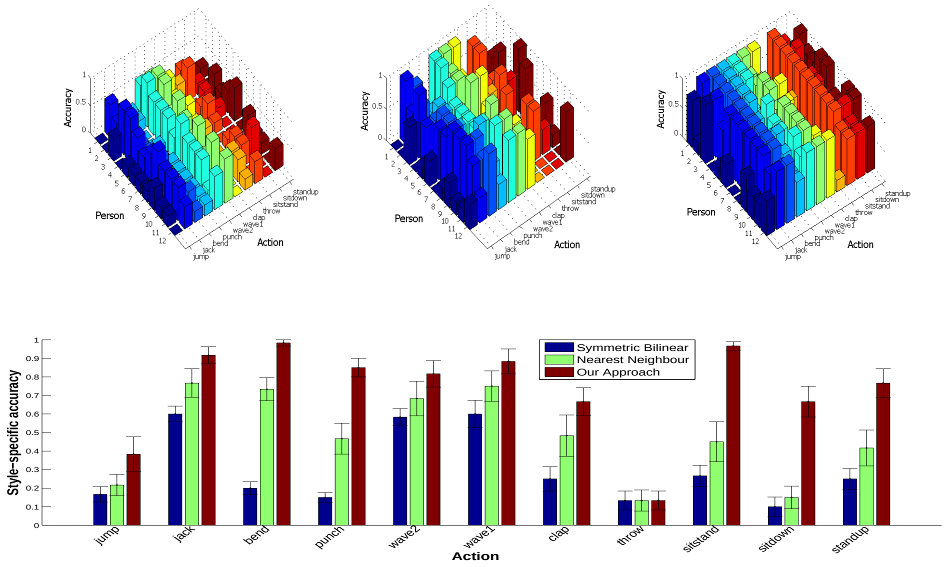


Figure 6.10: Results on MHAD-Depth: (a–c) accuracy for actor-action recognition by symmetric bilinear model, NN, and our approach respectively (d) Comparison of style recognition per action

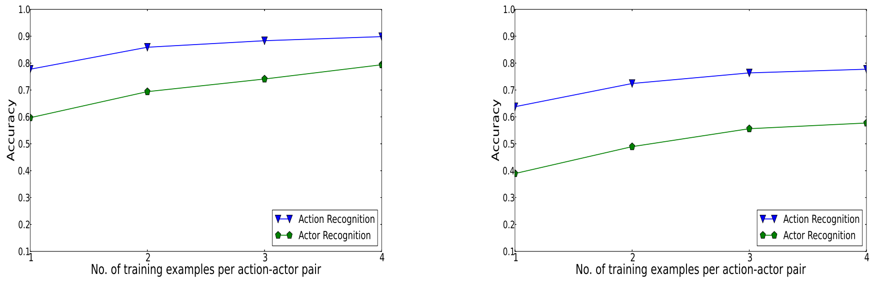


Figure 6.11: Effect of number of training instances per action-actor pair on individual recognition of actions or actors on (a) MHAD-Mocap (b) MHAD-Depth datasets

MHAD-Mocap data did not yield a significant gain, (c) the recognition rate of our hierarchical action-actor classification apporach is proportional to the rate of action recognition in the first step and with more samples per action-actor pair, our approach can achieve an accuracy which is near to that of regular action recognition, and (d) the proposed hierarchical approach outperforms other methods regardless of the amount of available data (Fig. 6.12).

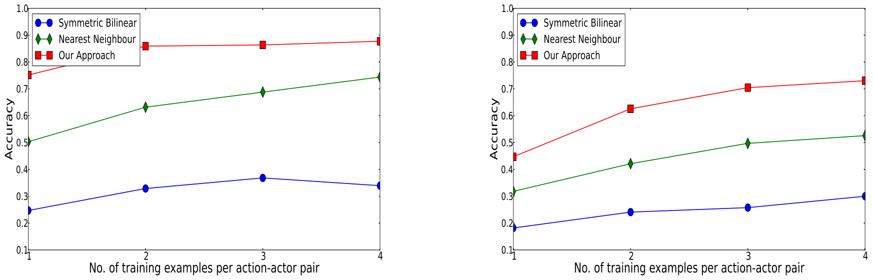


Figure 6.12: Effect of number of training instances per action-actor pair on multi-factor recognition on (a) MHAD-Mocap and (b) MHAD-Depth datasets

Task	**Approach**	**IXMAS**	**Berkeley-MHAD**	
		3D-Volumes	Mocap	Depth
action recognition	Nearest Neighbor	88.79%	89.85%	77.73%
actor recognition	Nearest Neighbor	58.48%	79.39%	57.73%
	Symmetric Bilinear	31.21%	33.94%	30.0%
action-actor recognition	Nearest Neighbor	58.67%	68.79%	47.73%
	Our Approach	88.79%	87.73%	73.03%

Table 6.1: Accuracies of different approaches

6.6 Conclusion and Future Directions

This chapter extends the applicability of human action recognition in 3D sequences. It discusses a novel but important problem of recognition of actions and the people performing those actions in 3D videos. We presented a hierarchical approach based on conventional action recognition and asymmetric bilinear modeling of styles. We were able to achieve high accuracy on different benchmark action datasets by using features that are based on motion cues. In particular, we demonstrated the capability of our approach to identify actions and people in sequences captured in the form of motion capture dynamics, 3D motion history volumes constructed by using a multi-camera system or depth images obtained from a single Kinect camera. To the best of our knowledge, this is the first such investigation in the area of human action recognition. In constrained scenarios, such as surveillance or Kinect-depth imagery, our approach is directly applicable. In the context of biomechanics, this chapter establishes that person identification is possible by recognizing execution styles of a number of human actions. In particular, we showed that person identification through motion dynamics is not limited to walking and running. We also showed that ou hierarchical approach outperforms symmetric bilinear modeling and nearest neighbor when the underlying factors result from the same generative process (motion cues in our case). This promises the applicability of our approach towards multifactor classification in different fields such as speech and facial emotions recognition.

Although the proposed system is evaluated on the readily available benchmark multi-actor multi-action datasets that have a maximum of 11 actions and 12 individuals, its performance to the larger datasets will depend at first hand on the discriminativeness of the actions. The experimental results show that our approach to action-actor classification can achieve an accuracy which is near to that of regular action recognition (Table 6.1). Seemingly, there are no theoretical and technical bounds on the applicability of asymmetric bilinear modeling to large scale recognition. However, it will be interesting to see how it scales to even larger and more versatile set of activities such as those containing unconstrained videos. A promising future direction can be extending our hierarchical framework to incorporate multiple factors including action, view, actor, scenario, camera motion, and visibility. This work may require databases larger than the existing realistic datasets such as HMDB [Kuehne et al., 2011] and UCF50 [Reddy and Shah, 2013] since they do not contain samples for many possible combination of the factors.

Summary

Although significant research efforts are now devoted to action recognition in emerging 3D environments (e..g. [Wang et al., 2012, Ofli et al., 2013]), the rich pose and motion information invites researchers to *scale up* the applications and the problem itself. Motivated by this objective, we treated the novel issue of recognizing human actions and the underlying execution styles (actors) in videos using motion dynamics only. While there exist profound studies in psychophysics [Cutting and Kozlowski, 1977, Thoroughhman and Shadmehr, 1999] which suggest that people tend to perform different actions in their own style, this chapter is among the first vision-based attempts that consider recognizing activities and performing styles for various human actions. We presented a hierarchical approach that is based on conventional action recognition and asymmetrical bilinear decomposition. In particular, we applied bilinear factorization on the tensorial representation of the action videos to characterize styles of performing different actions. Given a query sequence, we first apply a classical action classification to predict underlying action of the query video and then use this prediction to generate a style-specific basis for the query video using an asymmetric bilinear model. Finally, we compare this basis with the style-specific basis learned from training data in order to identify the most likely style. Through extensive experimentation on multiple depth and skeletal datasets, we showed that our hierarchical approach significantly improves results compared to naive nearest neighbor classification and symmetric bilinear modeling. A major contribution of this work is suggesting horizontal expansion of the problem of activity recognition. Seemingly, there are no theoretical and technical bounds on the applicability of asymmetric bilinear modeling to large-scale recognition. However, it will be interesting to see how it scales to even larger and more versatile set of activities such as those containing unconstrained videos. A salient direction of future research is extending the hierarchical bilinear framework to incorporate multiple factors including action, view, actor, scenario, visibility, and camera motion. This may eventually allow to apply such models to large-scale realistic action datasets.

Chapter 7

Learning Spatial Interest Regions from Videos for Action Recognition in Still Images

In Chapters 3 and 4, we observed how non-temporal approaches can be applied to efficient recognition of human activities in videos. Especially in Chapter 3, we presented a simple yet powerful approach to learn weights for the key poses that can optimally discriminate different action sequences. Complementing that work, this chapter presents a novel approach to action recognition in still images that exploits motion cues to determine salient image regions for discriminating different actions. This is of particular interest given that most approaches to human action recognition in still images are based on computing local descriptors in the vicinity of spatial key points. The key points either result from running a key point detector or from dense random sampling of pixel coordinates. Furthermore, they are not *a-priori* related to human activities and thus might not be very informative with regard to action recognition. Other approaches involve manual efforts and construct saliency maps using human visual attention or by making a set of discriminative postures called *poselets*.

We investigate the possibility and applicability of automatically identifying action-specific key points or regions of interest in still images based on information extracted from video data. This chapter presents our novel method for extracting spatial interest regions where we apply non-negative matrix factorization to optical flow fields extracted from videos. The resulting basis flows imply image regions that are specific to certain actions and therefore allow for an informed sampling of key points for feature extraction. We thus present a generative model for action recognition in still images that allows for characterizing joint distributions of regions of interest, local image features (visual words), and human actions. Experimental results shows that (a) our approach is able to extract interest regions that are greatly associated with those body parts most relevant for different actions and (b) our generative model achieves high recognition accuracy in action classification.

7.1 Introduction

Recognizing human actions in still images is a challenging task due to a number of factors such as lack of any motion cue or 3D shape information, partial occlusion, influence of texture, and noise. The problem has received considerable attention throughout the last decade, and efforts are still in progress. Corresponding research is motivated by promising applications in areas such as automatic indexing of large image repositories, automatic scene description, context-dependent object recognition, and pose estimation [Sun and Savarese, 2011, Johnson and Everingham, 2011, Weinland et al., 2011]. Recent approaches to action recognition (in still images) can be broadly divided into two main classes: (a) pose-based and (b) bag-of-features (BoF) approaches. Following the idea of *poselets* [Bourdev and Malik, 2009] – a notion of distributed part-based templates – pose-based approaches have recently been met with rekindled interest [Yang et al., 2010, Maji et al., 2011, Yao et al., 2011b]. However, the construction of poselets still requires a cumbersome procedure of manual annotation which impedes their use on large training sets. BoF approaches based on local descriptors are known for their state-of-the-art performance in object recognition and therefore have been adapted to action recognition [Deltaire et al., 2010]. In these approaches, the local image descriptors are typically computed in the vicinity of key points that result from low-level signal analysis or from dense or random sampling. Consequently, those key points are uninformative and independent of the activity depicted in an image.

