Thomas Mauthner

Learning Activity Recognition from Weakly Labeled Data

DOCTORAL THESIS

to achieve the university degree of
Doktor der technischen Wissenschaften

submitted to

Graz University of Technology

Supervisor

Prof. Dr. Horst Bischof
Institute for Computer Graphics and Vision
Graz University of Technology

Asst. Prof. Dr. Janez Perš
Faculty of Electrical Engineering
University of Ljubljana

Graz, Austria, November 2017

Für meine Frau Angela, die mich mit ihrer Liebe und Zuversicht immer unterstützt hat.

If you want to make an easy job seem mighty hard, just keep putting off doing it.

Wenn Du willst, dass Dir eine leichte Aufgabe richtig schwer erscheint, schieb sie einfach auf.

Olin Miller (1918 - 2002)

Abstract

The development of methods and algorithms for the detection and retrieval of activities in videos has been investigated intensively in recent years. A clear narrowing and annotation of activities is often controversial and in consequence of the large amount of data difficult to be carried out manually. The resultant ambiguities in terms of spatial, temporal and class-specific assignment of data to activities make a robust training of methods for classification or detection difficult. This thesis deals with different forms of ambiguities in the context of activities, and describes methods for more robust automatic assignment of video areas to activity labels. Learned temporal coherence is exploited for improved classification and detection of activities in videos. Furthermore, the thesis deals with the development of algorithms for the automatic determination of contextually prominent regions, as the basis for activity detection. The developed methods are applied in many different ways, evaluated against reference approaches and discussed with regard to future developments in the field of activity recognition.

Kurzfassung

Die Entwicklung von Methoden und Algorithmen zur Erkennung und Detektion von Aktivitäten in Videos wurde in den letzten Jahren intensiv vorangetrieben. Eine eindeutige Eingrenzung und Annotierung von Aktivitäten ist dabei oft schwierig und aufgrund der großen Datenmenge nicht manuell durchführbar. Die daraus resultierenden Mehrdeutigkeiten in Bezug auf räumlich, zeitliche und klassenspezifische Zuordnung von Daten zu Aktivitäten erschweren ein robustes Training von Methoden zur Klassifikation oder Detektion. Die vorliegende Dissertation behandelt unterschiedliche Formen von Mehrdeutigkeiten im Kontext von Aktivitäten, und beschreibt Methoden zur robusten automatischen Zuordnung von Videobereichen zu Aktivitätsbezeichnungen und der Ausnutzung gelernter zeitlicher Zusammenhänge für verbesserte Klassifikation und Detektion von Aktivitäten in Videos. Des Weiteren behandelt die vorliegende Arbeit die Entwicklung von Algorithmen zur automatischen Bestimmung von kontextuell herausragenden Regionen, als Basis für Aktivitätsdetektion. Die entwickelten Methoden werden in vielfältigen Evaluierungen mit Referenzansätzen verglichen und in Hinblick auf die zukünftige Entwicklung im Bereich Aktivitätserkennung diskutiert.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Place Date Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Ort Datum Unterschrift

Acknowledgments

Many people have accompanied me through my years at the Institute of Computer Graphics and Vision. We shared a really intense but interesting time, for which I am very thankful. First and foremost I would like to thank Prof. Bischof for giving me the opportunity to work within this active research area and especially for his support during the finalization of this thesis. I also want to thank Dr. Janez Perš for spending his knowledge and time for being my second supervisor. I consider myself fortunate to have had the opportunity of working with a number of colleagues who shared their knowledge with me during countless discussions, influenced my research and acted as co-authors. At this point I would like to thank Peter Roth, who is an outstanding personality and always inspired me with his tireless passion for research. Further I would like to mention Horst Possegger, a hard-working colleague and friend, who could always perfectly convert my abstract ideas into viable approaches. During all the years I have seen ups and downs and learned how important support and assistance between colleagues is, also apart from research related topics, to create a motivating and balanced working environment. Here I would like to emphasize Sabine Sterning, Stefan Kluckner, Georg Waltner, Peter Kontschieder and Mahdi Jampour, representative for all colleagues I had the pleasure to work with.

Contents

List of Figures

List of Tables

1 Introduction

Understanding human behavior is a key technology for future seamless connection between users and IT infrastructure. The full variability of possible human activities cannot be captured with sensors not including visual information. Therefore, activity related research has gained much interest in the computer vision community within recent years. Computer vision based activity recognition is a very diverse topic, concerning both the various fields of application and scenarios as well as manifold kinds of problem statements and input representations. Nowadays, we see activity or action recognition in various context and temporal extent. Ranging from temporally delimited actions (raise left arm), to simple concatenations of motion primitives (bend down and up) or temporal repetitions of actions (running, waving) up to more complex chronology of activities (cooking a certain meal). The meaning of an action can change by context information of objects (picking up box) or other surrounding actions (standing vs. queuing up). We can create a vast amount of scenarios and challenges with the above named points.