Significant research efforts (e.g. [Sharma et al., 2012, Bilen et al., 2013, Sharma et al., 2013]) are now being made towards enhancing action classification by determining salient or most relevant key points/patches in images. These approaches build on the observation that most people can infer human activities in still images just by looking at the posture or configuration of particular

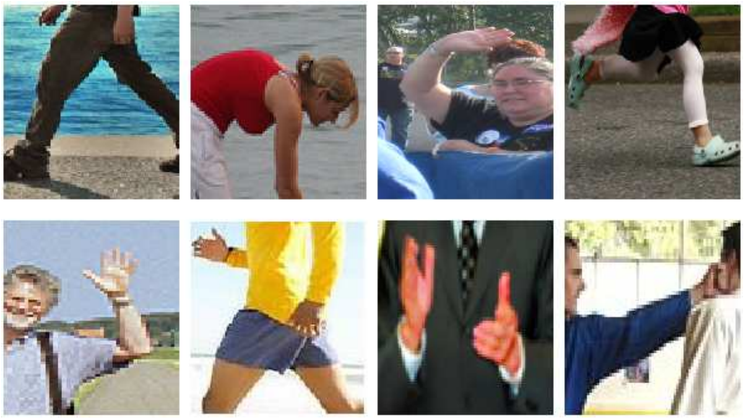


Figure 7.1: Examples of image patches in which human activities can be recognized even though the full body is not visible.

body parts. For instance, consider the images shown in Fig. 7.1 which one can interpret even without having a full view of the human body. This raises the question of whether it is possible to automatically learn or identify action-specific, informative, regions of interest in still images without having to rely on exhaustive mining of low-level image descriptors or labor-intensive annotations.

This chapter presents our attempt to answer this question and gives details of an efficient yet effective approach towards automatic learning of action specific regions of interest in still images. Based on the observation that activities are temporal phenomena (as they are characterized by articulation and movement of different body parts), we make use of information that is available from video analysis. Figure 7.2 presents a diagram of the components of the proposed approach towards determining action-specific regions of interest and subsequent image classification.

Given a set of training videos, each showing a single human performing some action, we compute optical flow fields and determine the magnitudes of the flow vectors in each frame of a video. We represent the set of all frames of flow magnitudes as a matrix and apply non-negative matrix factorization (NMF) to obtain basis flows. These basis flows are indicative of the position and configuration of different limbs or body parts whose motion characterizes certain activities. Viewed as images, the basis flows exhibit action-specific regions of interest and therefore allow for an *informed sampling* of interest points or regions for subsequent feature extraction. For action classification in still images, we formulated a generative probabilistic model that characterizes joint distributions of interest regions, local image descriptors (visual words) and human actions.

The major contributions of this chapter are the following: (i) we present a novel approach for determining discriminative spatial regions for action recognition in still images using simple videos; (ii) we apply NMF to determine action-specific regions of interest from motion flows; (iii) we incorporate action saliency maps based on videos and local spatial features of action images in a Bayesian framework for human action classification.

In Section 7.2, we review related work on human action recognition in still images. Section 7.3 describes our method of learning action-specific interest regions from videos. In Section 7.4, we present a generative model for action classification. In section 7.5, we evaluate both components of our approach. In particular, Section 7.5.1 evaluates the usefulness of regions of interest contained in basis flows for different actions by comparing correspondences between regions of interest that were automatically learned from videos and manually annotated locations of human body parts that are available from an independent set of still images. Section 7.5.2 shows that, even in the absence of any annotation of joints or body parts, our generative model achieves high accuracy for action classification in still images. Finally, Section 7.6 summarizes our work and results.

7.2 Related Work

Human action recognition in still images has been a topic of great interest to vision researchers. A number of approaches have been proposed in the last decade. Here, we restrict our discussion to the two arguably most popular approaches in the recent literature. In addition, we briefly review related matrix factorization methods. Similar to the case of videos, the idea of bags of visual words (BoWs[9]) is popular also in human action recognition in still images for its known simplicity, robustness, and good performance in content-based multimedia classification. Corresponding research treats an image as

[9] Throughout this chapter we will use the terms bag-of-features(BoF) and bag-of-words(BoW) interchangeably.

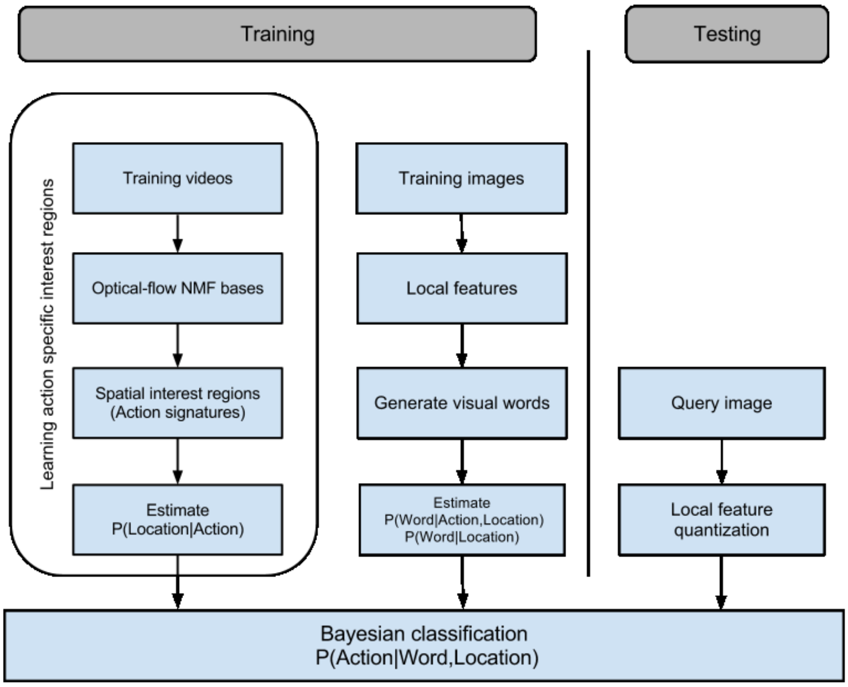

Training
Testing
Learning action specific interest regions
Training videos
Optical-flow NMF bases
Spatial interest regions (Action signatures)
Estimate P(Location|Action)
Training images
Local features
Generate visual words
Estimate P(Word|Action,Location) P(Word|Location)
Query image
Local feature quantization
Bayesian classification P(Action|Word,Location)

Figure 7.2: General diagram of our approach. In the training phase, we learn the actions' priors P(Location | Action) from training videos, and code book priors P(Word | Action,Location) from training images in order to perform human action recognition for new test queries in a fully Baysian setup.

a collection of independent visual descriptors computed at certain key point locations. Determining those key points is crucial within the BoW framework since it preselects image patches for subsequent classification. Naturally, one would like to focus only on those patches that are most discriminative.

BoW approaches such as [Matikainen et al., 2009, Laptev et al., 2008] based on key points detection [Harris and Stephens, 1988, Schmid et al., 2000, Lowe, 2004, Laptev, 2005, Jhuang et al., 2007, Willems et al., 2008, Bay et al., 2008], though generally discriminative, do not regard task specific objectives in key point localization. Rather, key points are determined from low-level properties of the image or video signal. Moreover, corresponding approaches typically assume key points to be independent and therefore fail to explain characteristic spatial and temporal layouts. Kovashka and Grauman [2010] addressed this limitation and proposed a mid-level representation that encodes spatial and temporal relationships among key points. The authors of [Gilbert et al., 2011, Liu et al., 2012] employed data mining to build high-level compound features from noisy and over-complete sets of low-level spatio-temporal features. In [Song et al., 2003], a triangular lattice of grouped point features was used to encode spatial layouts. Research presented in [Coates and Ng, 2011, Malinowski and Fritz, 2013, Sharma et al., 2012] explicated the importance of weighting local features while pooling in a way that regards the classification task in hand. Still, these approaches also center around low-level signal properties which do not necessarily provide an accurate account of the characteristics of an activity.

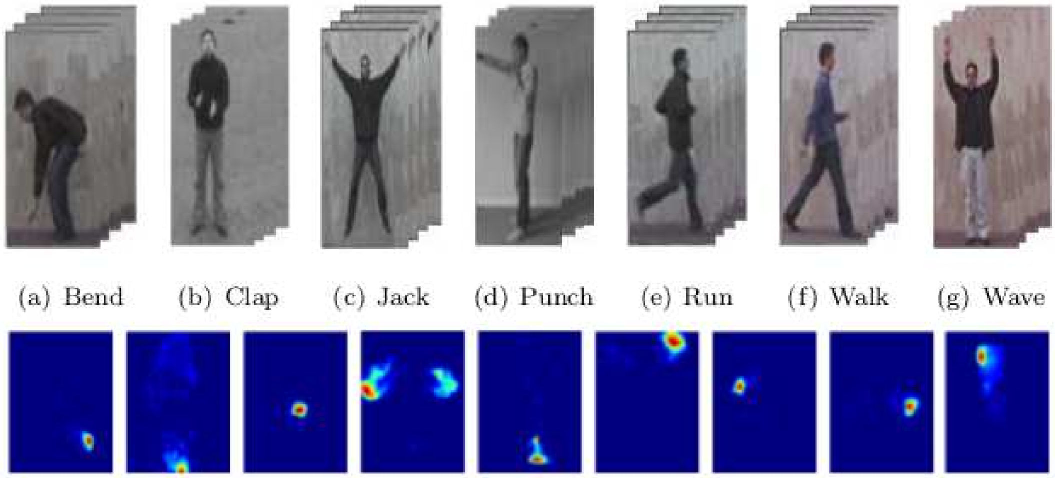
Some recent approaches proposed human-based fixation for sampling key points [Vig et al., 2012, Mathe and Sminchisescu, 2012, Itti and Koch, 2000]. Mathe and Sminchisescu [2012] proposed a saliency map learned from eye movements. Vig et al. [2012] presented a saliency-based descriptor for action recognition. These approaches show that using saliency maps learned from human fixation locations enhances the performance in comparison to other sampling techniques, while using an order of magnitude fewer feature descriptors. As opposed to these methods, our approach automatically learns the saliency maps from training videos (without human intervention) by analyzing their motion fields using NMF. Some object discovery approaches exploit temporal information for automatic detection of salient objects [Herbst et al., 2011, Garca et al., 2013]. For example, Herbst et al. [2011] consider a sequence over time and detect changes in two 3D maps for subtracting background (motion) and locating objects in a scene. Garca et al. [2013] also observe a scene over time, estimate so called proto-objects, and refine them to build object models. However, the extent to use motion information in order to build saliency map for human action recognition in still images is not yet explored.