Compared to related fields within computer vision like, e.g., facial gesture recognition or object detection, activity recognition is less definite and divided into several sub-fields. Nevertheless, it is one of the fastest growing topics with a variety of possible industrial applications from user interaction via gestures and significant motions for controlling human computer interfaces, to automatic content analysis in social media and TV broadcasts. Therefore, the research underlying to this thesis has seen many twists and turns as well as influences over the years.

Depending on the application and context, various kinds of activity related examples have been proposed over the years. This vast variety of approaches, modalities and datasets usually require individual preparation and annotation guidelines, depending on how the term activity is understood in a specific case. Coming from the area of tracking and sports game analysis, the focus of this thesis is human centered activity recognition as a sequence of detecting a person of interest or player, tracking them over time for building a temporal cue and finally recognizing what he or she is doing at a certain point of this temporal interval. Let us call this person of interest performing actions an "actor", for the remainder of this section. This actor centered approach rises several issues from which the research topics in this thesis have been derived from. Putting the actor in the focus seems to be logically, but a majority of activity recognition related work is in fact performing solely video classification without paying attention on foreground and background separation. One may argue that background and context defines the meaning of a motion sequence and is therefore a vital part of the activity. But related research has shown that algorithms are often biased by specific background features due to a lack of diversity in training and test data. Focusing on the actor(s) gets especially important if background becomes misleading or no global activity label is valid for a complete designated video. Furthermore, background can become negligible when it comes to activity recognition in surveillance and sports games analysis, where grass is always green, a pitch looks always the same or parking facilities give no further contextual information. Focusing on the actor implies a variety of challenges, which form the basic motivations of this thesis. But, what defines an activity, and when can we speak of weak annotated?

Let us define the root cause of this challenge, which is rather simple: we do not have enough information about the activities. Missing information leads to ambiguities, as an activity is defined by the three main questions: who, when and what? We can speak of weak annotation in the absence of at least one of this three needed points. Staying in the field of actor centered activities the question raises, how to get the information that we need to describe these three inputs? The first one defines the spatial parameters given as the position, image annotation or segmentation of an actor or group of actors. The second question concerns the temporal parameters, the information about begin, ending and maybe maximum of activity intensity. The last one defines the activity itself by a given activity class, group or label. Now let us recap what happens if one of these three constraining parameters is missing. In such circumstances we speak of an ambiguous labeled activity, meaning at least one parameter is open and has to be found by optimization. In other words, when we speak of weakly labeled data it is the opposite of fully labeled or fully supervised data, where all three constrains are given. Related to this three constrains, ambiguities can have the following reasons:

Missing spatial information lacks the description concerning spatial extent of an activity, defined by the actor(s) or objects of interaction. Annotation may be more complex compared to objects as it often varies over time. With solely given global activity class labels on video level, spatial (and temporal) annotation is missing at all.

Missing temporal information is mostly related to missing definition of the activity climax. Furthermore, unconstrained videos which are not directly limited to begin and end of one specific activity, but instead contain unrelated content before and after a certain activity of interest, fall into this category. Such uncertainties hamper training and make activity localization in videos an attractive research area.

Missing label information often comes with concurrent activities or interactions. As soon as several actors perform their motions in parallel simple general object detectors or proposals are only helpful for separating background motion. The task of learning the correct activity class associated to those proposals becomes really challenging.

In this thesis we have focused on approaches for tackling this ambiguities, and to allow for training from such weakly labeled data, missing one of the three constraining parameters of who, when and what. The results of this research are grouped into three related chapters of this thesis.

Starting with chapter 3, temporal relations of features and motions during activities are examined. As a result, two individual tasks are formulated. First, temporal feature weighting to support individual emphases of motion and appearance based information, for generating more discriminative activity representation. This approach tackles the temporal ambiguity and, in contrast to previous related approaches, allows for non-global feature weighting. Second, a concept for exploiting activity motion learned from training data is proposed, which supports unsupervised adaption of activity detectors to new datasets and background conditions. Instead of solving spatial ambiguities by manual annotation a forward-backward motion consistency based regularization is applied, showing much better predicting results compared to optical flow propagation.

Chapter 4 tackles the ambiguities of missing label association in the case of contemporary performed actions in parallel, or without clear begin and ending of activities within a video. Automatically finding the object of interest within a set of images is a widespread problem in computer vision, typically referred to as ”multiple-instance-learning” or ”non-negative-data-mining”. This task becomes more challenging when multiple different labels are possible per image or video. Such concurrent activities occur in several fields of application for activity recognition, e.g., sports, surveillance or human computer interaction. To avoid annotation effort, a ranking by pair-wise code-books training is developed, showing superior performance on multi-label ambiguity tasks.

Finally chapter 5 exploits the concept of saliency for finding activity related areas in videos, therefore tackling spatial and temporal ambiguities. Without adding any labeling information, the circumstance that activities form salient areas within the video volume, can be exploited to find activity relevant areas. Recent research has emphasized the need for analyzing salient information in videos to minimize dataset bias or to supervise weakly labeled training of activity detectors. Furthermore, we can find outcomes in related literature which emphasize success of weighting object motion information for activity classification. In contrast to previous approaches, the proposed saliency estimation method does not rely on training information given by either eye-gaze or annotation data, but rather propose a fully unsupervised algorithm to find salient regions within videos or images. Based on that, a novel video saliency detection method is presented, to support human activity recognition and weakly supervised training of activity detection algorithms. In general, the Gestalt principle of figure-ground segregation is enforced for both appearance and motion cues. Evaluations against state-of-the-art approaches on several datasets, including different challenging activity recognition problems for saliency algorithms, demonstrate favorable performance for activity detection and recognition.

The above mentioned research related chapters overlap concerning datasets, features, representation and learning approaches. Nevertheless, they address individual problems and different goals and can therefore be interpreted as individual projects. Thus individually selected related work and experimental evaluations are given at the beginning and end of each chapter, respectively. In addition, chapter 2 defines the common basis, by presenting the evolution of activity recognition during the time of the research underlying to this thesis. Furthermore, it highlights and reviews prominent state-of-the-art in activity related feature description, not mentioned in the individual research chapters.

2 Datasets, Problems and Solution: An Overview on Related Work

2.1 Evolution of Activity Recognition Benchmarks

The field of activity recognition has changed drastically over time, and evolution of datasets gives some insight how the field has changed. Changes may be driven by defining new demands and goals, by new approaches and insights making existing benchmarks obsolete or by new computational methods, which allows for processing more data in more detail and less time. The following summary gives an overview of datasets popular in the past, or recently applied for benchmarks in top-level publications, without claiming completeness. Within their survey paper from 2012, Chaquet et al. [28] already mentioned 68 activity related datasets. A more recent survey by Zhang et al. [192], covering solely the field of activity recognition form RGD-D data, mentions another 39 popular benchmarks. And reviewing recent conference publications gives a list of approximately 38 different datasets, mainly published after 2012. Therefore, the tremendous amount of existing datasets and benchmarks makes it hard to cover the whole field. Nevertheless, we try to describe the evolution within the field of vision based activity recognition and highlighting main developments and pitfalls. For further details we would like to refer to the following survey papers: [4, 28, 88, 123, 175, 192, 195], and the given references within the following sections for more details about specific datasets, approaches and results.

2.1.1 Human Centered Activity Recognition

In the early stages of vision-based activity recognition, research was focused on object-centered single-person activities. Within the seminal work by Bobick and Davis [16] in 2001, the authors proposed so called motion energy images (MEI) and motion history images (MHI). The term ”motion” in this case was estimated by aggregating foreground segmentation over time into either binarized (MEI) of weighted (MHI) representations, see Figure 2.1. Although their approach was solely applicable within static scenes with limited background variability, they already fused different view-points and created a real-time application for interaction of multiple users with a virtual avatar. This research has engaged the community until recently and has resulted in a lot of datasets like Weizmann [57] and KTH [142], which are still part of experimental evaluation in recent publications.

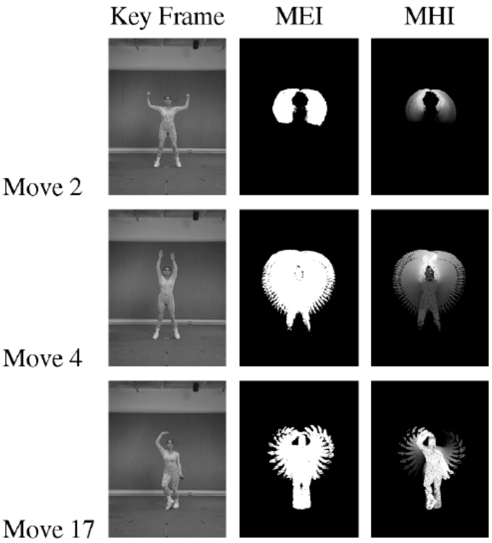


Figure 2.1: Examples for early activity descriptors MEI and MHI as described in [16].

One of the first really widespread activity benchmarks was the Weizmann dataset, recorded in 2005 by Blank et al. [12], which originally contained 81 low resolution videos (180 × 144) of nine subjects performing nine different actions: running, jumping in place, bending, waving with one hand, jumping jack, jumping sideways, jumping forward, walking, and waving with two hands. Subsequently, a tenth action, jumping on one leg (*skip*), was added [57]. The applied approach was based on space-time shape volumes and assumed convenient foreground segmentation. The videos were recorded on homogenous background with static cameras. Despite the fact that the release dates back several years, evaluations on this dataset were done until recently, e.g., [194]. A less well known subset has been later proposed under the name *Weizmann-Robustness*[1], with cluttered backgrounds and high irregularities within the performed activities, see Figure 3.9.

In 2004 the KTH dataset was proposed by Schüldt et al. [142], consisting of 600 videos (160 × 120), with 25 persons performing six human action in four different scenarios: outdoors (*s1*), outdoors with scale variation (*s2*), outdoors with different clothes (*s3*), and indoors (*s4*). This work proposed the detection of space-time interest points (STIP) as features for activity description followed by a classification with support vectors machines (SVM) for the first time, a pipeline that is still applied in the field of activity recognition until now.