Sampling techniques such as random sampling have also shown state-of-the-art action recognition performance. Nowak et al. [2006] empirically showed that random sampling provides equal or better activity classifiers than several sophisticated multi-scale interest point detectors; yet their work also illustrates that the most important aspect of sampling is the number of sample points extracted. Wang et al. [2009] states that dense sampling outperform all point detectors in realistic scenarios. However, recent work in [Gall et al., 2011] demonstrated that state-of-the-art action classification can also be obtained from only a few randomly sampled key points. It therefore appears to be an open issue whether to use dense or random sampling. It is, however, obvious that the success of dense sampling is bought at the expense of memory- and runtime-efficiency, whereas random sampling methods do not provide statistical guarantees as to their adequacy for the task at hand. Therefore, methods which mark a middle ground– namely informed sampling– seem to merit closer investigation.

Part-based approaches, too, are popular in research on human action recognition and were indeed shown to successfully cope with the PASCAL visual object recognition challenge[10]. Felzenszwalb et al. [2010] described a deformable model for human detection which was used to achieve state-of-the-art performance in action recognition on benchmark datasets [Deltaire et al., 2010]. Bourdev and Malik [2009] introduced exemplar-based pose representation, named *poselets*, for human detection. The term *poselet* denotes a set of patches with similar pose configurations. Maji et al. [2011] utilized poselets to identify human poses as well as actions in still images. Sun and Savarese [2011] proposed an articulated part-based model for human pose estimation and detection which adapts a hierarchical (coarse-to-fine, poselet-like) representation. Yang et al. [2010] exploited poselets as a coarse representation of a human pose and treated them as latent variables for action recognition. Despite their recent success, it is still questionable if these methods can make use of the favorable statistics of present-day large-scale datasets because the construction of suitable poselets requires extensive human intervention and manual labeling in the training phase.

Non-negative matrix factorization (NMF) is an unsupervised matrix factorization approach, which is often used to learn parts of objects [Lee and Seung, 1999]. Recent applications of NMF to human action recognition have shown that the extracted spatial parts entail semantic correspondence to human body parts. For example, Thurau and Hlavac [2008] employed NMF to learn a set of pose and background primitives for action recognition. In [Agarwal and Triggs, 2006] also, the human upper body pose was estimated through NMF. Eweiwi et al. [2013] presented a supervised approach to multiview human action recognition based on discriminative joint NMF. In these cases, NMF is employed to determine the part-based representation in image feature space. The work presented in this chapter, however, applies NMF to motion sequences in order

[10]http://pascallin.ecs.soton.ac.uk/challenges/VOC/



(a) Bend (b) Clap (c) Jack (d) Punch (e) Run (f) Walk (g) Wave

(h) Examples of basis vectors obtained from NMF

Figure 7.3: (a–g) Examples of training videos from the Weizmann and KTH datasets; (h) examples of basis flows obtained from applying NMF on optical flow fields.

to determine salient activity regions in person bounding boxes.

An empirical evaluation of pose- and appearance-based features is given in [Yao et al., 2011a]. The authors concluded that even for rather coarse pose representations, pose-based features either match or outperform appearance-based features. However, they acknowledge that appearance-based features still represent an ideal resort for cases of considerable visual occlusion. Accordingly, it appears worthwhile to study methods that allow for integrating both approaches into a single framework. Next, we discuss how we indeed exploit pose articulation from videos for the informed sampling of key points for appearance-based action recognition in images.

7.3 Learning Action-specific Interest Regions from Videos

Our approach identifies discriminative regions in the image plane and subsequently learns the relative importance of these regions for different actions. In order to identify salient spatial locations, we apply NMF to optical flow fields obtained from videos. Furthermore, we make use of NMF mixture coefficients in order to derive a generative probabilistic model that features joint distributions of local features, regions of interest, and human actions.

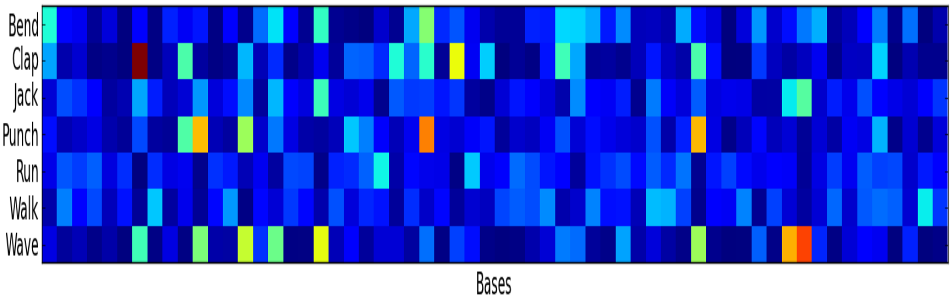


Figure 7.4: Relative importance of bases with respect to different actions as characterized by $P(\mathbf{w}_k|a_i)$. Note that action flows can be approximated by a small number of basis vectors.

7.3.1 Learning Basis Flows using NMF

For a given set of training videos of different actions, we determine optical flow magnitudes at each pixel within a bounding box of constant size surrounding a person visible in the video. Each frame can thus be transformed into a d dimensional non-negative vector $\mathbf{u}$. Let n_i represent the number of frames for an action $a_i \in \mathcal{A} = \{a_1, a_2, ..., a_r\}$ and let $N = \sum_{i=1}^{r} n_i$. We build a data matrix $\mathbf{U}$ of dimension $d \times N$ containing the flow magnitude vectors of all frames. The NMF of $\mathbf{U}$ yields K basis vectors, or *basis flows*, such that $\mathbf{U} \approx \mathbf{WH}$, where the columns of $\mathbf{W}_{d\times K}$ are non-negative basis elements and the columns of $\mathbf{H}_{K\times N}$ encode non-negative mixture coefficients.

In order to determine the factors $\mathbf{W}$ and $\mathbf{H}$, we apply the gradient descent algorithm according to [Lee and Seung, 1999]. This method is known to yield sparse basis elements, for it converges to vectors that lie in the facets of the simplicial cone spanned by the data (see the discussions in [Donoho and Stodden, 2004, Klingenberg et al., 2008, Thurau et al., 2011]). Accordingly, we can expect the resulting basis flows to be sparse in the sense that most entries of a basis element $\mathbf{w}_k$ will be (close to) zero and only a few entries will have noticeable values. Figure 7.3 (h) shows that this is indeed the case. It depicts pictorial representations of exemplary basis vectors $\mathbf{w}_k$ resulting from our NMF step. Note that, for each basis element, only a few pixels are larger than zero; in each case, these pixels apparently form distinct, more or less compact patches in the image plane.

7.3.2 Learning the Action-specific Importance of Basis Flows

Different actions are characterized by the articulation and movements of different body parts. The NMF basis vectors determined through factorization of frame-wise optical flow magnitudes appear to indicate image regions of importance for different actions. Here, we aim to learn the relative importance of different basis elements with respect to different actions. To this end, we consider the matrix **H** because its entries encode linear mixing coefficients required to reconstruct the vectors in **U** from the basis flows in **W**. Consequently, the columns of **H** represent the relevant importance of a basis for a given frame. Their (L_1) normalization to stochastic vectors allows us to estimate a joint probability distribution of actions and bases. The conditional probability of basis $\mathbf{w}_k$ given an action a_i is determined as:

$$P(\mathbf{w}_k|a_i) = \frac{\sum\limits_{f \in a_i} h_{kf}}{\sum\limits_{j=1}^{K} \sum\limits_{f \in a_i} h_{jf}} \tag{7.1}$$

Note in Fig. 7.4 that the resulting probability distribution, i.e. the weights of the basis elements w.r.t. different actions, again is sparse. Therefore, the distribution in Eq. (7.1) immediately allows us to determine how characteristic a certain basis flow is for an action. As an example, Fig. 7.5 shows the three highest ranking basis elements for a few exemplary actions from KTH and Weizmann datasets.

7.3.3 Action Signatures and Salient Regions

The probability distribution $P(\mathbf{w}_k|a_i)$ in Eq. (7.1) also allows us to consider *action signatures*, which we define to be the conditional expectations

$$\mathbf{s}_i = \sum_{k=1}^{K} P(\mathbf{w}_k|a_i)\mathbf{w}_k. \tag{7.2}$$

Computing and plotting action signatures s_i for different actions a_i, we find that characteristically different regions in the image plane are intensified for different actions. Figure 7.6 shows examples of action signatures which we obtained from basis flows extracted from the Weizmann and KTH datasets.

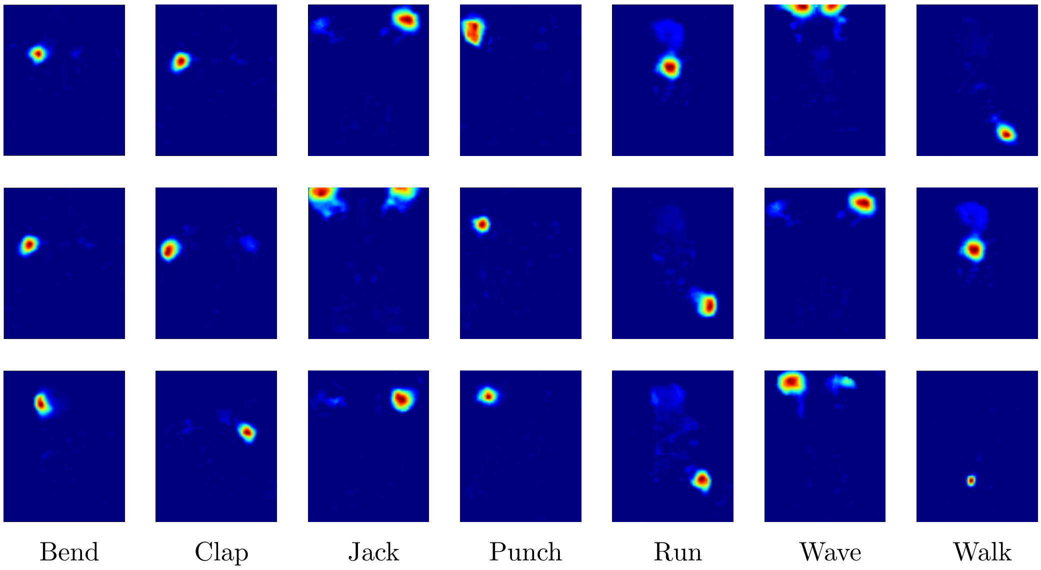


Figure 7.5: Top 3 bases for selected actions based on $P(\mathbf{w}_k|a_i)$. Note that these bases are sparse and shared among all actions with different mixing coefficients based on their contribution to their corresponding actions

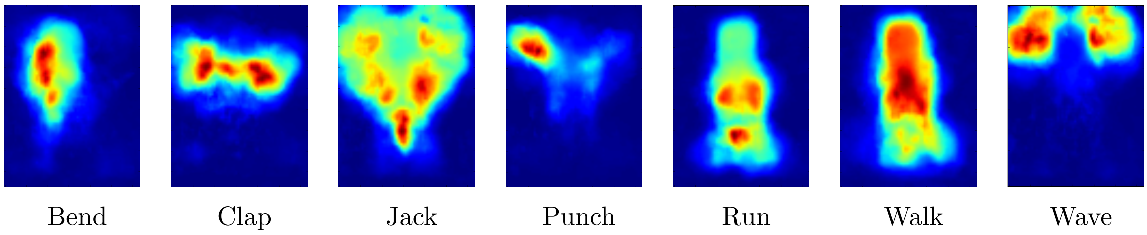


Figure 7.6: Examples of action signatures resulting from equation (7.2).