[1] http://www.wisdom.weizmann.ac.il/vision/VideoAnalysis/Demos/SpaceTimeActions/DB/robust-deform.zip

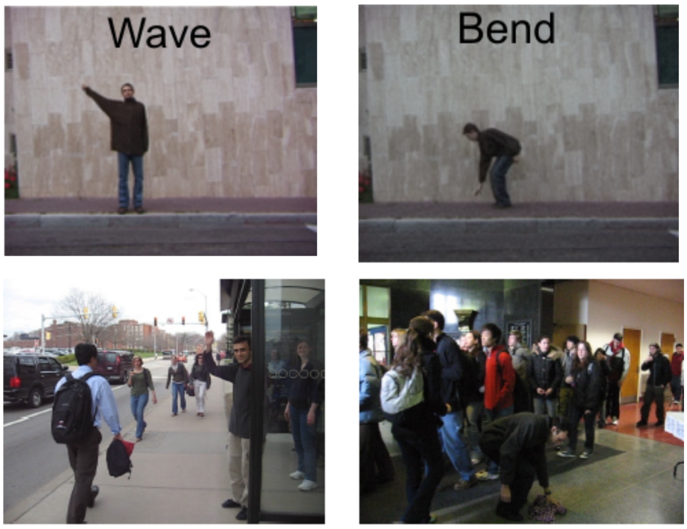


Figure 2.2: Activity exemplars of Weizmann (top-row) and related videos from CMU-crowded with background noise (bottom-row).

Activities in Weizmann and KTH are mainly repetitive, meaning no specific start or end time needs to be defined. Furthermore, they are carried out on homogenous backgrounds and without disturbing background motions or other concurrent activities. Nevertheless, they are still under the top cited dataset within the area of activity recognition.

Much more challenging artifacts were introduced with the CMU-crowded published by Ke et al. [84] in 2007 and MSR-activity datasets (I) and (II) proposed by Yuan et al. [187] and Cao et al. [24], respectively. Their videos include the same activities as Weizmann and KTH, but with more challenging background and the additional task of activity detection rather than just classifying the video. CMU crowded consists of approximately 20 minutes of low resolution videos (160 × 120 pixel), captured by hand-held cameras. Therefore, extracting features from this video quality is challenging. Furthermore, strong background noise and motion (pedestrian and traffic) complicate the action detection. Similar to the Weizmann dataset, annotated activities are ”bend”, ”wave” and ”jumping-jack”. For that reason, the dataset was often applied for showing transfer learning capabilities.

Similar to that, the MSR action datasets contains three activities from the KTH set, namely ”clapping”’, ”waving” and ”boxing” within an urban environment or indoor scenes. The videos were taken with a static camera, minimizing motion artifacts compared to CMU-crowded, with 320 × 240 resolution. Overall 70 videos were collected in both MSR sets, with the indention of training on KTH data and solely testing the detection capabilities on MSR videos. All threes sets were designed for detection rather than recognition as several activities could appear within on video (especially within MSR), and background activities and motion would be too ambiguous.

Further surveillance related datasets can be found in the literature. From indoor

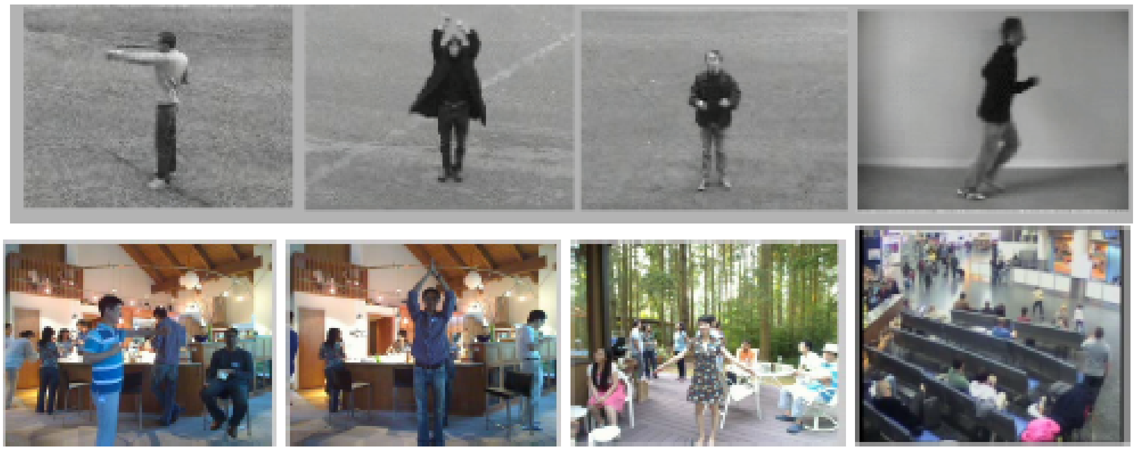


Figure 2.3: Activity exemplars of KTH (top-row) and related videos from MSR with background noise (bottom-row). Image taken from [24].

scenarios, like the CAVIAR[2] project (Context Aware Vision using Image-based Active Recognition) started in 2003, to outdoor scenarios like the BEHAVE interactions test case scenarios by [15], different perspectives on human interactions in UT-interaction [139] and elevated positions or drones in UT-Tower. More recent benchmarks and results can be found on the VIRAT[3] [119] data and most recent the MERL Shopping dataset by Singh et al. [150].

A complete new problem formulation was proposed in the dataset collected by Xu et al. [179], named the A2D dataset and benchmark for action recognition and segmentation with multiple classes of actors. The authors define the problem of actor-action tuples, with 7 different actor classes and 9 different actions possible, see Figure 2.4 for examples. The dataset also provides with pixel-wise actor and action annotation and one third of the videos contains more than one actor or action, which triggers the task of weakly labeled activity recognition, e.g., [181].

2.1.2 Video Classification

Over the years, the focus of activity recognition has changed from detecting activities in surveillance-like videos, to general video recognition based on TV broadcast or YouTube video collections. While activity detection was one of the main goals in above mentioned surveillance related datasets, still a quite challenging task in real world scenarios, the number of actions was limited to daily motions and interaction with other humans or objects. Over time, the focus of research shifted more towards video classification with larger numbers of different action classes.

One dataset defining this new kind of activity recognition problem, was the UCF-

[2] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/caviar.htm

[3] http://www.viratdata.org/

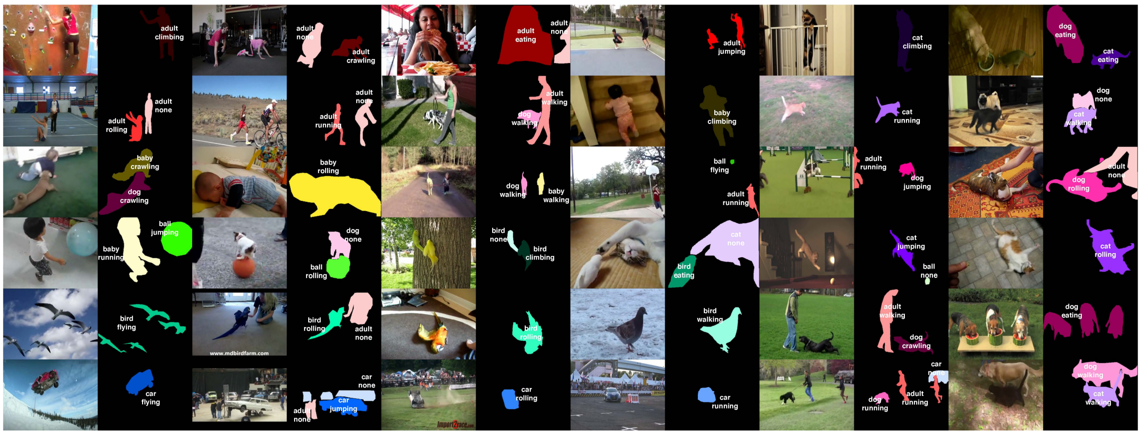


Figure 2.4: A2D dataset for actor-action recognition. The actor concept is extended to moving animals and objects on pixel level, while multiple actors doing different actions may be present. Image taken from [179].

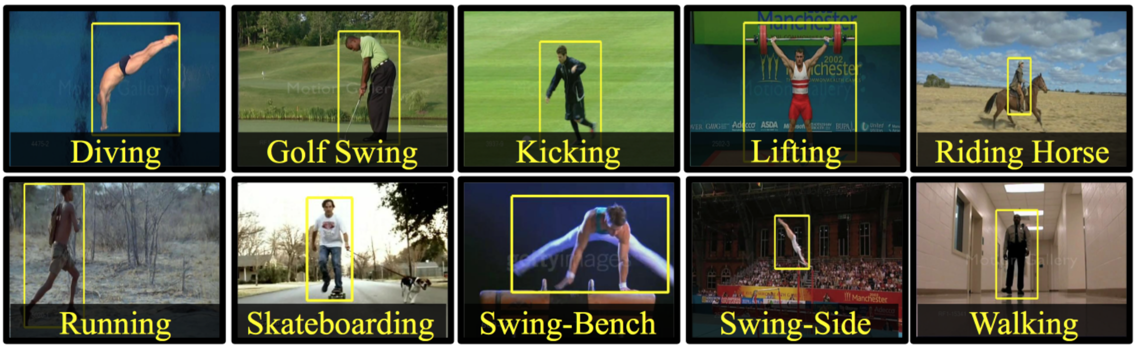


Figure 2.5: Activity exemplars of UCF sports dataset with bounding-box annotation from [133].

sports dataset by Rodriguez et al. [133] in 2008. UCF Sports is a collection of low-quality television broadcasts, containing around 150 videos of various sports. This data set depicts challenging scenarios including camera motion, cluttered backgrounds, and non-rigid object deformations. Furthermore, it provides ground truth bounding box annotations for all activities, namely "diving", "golf", "swinging-side", "kicking", "lifting", "horse back riding", "running", "skating", "swing-bench", "walking". Originally evaluated solely for recognition tasks, in a leave-one-out cross validation scheme, authors mentioned the strong intra-class similarity of several actions within the dataset. As a results, recognition results were more related on a designated background rather than the activity itself. As a consequence, authors proposed different evaluation concepts as 5-fold cross validation or predefined train-test splits. Later, action detection was done too on the UCF-sports dataset. A detailed analysis and overview of published results on the UCF-Sports dataset is given by Soomro and Zamir in [154].