Apparently, these action signatures may serve two purposes. Firstly, they provide us with a prior distribution for the sampling of interest points from still images showing people in order to compute action-specific local features for activity classification. Secondly, action signatures may be used as templates or filter masks for pose-based activity recognition. Regarding the former, each action signature s_i, i.e. a $d-$dimensional vector in the image space, can be used to derive an action-specific spatial saliency for an image region l_k, namely

$$P(l_k|a_i) = \sum_{j \in l_k} s_i. \tag{7.3}$$

7.4 Action Classification in Still Images using Spatial Interest Regions

In this section, we describe a Bayesian framework for action classification that combines the Bag-of-Words approach used for still images with action signatures learned from videos. For a given set of training images $\mathbf{F} = \{(\mathbf{f}_i, y_i), i = 1, 2, ...M\}$ where $y_i \in \mathcal{A}$, each image is first divided into a set L of cells (or locations) and a local histogram of oriented gradient are extracted for each of the locations. A vocabulary of visual words $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m\}$ is then learned using k-means clustering based on the L_2 norm. Thus, each training image is represented as a vector of $|L|$ visual words.

Our classification approach considers the likelihood of an action given spatial locations (with their relative importance) and visual words for those locations. This likelihood $P(a_i|\mathbf{v}_j, l_k)$ is estimated as

$$\begin{aligned} P(a_i|\mathbf{v}_j, l_k) &= \frac{P(\mathbf{v}_j|a_i, l_k)P(a_i|l_k)}{P(\mathbf{v}_j|l_k)} && (7.4) \\ &= \frac{P(\mathbf{v}_j|a_i, l_k)P(l_k|a_i)P(a_i)}{P(\mathbf{v}_j|l_k)P(l_k)} && (7.5) \\ &= \alpha\frac{P(\mathbf{v}_j|a_i, l_k)P(l_k|a_i)}{P(\mathbf{v}_j|l_k)} && (7.6) \end{aligned}$$

where α is the normalization factor and priors $P(a_i)$ and $P(l_k)$ are assumed to be uniformly distributed. The three conditional probabilities on the right hand side of Eq. (7.6) are estimated from the training data ($\mathbf{U}$ and $\mathbf{F}$).

$P(l_k|a_i)$ represents the action-specific importance of each region and is derived from action signatures that were learned from action videos (see Eq. 7.3). The term $P(\mathbf{v}_j|l_k)$ denotes the likelihood of visual word v_j given location l_k and is determined by

$$P(\mathbf{v}_j|l_k) = \frac{\sum\limits_{\mathbf{f}\in\mathbf{F}} \chi(\mathbf{v}_j, l_k)}{|\mathbf{F}|} \qquad (7.7)$$

where χ is an indicator function that has value 1 only if v_j is assigned to l_k.

While the measure $P(\mathbf{v}_j|l_k)$ indicates the overall probability of the occurrence of a visual word at a specific location, the action-specific likelihood $P(\mathbf{v}_j|a_i, l_k)$ further specifies the relevant importance of visual words at different locations for different actions. This is achieved by computing

$$P(\mathbf{v}_j|a_i, l_k) = \frac{\sum\limits_{\mathbf{f}\in\mathbf{F}^{(a_j)}} \chi(\mathbf{v}_j, l_k)}{\left|\mathbf{F}^{(a_j)}\right| |L|} \tag{7.8}$$

where $\mathbf{F}^{(a_j)} \subset \mathbf{F}$ are those training images that contain examples of action a_i.

In summary, our generative model is composed of three components: $P(Word|Location)$, $P(Location|Action)$ and $P(Word|Action, Location)$ (See Eq. 7.6). Among these, the factor $P(Location|Action)$ is learned by NMF of action videos while $P(Word|Location)$ and $P(Word|Action, Location)$ are learned by image features. The generative nature of our framework makes it flexible enough to adapt to different kinds of variations and constraints. For example, if the training videos are not available for some action, a uniform distribution can be assigned to $P(Location|Action)$ and in this way our approach reduces to the standard BoW model without a saliency map for those actions. Moreover, any other saliency approach can be used to determine $P(Location|Action)$. Our experimental results, however, show that learning action signatures from videos results in more informative saliency maps compared to those based on low-level key point detection in spatial space.

7.5 Experimental Results

Experimental evaluation of our approach mainly addresses two tasks: (i) the matching of video-based action-specific regions of interests to important body parts in still images (Section 7.5.1) and (ii) the classification of action images using regions of interest or signatures (Section 7.5.2).

In order to learn action-specific regions of interest, we used videos of different actions available in the Weizmann and KTH datasets. As these videos show little change in background and viewpoint, they allowed us to focus on estimating the importance of different body parts for different actions. In particular, we considered the following actions: *Bending*, *Claping*, *Jacking*, *Punching*, *Running*, *Walking*, and *Waving*. We used the bounding boxes provided by [Yao et al., 2010] and resized them to a common size of 96 × 64 pixels. To determine optical flows, we considered the methods due to Lucas-Kanade [Lu-

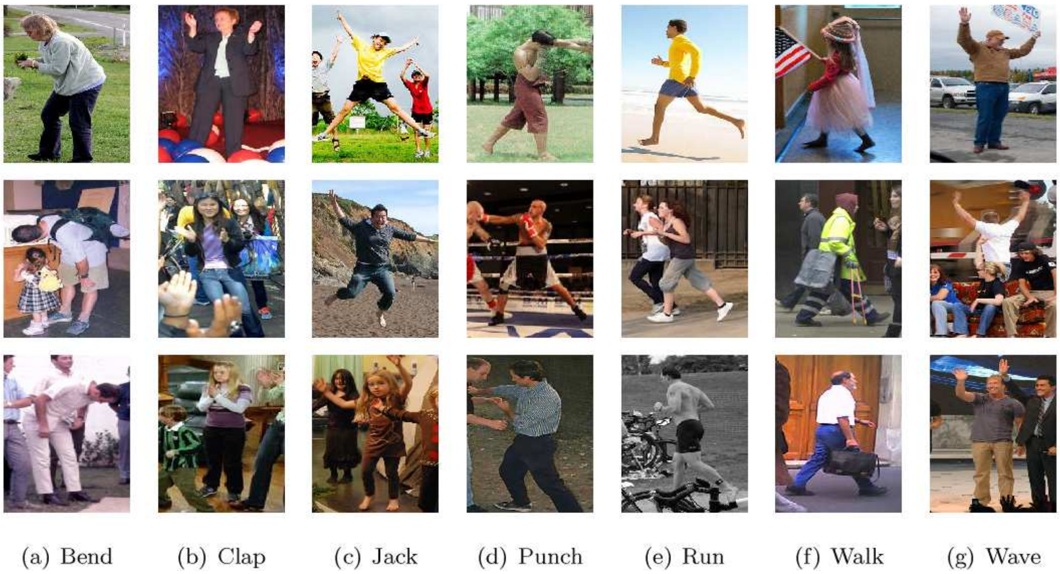


Figure 7.7: Examples of images showing different actions.

cas and Kanade, 1981] and Farnebäck [Farnebäck, 2003]. In both cases, we used the corresponding OpenCV implementations. However, similar to [Wang et al., 2011], we finally adopted the Farnebäck algorithm as we observed a higher efficiency and robust performance in the extraction of our actions signatures. All of the results reported in this section were obtained using 200 basis flows $\mathbf{w}_i$.

In order to evaluate the proposed approach on the target domain, i.e. still images, we collected 270 images from the H3D [Bourdev and Malik, 2009] and the VOC2011 [Everingham et al.] datasets, which we also resized to a resolution of 96 × 64 pixels. Each of these images shows a person performing an action. Figure 7.7 gives several example images for each action. It is obvious that most of the images include background clutter and occlusion.

7.5.1 Interest Regions and Salient Body Parts

We evaluated how far regions of interest extracted by our approach described in Section 7.3 correspond to locations of human body parts in real images. In this regard, we exploited the manually annotated positions of limbs or joints that are available in the H3D and VOC2011 datasets. In particular, we computed the joint probability distribution of actions, interest regions, and body parts.

Given the locations of a body part b_j in an image of action a_i, we have

$$\begin{aligned} P(b_j, \mathbf{w}_k, a_i) &= P(b_j|\mathbf{w}_k, a_i)P(\mathbf{w}_k|a_i)P(a_i) \\ &= P(b_j|\mathbf{w}_k)P(b_j|a_i)P(\mathbf{w}_k|a_i)P(a_i) \end{aligned} \tag{7.9}$$

where $P(b_j|\mathbf{w}_k)$ is chosen to be inversely proportional to the Euclidean distance between the location of b_j and the center of a region in $\mathbf{w}_k$. The prior $P(a_i)$ is assumed to be uniform. The conditional distribution $P(b_j|a_i)$ is obtained by marginalizing over the K bases and all training images corresponding to action a_i. Thus, $P(b_j|a_i)$ can be understood to encode the relative importance of different body parts for an action a_i.

We made use of all 270 annotated images and determined the joint distribution of actions, interest regions, and body parts. For each of the selected action categories, we are interested in estimating the most likely location of 13 body parts or joints including, for example, the head, feet, knees, hips, shoulders, elbows, and hands.

We compared the interest regions resulting from our approach to key points extracted by two popular detectors, the Harris corners detector [Harris and Stephens, 1988] and the SIFT key points detector [Lowe, 2004]. In each case, we selected key points with the highest response in every image, assigned them to their nearest annotated body part, and normalized the resulting histogram. In this way, we obtained a stochastic vector for each action by iterating over all images of that action – thus mimicking the conditional distribution $P(b_j|a_i)$ discussed in Section 7.3.

Figure 7.8 compares results from our method for extracting interesting regions from video data to the ones obtained from using Harris and SIFT key points. The visualization emphasizes the relative importance of the body parts for a particular action given different sampling schemes. The size of the plotted body part corresponds to the frequency or the importance of locations around that part. The stick figures are shown in a standing pose only for better visualization i.e. in order to avoid occlusion of some joints due to large size of others.