Sports videos in general became an important topic for action datasets. Beside the already mentioned UCF-Sports, the OlympicSports dataset was proposed in 2010 by Niebles

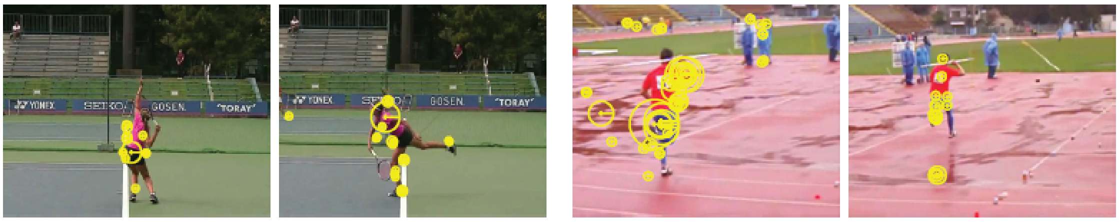


Figure 2.6: Video frames from Olympic Sports dataset with superimposed spatial-temporal interest point detections. Images taken from [116].

et al. [116]. It contains 50 videos from 16 different sports, see Figure2.6. Originally only activity labels per video were given, and the dataset was applied for evaluating recognition results on video level.

The CRCV[4] (Center for Research in Computer Vision) and University of Central California (UCF) generated several action datasets over time, with growing complexity. Subsequent to the UCF-sports action set, UCF11 containing already 11 activities from YouTube videos was published. Videos are organized in 25 groups, where videos in one group may share similar backgrounds and features, since the videos in a group may be obtained from one single video. Splitting training and testing on such groups should give less classification bias to activity related backgrounds. A further extension of this dataset named UCF50, related to the increased number of 50 action categories was published by Reddy and Shah [131]. The video collection was increased a third time named UCF101 [155], containing 101 action categories, additionally divided into five types, namely human-object interaction, body-motion only, human-human interaction, playing musical instruments and sports, see Figure 2.7. UCF101 was the basis for the first THUMOS challenge on activity recognition in untrimmed videos in 2013. Besides recognition on video level, additional ground-truth annotation is given for evaluating activity detection too.

At the same time the authors of [91] published the HMDB51[5] (human motion data base containing 51 action categories). As a novelty to previous datasets, HMBD51 contains additionally meta labels describing the properties of each video. Such meta information contains information about visible body parts (head, upper, lower, full-body), the camera motion (static or motion), camera view (front, back, left, right), number of people involved in the action (one , two, three) and the video quality (good, medium, ok). The authors also proposed a stabilized version with compensated or removed camera motion via RANSAC based stabilization. With a tremendous amount of labeling effort, the authors of [75] generated the JHMDB51[6] (joint-annotated HMDB), containing 21 action classes of JHMDB with ground truth labeling of 15 body joints and body segmentation and motion, as so called puppet-mask and puppet-flow. This work tried to evaluate the influence of fore-

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

[5] http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

[6] http://jhmdb.is.tue.mpg.de

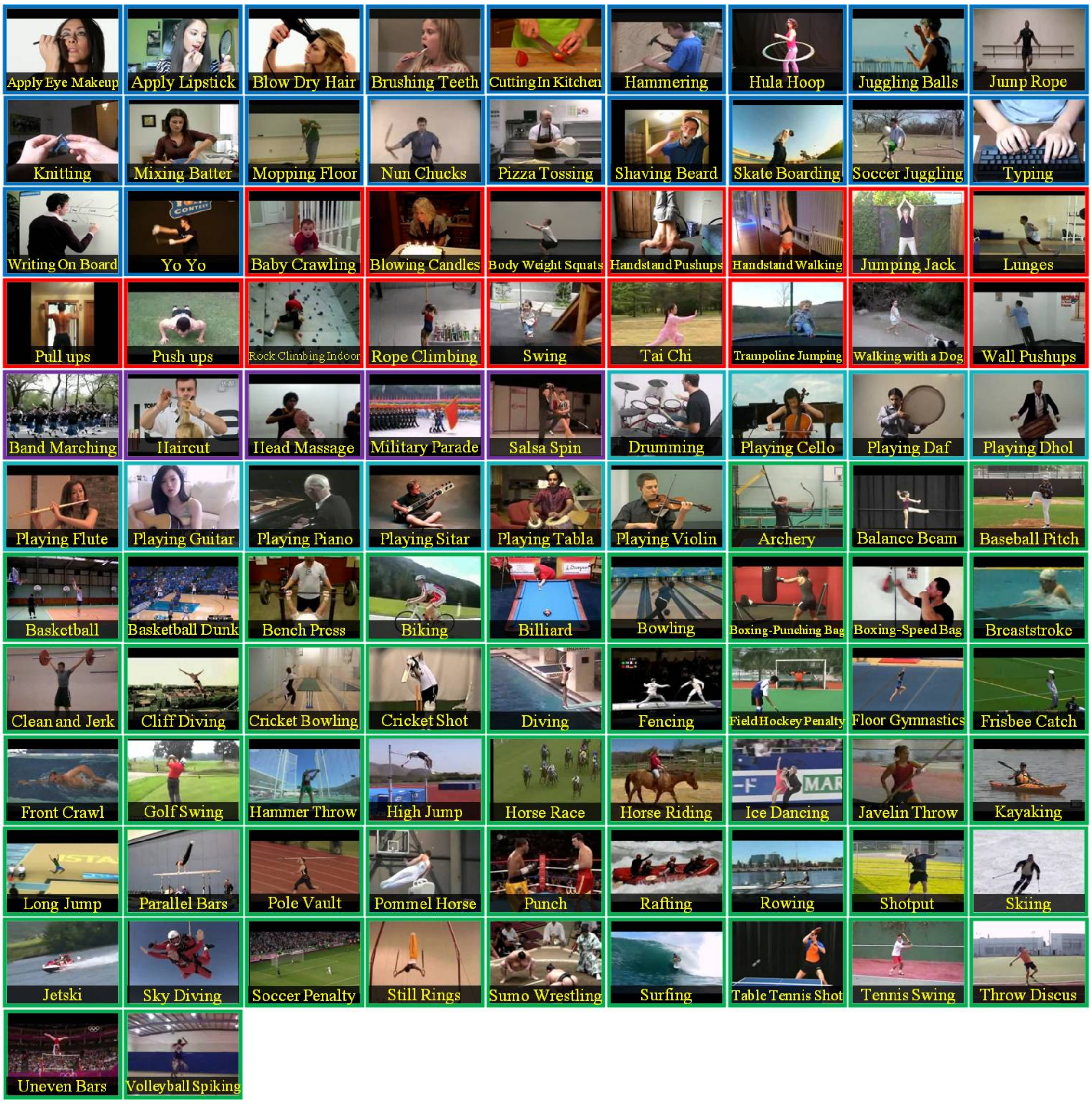


Figure 2.7: Action classes and categories of the UCF101 dataset. Colored frames depict different categories: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments and Sports. Image taken from [155].

ground knowledge on action recognition and to bridge the gap between MOCAP (motion capturing) datasets, captured under controlled environmental settings, and video based action recognition. The findings in [75] suggested that motion based action recognition benefits from knowledge of foreground related optical flow, and that pose-based features outperforms classical low level appearance and motion features. Another interesting finding was that superior optical-flow benchmark algorithms does not lead to better action recognition results.

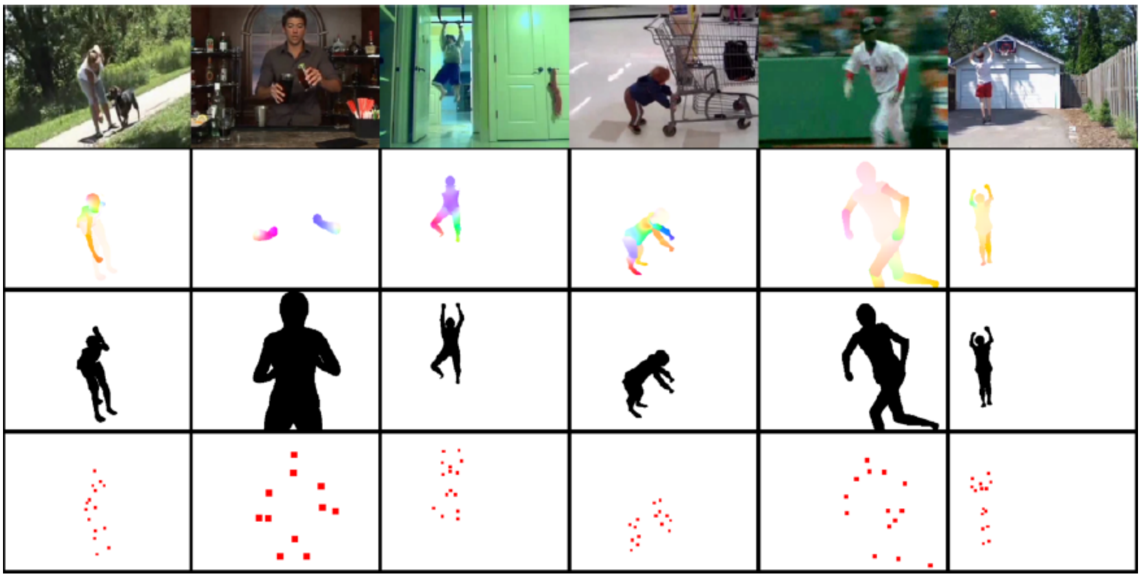


Figure 2.8: Examples from JHMDB videos. Top original video from HMDB dataset. Second row showing the puppet flow, computed from the manually annotated puppet masks in the third row. Bottom depicts positions of annotated body joints. Images taken from [75].