We observe that, in the case of Harris and SIFT key points, head and feet dominate other limbs regardless of action (Fig. 7.8 rows 1 and 2). Moreover,

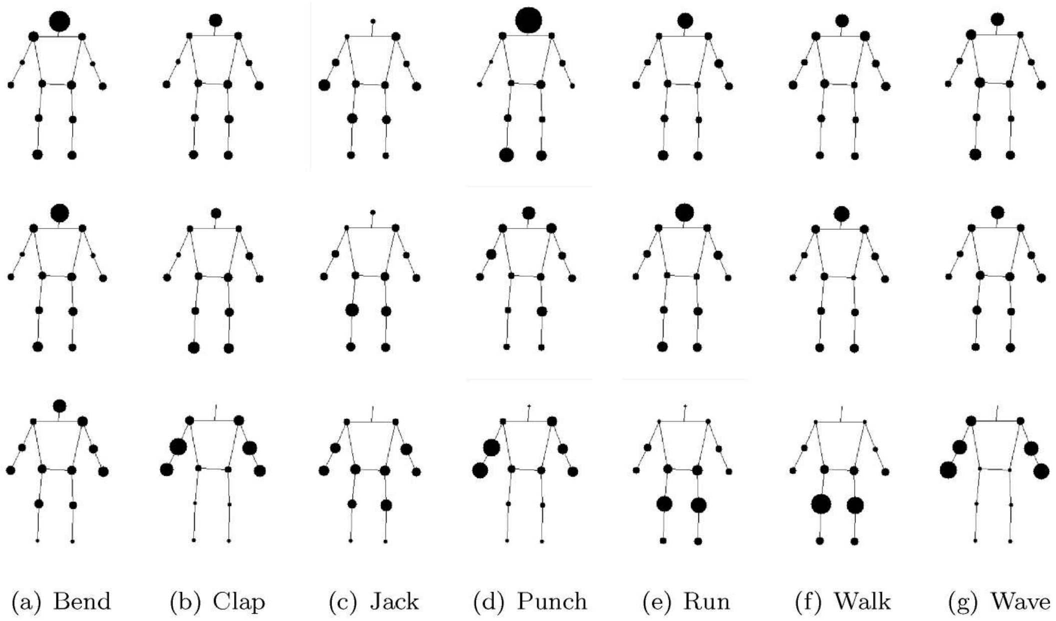


Figure 7.8: Stick figures depicting the relevance of different body parts for different actions. Important key points computed using the Harris detector (first row) and SIFT detector (second row) hardly correlate to action-specific body parts; interest regions from our approach correlate better (third row).

in these cases, the probabilities for other body parts are almost uniformly distributed and do not convincingly relate to different actions. For example, body parts naturally related to the activity of *clapping*, i.e. elbows and hands, achieve rather low scores compared to other limbs or parts.

Our approach, on the other hand, exhibits logically coherent relationships between body parts and actions (Fig. 7.8 third row). Compare, for example, the varying importance of different body parts for *clapping* and *running*. Clearly, the lower body parts are dominant for the action of running while the arms are of higher importance for the action of clapping. From the perspective of body parts, observe that, for instance, the head is less relevant for actions such as *claping* or *running* compared to *bending*. We therefore expect that this favorable property of our approach can ultimately be used to establish rigorous and discriminative action models through an informed sampling phase that focuses on the distinctive patterns of an action rather than on random or coarse sampling.

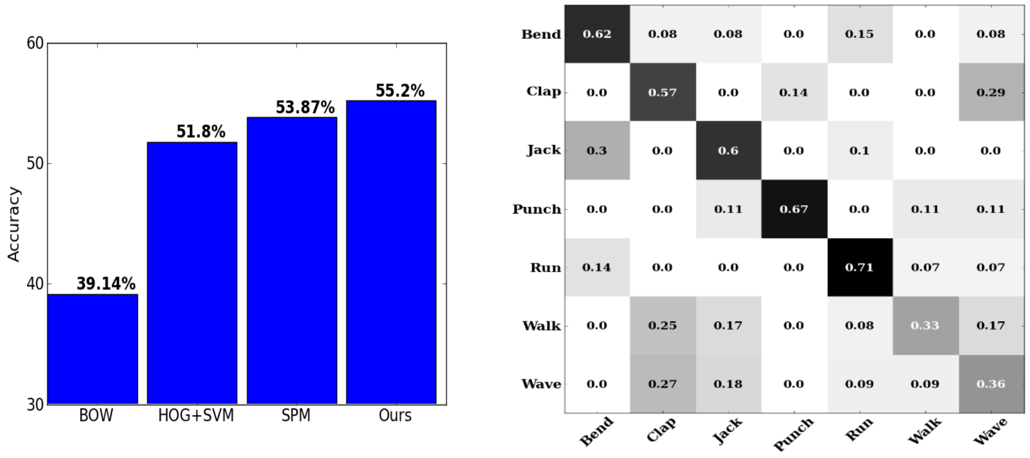


Figure 7.9: (left) Classification accuracies using various approaches. (right) Confusion matrix for action classification by the proposed approach

7.5.2 Action Classification

After establishing the effectiveness of our approach in identifying salient regions (body parts) for different actions, we then evaluated its utility for action recognition in still images. To this end, we used all images in our dataset without considering any annotation of body parts or joints. In the training phase, we again divided each image into L rectangular regions (see Section 7.4) where we used a 16×16 grid of overlapping cells. Within each cell, we extracted an L_2 normalized 6-bin local histogram of oriented gradients. To compute an optimal vocabulary $\mathbf{V}$ of visual words, we considered different numbers of words and observed good performance for 64 to 120 words. Next, we estimated the probability distributions $P(Word|Location)$ and $P(Word|Action, Location)$ (see Eq.7.6). Note again that $P(Location|Action)$ is determined by action signatures learned from videos.

To classify a given query image, we identify the best matching visual word $\mathbf{v}(l_k)$ at each location and assign the query image the action with the highest likelihood as follows

$$\text{argmax}_{a_i} \sum_{k=1}^{L} P(a_i|\mathbf{v}(l_k), l_k) \tag{7.10}$$

$$= \text{argmax}_{a_i} \sum_{k=1}^{L} \frac{P(\mathbf{v}(l_k)|a_i, l_k)P(l_k|a_i)}{P(\mathbf{v}(l_k)|l_k)} \tag{7.11}$$

In a 5-fold cross validation, our approach achieved an average accuracy of about 55.2% for action recognition. We compared our approach to the standard BoW approach based on spatial pyramid binning (SPM) and to other global template matching techniques using the histograms of oriented gradients (HOG). For SPM, we densely sampled local features every 6 pixels at multiple scales and computed SIFT features [Lowe, 2004]. Then we constructed a *code book* using K-means and used the code book to encode the extracted local features from the image. The codes were pooled afterwords over three levels of the spatial pyramid of the image plane per [Lazebnik et al., 2006]. Figure 7.9 compares our approach to these baseline methods. The confusion matrix obtained from using our approach is shown in Fig. 7.9. Note that most of the ambiguity is due to the actions of *waving*, *jacking*, and *clapping*, as all of them share similar body part appearances. On the other hand, actions such as *punching* and *bending* were accurately classified by our approach.

Note that the closest SPM model depends on a dense (uniform) sampling to extract local image features. This produces large numbers of local image descriptors which may be unnecessary for action recognition. Given a test image, our approach would require determining features only at most salient locations, whereas SPM-based models need to compute all features at all location at multiple levels. Existing literature suggests that although SPMs are better than HOG+SVM and BOW, saliency information can be used to achieve similar or higher performance by using fewer features [Vig et al., 2012, Mathe and Sminchisescu, 2012]. This is also affirmed by our results, where we showed that the proposed saliency approach could achieve results better than SPM.

Finally, we evaluated the impact of the number of training videos on our saliency map $P(\text{location}|\text{action})$. In the extreme case, when training videos are not present, our model assigns a uniform distribution of location importance to $P(\text{location}|\text{action})$. As discussed earlier, this is similar to an orderless BoW model, and as anticipated, the performance was close to it (42.71%). By using only two training videos per action, our model achieved an accuracy as high as HOG+SVM (51.02%). The best performance of 55.2% was obtained by considering only 4 videos per action. Adding more training videos did not significantly improve the overall classification. To conclude, it appears that a few videos are sufficient for constructing discriminative action signatures that cover all different execution styles of an action.

7.6 Conclusion

We have presented a novel approach to human action recognition in still images based on the notion of regions of interest. Since human activities are inherently dynamic phenomena, we analyzed optical flow fields extracted from simple video sequences showing human activities in order to learn about salient regions for action recognition. We employed non-negative matrix factorization to obtain sets of basis flows which were found to be indicative of the location of different limbs or joints in different actions. Then we exploited this saliency in a generative Bayesian model for action classification which integrates information as to regions of interests and local spatial image features.

Through experimental validation, we found a clear relationship between regions of interest determined by our approach and action-specific body parts. Our approach achieves higher action recognition accuracies than three recent baseline methods. This is noteworthy since our approach fundamentally differs from existing approaches for action recognition in still images. Firstly, although we consider rather low-level signal properties of videos of activities, the characteristics of optical flow enable us to identify locations of body parts whose articulation distinguish an action from others. Our approach, unlike common bag-of-features approaches, facilitates an informed sampling of key points in still images. Secondly, the concept of action signatures provides probabilistic templates for pose-based recognition. Compared to common approaches based on distributed pose representations, our approach does not require thorough manual annotation of images or frames and thus offers better scalability and convenience for large datasets. Also, compared to conventional part-based approaches, our approach does not assume an underlying elastic model of the body but provides priors even for cluttered scenes or images of partly occluded human bodies. To conclude we have established a baseline for video-based feature selection and classification towards action recognition in still images.

Given spatial features that could better encode individual body parts, our approach may perform even better than it does now. While our approach determines action signatures by NMF of controlled videos, our generative framework allows the use of saliency $P(Location|Action)$ derived by any other method, e.g. by Kinect skeletal data.

Summary

In this chapter, we discussed the issue of human action recognition in still images using automatically generated saliency maps. The state-of-the art approaches to action recognition in still images either compute a large number of features on multiple spatial levels (SPM); make use of *poselets*, the manually determined patches representing (partial) human pose; or use a low-level key point detector in appearance space (e.g. Harris and SIFT) and apply the Bag-of-Word model for image representation. These approaches either involve manual efforts (e.g. poselets) or suffer from uninformedness of the key points (e.g. appearance-based key points). We have proposed a saliency-based approach that exploits controlled videos to determine the respective salient regions for different human actions. These saliency maps, called action signatures, are built automatically from non-negative matrix factorization of optical flow fields. Unlike *poselets*, this approach involves no manual effort and therefore offers better scalability and convenience on large datasets. Our empirical results show that, compared to Harris and SIFT key points, the action signatures better correspond to the location of salient limbs or joints in different actions in target images. Accordingly, we presented a generative Bayesian framework which integrates information as to regions of interests, local spatial image features, and human actions. Through experimental validation on a challenging image set, we showed that our approach achieves higher action recognition accuracies than three recent baseline methods.

Chapter 8

Conclusion and Future Perspectives

8.1 Conclusion

Human activity recognition is a growing area of research. Major components of an automatic activity recognition system include data sources, feature extraction, feature representation, model building, and classification. A great amount of research has been devoted to this field in the last decade. These efforts have led to: (a) sophisticated feature extraction and saliency approaches such as optical flow, spatial or spatio-temporal interest points, and dense trajectories (b) informative local or global feature representations such as silhouette contours, local binary patterns, and motion boundary histograms, and (c) discriminative or generative classification models such as latent support vector machines, convolutional neural networks, and hidden Markov models. Over time, some of these components have become the *de facto* standards in certain domains. For example, support vector machine classification with Bag-of-Words features representations is often used for human action recognition in large unconstrained data without paying attention to the underlying distribution of the sparse high dimensional data. As another example, appearance-based corner points and manually labeled *poselets* are employed frequently in image classification. This trend is ongoing, and as a result, issues of scalability and efficiency remain open.

The efficiency and scalability of vision-based human action recognition systems also needs to be addressed seriously because past few years have seen rapid developments in terms of advancement in image capturing devices, the availability of large-scale of multimedia data, and the emergence of sophisticated application areas. The demand to develop efficient large-scale activity recognition systems such as visual surveillance and content-based image/video retrieval is inevitable. This thesis treats the problem of action recognition from different application perspectives, discusses important issues of scalability and efficiency, and proposes several simple yet powerful methods to human action recognition. For this purpose, we categorize activity recognition in four scenarios of increasing complexity: controlled scenarios such as indoor video surveillance, uncontrolled and unconstrained video databases such as YouTube, multimodal emerging environments such as those captured through Kinect sensory, and still images. Activity recognition in these scenarios poses challenges of different natures and scales. So no single common framework can be suitable to all. Therefore, our scientific investigations are based on general principles, e.g. simple features, discriminative poses, and latent factors, as well as on domain-specific constraints, e.g. high dimension low sample size data. We have achieved state-of-the art activity recognition performance on a number of benchmark datasets representing various levels of complexity.

For action recognition in a controlled video environment, where person localization and background subtraction can be reliably achieved, we proposed a novel key pose based method in Chapter 3. Each class is represented by a collection of key poses obtained by k-mean clustering of corresponding frames. It turns out that representing a human pose by a simple contour-based feature can achieve high recognition accuracy. Further, we devised a mutual information weighting scheme that determines most discriminative key poses based on inter-class and intra-class variation. Experimental evaluation on single and multi-view benchmark datasets show that learning weights (latent factors) for key poses of different actions enhances classification performance. Chapter 4 extended the applicability of pose-based classification to large-scale person identification by gait sequence. We have achieved state-of-the-art recognition performance on a multi-view gait dataset with 124 classes. Efficient feature extraction, low dimensionality, and instance-based classification leads to real-time classification. Such demonstration of accuracy and efficiency is of great importance towards large-scale activity recognition problems (e.g. surveillance) that are characterized by involving missing data, observational latency, and online classification.

The demand of activity recognition in more complex scenarios, i.e. large databases of unconstrained images and videos, has recently led a paradigm shift from traditional pose-based activity recognition approaches to feature-intensive approaches. Currently, most successful approaches in this domain extract vast amount of spatio-temporal features around interest points or dense trajectories and apply off-the-shelf classifiers [Laptev et al., 2008, Kliper-Gross et al., 2012, Wang et al., 2013] without paying attention to the underlying distribution of the high dimensional data. Chapter 5 discussed the issues of classification of such high dimension low sample size data (HDLSS). It turns out that most traditional proximity-based methods such k-Nearest Neighbors and Nearest Convex Hull suffer from lack of neighborhood structure in HDLSS feature spaces. The discriminative methods such as SVMs, on the other hand, lead to severe over-fitting. Based on the statistical studies that prove that HDLSS data lie on a simplex [Hall et al., 2005, Donoho and Tanner, 2005], we suggest representing each class as an affine hull spanned by its instances. Compared to the tight representations such as convex hull and hyperdisk, an affine hull offers a loose structure without bounding the class regions. This loose structure becomes of vital importance in HDLSS, where the likelihood of new data to lie within an existing neighborhood is negligible. The existing SVD- based method of nearest affine hull classification [Cevikalp et al., 2008] is time and computation intensive and may not be applicable to large-scale problems. To this end, we proposed a novel approach: NAH-lsq based on QR factorization and least squares. The QR factorization takes advantage of the fact that the HDLSS data is usually full rank. Extensive experimentation on 5 different datasets and 8 different methods shows that NAH-lsq outperforms other instance-based methods and exhibits performance competitive or superior to SVMs. For online settings (a major application area), the proposed NAH-lsq method is faster than online-SVMs, as SVMs would need complete retraining. To our knowledge, this is the first such study in the domain of human activity recognition. We expect that this investigation of classifiers and our experimental results will motivate activity recognition researchers to revisit instance-based classification and simple representation schemes.

A significant recent development in computer vision is the introduction of low-cost 3D image capturing, e.g. Kinect. Much research is now devoted to recognize human actions in such skeletal or depth data. Most successful approaches in this domain build on principles of existing 2D pose- or part-based methods [Wang et al., 2012, Ofli et al., 2013]. Nevertheless, high recognition performance is achieved by these methods by exploiting the rich pose and motion cues. The richness of such imagery also invites researchers to think of *out-of-the-box* applications. In this line, we investigated an interesting area motivated by studies in psychophysics which state that the people tend to perform different actions in their own personal manner. A novel hierarchical approach was introduced in Chapter 6 that enables the recognition of both actions and actors using only motion cues from 3D data. The proposed approach is based on conventional action recognition and bilinear modeling of contents and styles [Tanenbaum and Freeman, 2000]. Experimental evaluation on different kinds of activity sequences (motion capture, skeletal, motion history volumes) affirms that people tend to perform different actions in different styles and that our approach is most suitable for this problem [Cheema et al., 2013]. Furthermore, it can be directly applied to in-depth human action recognition in interactive environments such as multi-player video games or smart homes.

Finally, we showed that how information from controlled videos can be used to efficiently classify unconstrained action images (Chapter 7). Common approaches to activity recognition in still images build histograms of local features determined around key points that are selected through running a key point detector or by dense random sampling of pixel coordinates. These key points are not a-priorily related to human activities and thus might not be very informative for action recognition. We proposed a novel approach that applies non-negative matrix factorization (NMF) to optical flow fields from simple action videos to find salient regions in image space. The NMF produces a set of sparse basis which are then combined to determine action-specific saliency maps called action signatures. Consequently, a generative Bayesian framework based on action signatures, local spatial image features, and human actions was presented. Experimental evaluation on a challenging image dataset showed that our identified interest regions (peaks of action signatures), when compared to appearance-based key points, are highly correlated to those body parts that characterize corresponding actions. Accordingly, high accuracy was achieved for action classification in images. Unlike *poselets*, we do not require manually labeled template patches and offer better scalability and convenience on large datasets. Since acquisition of skeleton data is becoming efficient and reliable (e.g. through by OpenKinect [11] wrappers for Kinect), our generative

[11] http://openkinect.org

framework can be directly used to build action signatures from skeletal data in future.

To conclude, we have presented several novel approaches for efficiently recognizing human activities in different scenarios. Our research mainly benefits from efficient feature representations, instance-based learning, and latent factor models. In particular, we showed how latent factor models such as QR factorization, bilinear factorization, and non-negative matrix factorization can be employed for efficient and informed human action recognition. We also highlight limitations of main-stream approaches e.g. due to high dimension, low sample size video data; as well as opportunities of future research e.g. multi-modal and multi-factor human activity recognition.

8.2 Future Perspectives

Throughout our research, we have posed novel questions and presented our research on those issues. Some ideas, such as action recognition using discriminative key poses or contour-based features [Cheema et al., 2011], are already receiving significant attention from the vision community [Chaaraoui et al., 2012, Liu et al., 2013, Climent-Prez et al., 2013, Zanfir et al., 2013]. We believe that there remain challenges and opportunities for future research at each scale.

Our approach to classification of HDLSS human activity data (Chapter 5) is arguably the first such attempt. We have shown that affine hull-based representation of low sample size data can remedy the issues caused by lack of proper neighborhood in HDLSS data. This approach can be tailored to other domains such as video set classification in *Big Data* (e.g. YouTube portal). For example, large-scale classification of weakly labeled YouTube videos using video co-watch data can benefit from affine hull-based representation. Note that several image set classification methods (e.g. [Hu et al., 2012] and [Wu et al., 2013]) already employ affine hull representation to project image subsets as points on a Grassmanian manifold for classification. Harandi et al. [2013] modeled Auto Regression Moving Average (ARMA) features along a simple video as an affine subspace, represented it as a point on a Grassmanian manifold, and employed discriminant analysis for action recognition. However, their frame-by-frame feature representations may become cumbersome for unconstrained videos where the compact bag-of-features representation is more practical. There are hardly any approaches to video set classification, in particular to human action recognition in Big Data. However, with access to computational resources and a large amount of (weakly) labeled video data, affine hull-based representation may be of great impact in future research on Big Vision[12].

Another promising area of future research is multi-factor analysis of activity videos. We have shown that it is possible to determine multiple factors, namely actions and actors, that characterize a human motion sequence in 3D (Chapter 6). Extending our hierarchical bilinear framework towards incorporating multiple factors, including action, view, actor, scenario, camera motion, and visibility, is a natural next step. In fact, Cuzzolin [2014] has recently proposed multilinear classifiers which use higher order singular value decomposition (HOSVD) and asymmetric modeling. While that model has shown robust results on gait recognition, it would be interesting to see how such multilinear models behave for action recognition in unconstrained images and videos. Such an investigation may need databases larger than the existing realistic datasets such as HMDB [Kuehne et al., 2011] and UCF50 [Reddy and Shah, 2013], since they do not contain samples for all possible combinations of the factors.

[12] https://sites.google.com/site/bigvision2012/

BIBLIOGRAPHY

A. Agarwal and B. Triggs. A local basis representation for estimating human pose from cluttered images. In *Asian Conference on Computer Vision*, 2006.

J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):16:1–16:43, 2011.

J. Ahn, J. S. Marron, K. M. Muller, and Y Chi. The high-dimension, low sample-size geometric representation holds under mild conditions. *Biometrika*, 94(3):160–766, 2007.

S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *International Conference on Computer Vision*, 2007.

H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

S. Baysal, M. C. Kurt, and P. Duygulu. Recognizing human actions using key poses. In *International Conference on Pattern Recognition*, 2010.

K. P. Bennett and E. J. Bredensteiner. Duality and geometry in svm classifiers. In *International Conference on Machine Learning*, 2000.

H. Bilen, V. P. Namboodiri, and L. J. Van Gool. Object and action classification with latent window parameters. *International Journal of Computer Vision*, 106(3):237–251, 2013. doi: 10.1007/s11263-013-0646-8.

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

A. Bobick and J. Davis. The recognition of human movement using temporal templates. *Transactions on Pattern Analysis and Machine Intelligence*, 23 (3):257–267, 2001.

A. Bolivar-Cime and J. S. Marron. Comparison of binary discrimination methods for high dimension low sample size data. *Journal of Multivariate Analysis*, 115:108–121, 2013.

A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6:1579–1619, 2005.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, 1992.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *International Conference on Computational Statistics*, 2010.

L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *International Conference on Computer Vision*, 2009.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

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

H. Cevikalp and B. Triggs. Large margin classifiers based on convex class models. In *ICCV Workshops*, 2009.

H. Cevikalp and B. Triggs. Hyperdisk based large margin classifier. *Pattern Recognition*, 46(6):1523–1531, 2013.

H. Cevikalp, B. Triggs, and R. Polikar. Nearest hyperdisk methods for high-dimensional classification. In *International Conference on Machine Learning*, 2008.

H. Cevikalp, B. Triggs, H. S. Yavuz, Y. Kk, M. Kk, and A. Barkana. Large margin classifiers based on affine hulls. *Neurocomputing*, 73(16-18):3160–3168, 2010.

A. A. Chaaraoui, P. Climent-Prez, and F. Flrez-Revuelta. An efficient approach for multi-view human action recognition based on bag-of-key-poses. In *Third International Workshop on Human Behavior Understanding*, 2012.

J. M. Chaquet, E. J. Carmona, and A. Fernandez-Caballero. A survey of video datasets for human action and activity recogniotion. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.

E. Chaung and C. Bregler. Mood swings: expressive speech animation. *ACM Transactions on Graphics*, 24(2):331–347, 2005.

M. S. Cheema, A. Eweiwi, and C. Bauckhage. Human activity recognition by separating style and content. *Pattern Recognition Letters*, In Press, 2013. doi: http://dx.doi.org/10.1016/j.patrec.2013.09.024.

M.S. Cheema, A Eweiwi, and C. Bauckhage. Who is doing what? simultaneous recognition of actions and actors. In *International Conference on Image Processing*, 2012a.

M.S. Cheema, A Eweiwi, and C. Bauckhage. Gait recognition by learning distributed key poses. In *International Conference on Image Processing*, 2012b.

S. Cheema, A. Eweiwi, C. Thurau, and C. Bauckhage. Action recognition by learning discriminative key poses. In *ICCV Workshops*, 2011.

C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian. Frame difference energy image for gait recognition with imcomplete silhouettes. *Pattern Recognition Letters*, 30(11):977–984, 2009.

C. Chen, J. Liang, and X. Zhu. Gait recognition based on improved dynamic bayesian networks. *Pattern Recognition*, 44(4):988–995, 2011.

H. S. Chen, H.T. Chen, Y. W. Chen, and S. Y. Lee. Human action recognition using star skeleton. In *International Workshop on Video Surveillance and Sensor Networks*, 2006.

P. Climent-Prez, A. A. Chaaraoui, J. R. Padilla-Lpez, and F. Flrez-Revuelta. *Optimal joint selection for skeletal data from RGB-D devices using a genetic algorithm*, volume 7630 of *Advances in Computational Intelligence: Lecture Notes in Computer Science*. Springer, 2013.

A. Coates and A. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *International Conference on Machine Learning*, 2011.

R. T. Collins, R. Gross, and S. Jianbo. Silhouette-based human identification. In *International Conference on Face and Gesture Recognition*, 2002.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and application. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):257–267, 2000.

J. E. Cutting and L. T. Kozlowski. Recognizing friends by their walk: gait perception without familiarity cues. *Bulletin of the Psychonomic Society*, 9 (5):353–356, 1977.

F. Cuzzolin. Using bilinear models for view-invariant action and identity recognition. In *International Conference on Computer Vision and Pattern Recognition*, 2006.

F. Cuzzolin. Multilinear classifiers. Submitted to Transactions on Pattern Analysis and Machine Intelligence, 2014.

N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*, 2006.

Y. Dedeoglu, B. Ugur Toereyin, U. Gueduekbay, and A. E. Cetin. Silhouette-based method for object classification and human action recognition. In *ECCV workshop on Human Computer Interaction*, 2006.

V. Deltaire, I. Laptev, and J. Sivic. Recognizing human actions in still images: A study of bag-of-features and part-based representations. In *British Machine Vision Conference*, 2010.

P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Neural Information Processing Systems*, 2004.

D. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 102(27):9452–9457, 2005.

Y. Du, F. Chen, and W. Xu. Human interaction representation and recognition through motion decomposition. *Signal Processing Letters*, 14(12):952–955, 2007.

A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *International Conference on Computer Vision*, 2003.

A. M. Elgammal and C. S. Lee. Separating style and content on a nonlinear manifold. In *International Conference on Computer Vision and Pattern Recognition*, 2004.

H. J. Escalante, I. Guyon, V. Athitsos, P. Jangyodsuk, and J. Wan. Principal motion components for gesture recognition using a single-example. http://arxiv.org/abs/1310.4822, 2013.

M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html.

Abdalrahman Eweiwi, Muhammad Shahzad Cheema, and Christian Bauckhage. Discriminative joint non-negative matrix factorization for human action classification. In *German Conference on Pattern Recognition, Germany*, 2013.

G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*, 2003.

A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

P. Felzenszwalb, R. Girschick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. S. Lempitsky. Hough forests for object detection, tracking, and action recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 33:2188–2202, 2011.

G. Martin Garca, S. Frintrop, and A. B. Cremers. Attention-based detection of unknown objects in a situated vision framework. *German Journal of Artificial Intelligenz*, Springer, 2013.

A. Gilbert, J. Illingworth, and R. Bowden. Action recognition using mined hierarchical compound features. *Transactions on Pattern Analysis and Machine Intelligence*, 33:883 – 897, 2011.

M. Goffredo, I. Bouchrika, J. N. Carter, and M. S. Nixon. Self-calibrating view-invariant gait biometrics. *Transactions on Systems, Man and Cybernetics*, 40(4):997–1008, 2010.

B. Gruenbaum. *Convex polytopes, Graduate Texts in Mathematics*, chapter Neighborly polytopes, pages 122–129. Springer, New York, 2nd edition, 2003.

S. Hadfield and R. Bowden. Hollywood 3D: Recognizing actions in 3D natural scenes. In *International Conference on Computer Vision and Pattern Recognition*, 2013.

P. Hall, J. S. Marron, and A. Neeman. Geometric reprsentations of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, 2005.

J. Han and B. Bhanu. Individual recognition using gait energy image. *Transactions on Pattern Analysis and Machine Ingelligence*, 28(2):316–322, 2006.

M. T. Harandi, C. Sanderson, S Shirazi, and B. C. Lovell. Kernel analysis on grassmann manifolds for action recognition. *Pattern Recognition Letters*, 34 (15):1906–1915, 2013.

C. Harris and M. Stephens. A combined corner and edge detection. In *Alvey Vision Conference*, 1988.

E. Herbst, P. Henry, X. Ren, and D. Fox. Toward object discovery and modeling via 3-d scene comparison. In *International Conference on Robotics and Automation*, 2011.

Y. Hu, A. S. Mian, and R. Owens. Face recognition using sparse approximated nearest points between image sets. *Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1992–2004, 2012.

N. Ikizler-Cinbis and D. A. Forsyth. Searching for complex human activities with no visual examples. *International Journal of Computer Vision*, 30(3): 337–357, 2008.

N. Ikizler-Cinbis, R. G. Cinbis, and S. Sclaroff. Learning actions from the Web. In *International Conference on Computer Vision*, 2009.

A. Iosifidis, A. Tefas, and I. Pitas. Person specific activity recognition using fuzzy learning and discriminant analysis. In *European Signal Processing Conference*, 2011.

A. Iosifidis, A. Tefas, and I. Pitas. Activity-based person identification using fuzzy representation and discriminant learning. *Transactions on Information Forensics and Security*, 7(2):530–542, 2012.

L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10–12):1489–1506, 2000.

Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):852–872, 2000.

A. Jalal, M. Z. Uddin, J. T. Kim, and T. S. Kim. Recognition of human home activities via depth silhouettes and transformation for smart homes. *Indoor and Built Environment*, 21(1):184–190, 2011.

H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *International Conference on Computer Vision*, 2007.

H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *International Conference on Computer Vision*, 2013.

H. Jiang and D. R. Martin. Finding actions using shape flows. In *European Conference on Computer Vision*, 2008.

Y. Jiang, Q. Dai, X. Xue, W. Liu, and C. Ngo. Trajectory-based modeling of human actions with motion reference points. In *European Conference on Computer Vision*, 2012.

Y. Jiang, S. Bhattacharya, S. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, 2(2):73–101, 2013.

L. Jianyi and Z. Nanning. Gait history image: A novel temporal template for gait recognition. In *International Conference on Multimedia and Expo*, 2007.

S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *International Conference on Computer Vision and Pattern Recognition*, 2011.

S. Ke, H. L. Thuc, Y. J. Lee, J. N. Hwang, J. H. Yoo, and K. H. Choi. A review on video-based human activity recognition. *Computers*, 2(2):88–131, 2013.

Y. Ke., R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *International Conference On Computer Vision*, 2005.

J. Kilner, J-Y.Guillemaut, and A. Hilton. 3D action matching with key-pose detection. In *CVPR Workshop on Search in 3D and Video*, 2009.

A. Kläser, M. Marszaek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, 2008.

S. Klement, A. M. Mamlouk, and T. Martinetz. Reliability of cross-validation for svms in high-dimensional, low sample size scenarios. In *International Conference on Artificial Neural Networks*, 2008.

B. Klingenberg, J. Curry, and A. Dougherty. Non-negative matrix factorization: Ill-posedness and a geometric algorithm. *Pattern Recognition*, 42(5): 918–928, 2008.

O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf. Motion interchange patterns for action recognition in unconstrained videos. In *European Conference on Computer Vision*, 2012.

A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *International Conference on Computer Vision and Pattern Recognition*, 2010.

H. Kuehne, H. Jhaung, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *International Conference on Computer Vision*, 2011.

W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Pairwise shape configuration-based PSA for gait recognition under small viewing angle change. In *International Conference on Advanced Video and Signal-Based Surveillance*, 2011.

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

I. Laptev and T. Lindeberg. Space-time interest points. In *International Conference on Computer Vision*, 2003.

I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *International Conference on Computer Vision and Pattern Recognition*, 2006.

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. In *International Conference on Computer Vision and Pattern Recognition*, 2011.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–799, 1999.

E. Lee and L. J. Schulman. Clustering affine subspaces: Hardness and algorithms. In *ACM-SIAM Symposium on Discrete Algorithms*, 2013.

W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPR Workshop on Human Activity Understanding from 3D Data*, 2010.

J. Liu, Y. Yang, I. Saleemi, and M. Shah. Learning semantic features for action recognition via diffusion maps. *Computer Vision and Image Understanding*, 116(3):361–377, 2012.

L. Liu, Y. Yin, and W. Qin. Gait recognition based on outermost contour. In *International Conference on Rough Sets and Knowledge Technology*, 2010.

L. Liu, L. Shao, X. Zhen, and X. Li. Learning discriminative key poses for action recognition. *Transactions on Cybernetics*, 43(6):1860 – 1870, 2013.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Imaging Understanding Workshop*, 1981.

S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *International Conference on Computer Vision and Pattern Recognition*, 2011.

M. Malinowski and M. Fritz. Learning smooth pooling regions for visual recognition. In *British Machine Vision Conference*, 2013.

M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *International Conference on Computer Vision and Pattern Recognition*, 2009.

F. Martnez-Contreras, C. Orrite-Urunuela, J. Herrero J., H. Ragheb, and S. A. Velastin. Recognizing human actions using silhouette-based hmm. In *International Conference on Advanced Video and Signal Based Surveillance*, 2009.

S. Mathe and C. Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *European Conference on Computer Vision*, 2012.

P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: action recognition through the motion analysis of tracked features. In *ICCV Workshop on Video-Oriented Object and Event Classification*, 2009.

R. Munoz-Salinas, R. Medina-Carnicer, F. J. Madrid-Cuevas, and A. Carmona-Poyato. Depth silhouettes for gesture recognition. *Pattern Recognition Letters*, 29(3):319 – 329, 2008.

F. Murtagh. The remarkable simplicity of very high dimensional data: Application of model-based clustering. *Journal of Classification*, 26(3):249–277, 2009.

G. I. Nalbantov, P. J. F. Groenen, and J. C. Bioch. Nearest convex hull classification. Technical report, Econometric Institute Erasmus University Rotterdam, 2007.

P. Natarajan and R. Nevatia. View and scale invariant action recognition using multiview shape-flow models. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

B. Ni, G. Wang, and P. Moulin. Human activity detection from RGBD images. In *ICCV Workshop on Consumer Depth Cameras for Computer Vision*, 2011.

J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learing of human action categories using spatial-temporal words. In *British Machine Vision Conference*, 2006.

F. Niu and M. Abdel-Mottaleb. HMM-based segmentation and recognition of human activities from video sequences. In *International Conference on Multimedia and Expo*, 2005.

E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*, 2006.

F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley MHAD: A comprehensive multimodal human action database. In *IEEE Workshop on Applications of Computer Vision*, 2013.

N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *International Conference on Multimodal Interfaces*, 2002.

M. Perera, S. Kudoh, and K. Ikeuchi. Keypose and style analysis based on low-dimensional representation. *Journal of Information Processing Society of Japan*, 50:1234–1249, 2009.

Ronald Poppe. A survey of vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

H. Ragheb, S.A. Velastin, P. Remagnino, and T. Ellis. Human action recognition using robust power spectrum features. In *International Conference on Image Processing*, 2008.

R. Rammal, J. C. Angles D'auriac, and B. Doucot. On the degree of ultrametricity. *Le Journal de Physique – Letters*, 46(20):945–952, 1985.

K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vison and Applications*, 24(5):971–981, 2013.

M. Rohrbach, S. Amin, M Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *International Conference on Computer Vision and Pattern Recognition*, 2012.

P. M. Roth, T. Mauthner, I. Khan, and H. Bischof. Efficient human action recognition by casecade linear classification. In *International Conference on Computer Vision*, 2009.

S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *International Conference on Computer Vision and Pattern Recognition*, 2012.

C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

C. Schueldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *International Conference on Pattern Recognition*, 2004.

P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *International Conference on Multimedia*, 2007.

Kazuyuki Sekitani and Yoshitsugu Yamamoto. A recursive algorithm for finding the minimum norm point in a polytope and a pair of closest ppoint in two polytopes. *Mathematical Programming*, 61:233–249, 1993.

G. Sharma, F. Jurie, and C. Schmid. Discriminative spatial saliency for image classication. In *International Conference on Computer Vision and Pattern Recognition*, 2012.

G. Sharma, F. Jurie, and C. Schmid. Expanded parts model for human attribute and action recognition in still images. In *International Conference on Computer Vision and Pattern Recognition*, 2013.

D. Shin, H-S. Lee, and D. Kim. Illumination-robust face recognition using ridge regressive blinear models. *Pattern Recognition Letters*, 29(1):49–58, 2008.

Ji. Shuiwang, X. Wei ; Y. Ming, and Y. Kai. 3D convolutional neural networks for human action recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

S. Singh, S. A. Velastin, and H. Ragheb. MuHAVi: A multicamera human action video dataset for the evaluation of action recognition methods. In *International Conference on Advanced Video and Signal Based Surveillance*, 2010.

V. K. Singh and R. Nevatia. Action recognition in cluttered dynamic scenes using pose-specific part models. In *International Conference on Computer Vision*, 2011.

C. Sminchisescu, A. Kanaujia, and D. N. Metaxas. Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding*, 104(2–3):210–220, 2006.

Y. Song, L. Goncalves, and P. Perona. Unserpervised learning of human motion. *Transactions on Pattern Analysis and Machine Intelligence*, 25(7):814 – 827, 2003.

K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human action classes from videos in the wild. Technical Report CRCV-TR-12-01, University of Centeral Florida, 2012.

R. Souvenir and J. Babbs. Learning the viewpoint manifold for action recognition. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *International Conference on Computer Vision*, 2011.

J. B. Tanenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

E. Taralova, F. T. Frade, and M. Hebert. Source constrained clustering. In *International Conference on Computer Vision*, 2011.

D. M. J. Tax and R. P. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004.

M. Tenorth, J. Bandouch, and M. Beetz. The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *ICCV Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences*, 2009.

K. A. Thoroughhman and R. Shadmehr. Electromyographic correlates of learning an internel model of reaching movements. *Joural of Neuroscience*, 19(19): 8573–8588, 1999.

C. Thurau. Behavior histograms for action recognition and human detection. In *ICCV Workshop on Human Motion*, 2007.

C. Thurau and V. Hlavac. Pose primitive based human action recognition in videos or still images. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage. Convex nonnegative matrix factorization for massive datasets. *Knowledge And Information Systems*, 29(2):457–478, 2011.

S. Todorovic. Human activities as stochastic kronecker graphs. In *European Conference on Computer Vision*, 2012.

A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *International Conference on Computer Vision and Pattern Recognition*, 2011.

K. N. Tran, I. A. Kakadiaris, and S. K. Shah. Modeling motion of body parts for action recognition. In *British Machine Vision Conference*, 2011.

P. Turaga, R. Chellappa, V. S. Subrahmanian, and Octavian Udera. Machine recognition of human activities: a survey. *Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.

A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

A. Vieira, E. Nascimento, G. Oliveira, Z. Liu, and M. Campos. Stop: Space-time Occupancy Patterns for 3D action recognition from depth map sequences. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications.*, 2012.

E. Vig, M. Dorr, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *European Conference on Computer Vision*, 2012.

P. Vincent and Y. Bengio. K-local hyperplane and convex distance nearest neighbor algorithms. In *Neural Information Processing Systems*, 2001.

H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *British Machine Vision Conference*, 2009.

H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *International Conference on Computer Vision and Pattern Recognition*, 2011.

H. Wang, A. Kläser, C. Schmid, and C.L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

J. Wang, M. She, S. Nahavandi, and A. Kouzani. A review of vision-based gait recognition methods for human identification. In *International Conference on Digital Image Computing: Techniques and Applications*, 2010.

J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *International Conference on Computer Vision and Pattern Recognition*, 2012.

L. Wang and D. Suter. Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model. In *International Conference on Computer Vision and Pattern Recognition*, 2007.

L. Wang, T. Tan, W. Hu, and H. Ning. Automatic gait recognition based on statistical shape analysis. *Transactions on Image Processing*, 12(9):1–13, 2003a.

L. Wang, T. Tan, H. S. Ning, and W. M. Hu. Silhouette analysis-based gait recognition for human identification. *Transactions on Pattern Analysis and Machine Ingelligence*, 25(12):1505–1518, 2003b.

Y. Wang, K. Haung, and T. Tan. Human activity recognition based on R transform. In *International Conference on Computer Vision and Pattern Recognition*, 2007.

D. Weinland and E. Boyer. Action recognition using exemplar-based embedding. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2):249–257, 2006.

D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224–241, 2011.

G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European Conference on Computer Vision*, 2008.

P. Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149, 1976.

Y. Wu, M. Minoh, and M. Mukunok. Collaboratively regularized nearest points for set based recognition. In *British Machine Vision Conference*, 2013.

C. Y. Yam, M. S. Nixon, and J. N. Carter. Gait recognition by walking and running: A model-based approach. In *Asian Conference on Computer Vision*, 2002.

J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden Markov model. In *International Conference on Computer Vision and Pattern Recognition*, 1992.

W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *International Conference on Computer Vision and Pattern Recognition*, 2010.

A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *International Conference on Computer Vision and Pattern Recognition*, 2010.

A. Yao, J. Gall, G. Fanelli, and L. Van Gool. Does human action recognition benefit from pose estimation? In *British Machine Vision Conference*, 2011a.

A. Yao, J. Gall, and L. Van Gool. Coupled action recognition and pose estimation from multiple views. *International Journal of Computer Vision*, 100 (1):16–37, 2012.

B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human aciton recognition by learning bases of action attributes and parts. In *International Conference on Computer Vision*, 2011b.

M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall. *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, chapter A survey on human motion analysis from depth data, pages 149–187. Springer Lecture Notes in Computer Science, 2013.

A. Yilmaz and M. Shah. Actions sketch: A novel action representation. In *International Conference on Computer Vision and Pattern Recognition*, 2005.

G. Yu, J. Yuan, and Z. Liu. Propagative Hough voting for human activity recognition. In *European Conference on Computer Vision*, 2012.

S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *International Conference on Pattern Recognition*, 20006. URL `http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp`.

M. Zanfir, M. Leordeanu, and C. Sminchisescu. The Moving Pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In *International Conference on Computer Vision*, 2013.

L. Zhang and X. Lin. Some considerations of classification for high dimension low-sample size data. *Statistical Methods in Medical Research*, 2011. published online DOI:10.1177/0962280211428387.

Y. Zhang, N. Yang, W. Li, X. Wu, and Q. Ruan. Gait recognition using procrustes shape analysis and shape context. In *Asian Concference on Computer Vision*, 2009.

Z. A. Zhu, W. Chen, G. Wang, C. Zhu, and Z. Chen. P-packSVM: Parallel primal gradient descent kernel SVM. In *International Conference on Data Mining*, 2009.