The THUMOS[7] challenge has influenced research in the area of action recognition over several years. Starting with the UCF-101 dataset in 2013, the challenge has extended the video summary piecewise in 2014 and 2015. While previous datasets consisted of trimmed videos, were the activities is present the whole video, THUMOS challenges in 2014 and 2015 allows trimmed videos, based on UCF-101, for training only and validation and test data consisting solely of untrimmed videos. In addition, a background set is introduced containing guaranteed non of the 101 actions, needed for training the capability of temporal action detection in untrimmed videos. Furthermore, bounding box annotations are given for 24 selected actions. The THUMOS13 challenge already organized video based classification for all 101 classes and detection performance on selected 20 classes. Interestingly no action detection submissions had been made in 2013, compared to 16 teams submitting to the classification challenge. This fact shows the complexity of action detection compared to global video classification. In 2014, the number teams competing in action detection increased already to 3 compared to 11 classification teams, while in 2015 again only 1 team submitted detection results compared to again 11 competitors on classifications.

Two more outstanding datasets, with respect to their video content or defined tasks, are the Hollywood and ASLAN datasets. Hollywood and Hollywood2 [107] consist solely of Hollywood movie sequences, containing 12 action categories and 10 scene categories. Kliper-Gross et al. [88] proposed the Action Similarity Labeling (ASLAN) challenge, a binary pair-matching task to decide if two videos belong to the same action category or

[7] http://www.thumos.info/results.html

not.

Most recent datasets and related benchmarks have again increased the amount of data and action classes significantly. The Sports-1M database published by Karpathy et al. [83] contains already an amazing number of 1.133.158 video URLs which have been annotated automatically with 487 sports labels using the YouTube Topics API. One year later Jiang et al. [78] proposed the Fudan-Columbia Video Dataset (FCVID), containing $91,223$ Web videos annotated manually according to 239 categories which are organized in a hierarchy of 11 high-level groups. In contrast to Sports-1M, FCVID dataset contains categories over a vast range of daily living scenarios.

Current state-of-the-art action recognition and detection benchmarks are the ActivityNet[8] challenge and the YouTube8m[9]. ActivityNet data was first release by Caba et al. [44] in 2015 in smaller version (1.2 release $10-2015$, 100 action classes and $10k$ videos). ActivityNet challenge started in 2016 within a CVPR-Workshop. Around $20k$ videos distributed among 200 activity categories, distributed among training, validation and testing in 50%, 25%, and 25% respectively (ActivityNet 1.3 relesase $03-2016$). In addition, the organizers pe-computed features for global video representation, namely ImageNetShuffle-CNN, Motion Boundary Histogram and frame-wise features C3D : Generic Features for Video Analysis. Overall 22 video classification results were submitted, with the winning categories already applying fine-tuned neural networks, and 6 teams tackling the temporal action detection task in untrimmed videos. Methods based on provided feature cannot be found not under the top 3 for classification, while the winning approach for detection was as well based on fine-tuned neural networks. In 2017, the ActivityNet benchmark added more additional challenges and datasets. First, the "Trimmed Action Recognition" on Kinetics[10] human action video dataset (approximately $300,000$ video clips, and covers 400 human action classes with at least 400 video clips for each action class. Each clip lasts around $10s$ and is labeled with a single class. Second, the "Temporal Action Proposals" challenge based on ActivityNet, and finally the "Dense-Captioning Events in Videos" challenge based on ActivityNet Captions dataset.

The currently largest activity dataset is the YouTube8m[11], proposed by [2] in 2016, an recently bench-marked within a CVPR 2017 workshop. The YouTube8m contains the incredible number of more than 8 million videos (given as weblinks) with labels from a diverse vocabulary of 4716 visual entities, with an average of 3.4 labels per video. Labels are generated automatically given as "Knowledge Graph" entities. As this tremendous amount of data would be unfeasible to be handled for most researchers, the team released precomputed features based on publicly available TensorFlow inception network [1] trained on ImageNet [71]. Nevertheless, the total size frame-level features is 1.71 Terabytes and total size of the video-level features is still 31 Gigabytes.

[8] http://activity-net.org/

[9] https://research.google.com/youtube8m/

[10] https://deepmind.com/research/open-source/open-source-datasets/kinetics/

[11] https://research.google.com/youtube8m/index.html

In comparison to general video classification datasets, the number of available multi-camera activity recognition datasets is fairly limited. The IXmas [174] dataset was captured under controlled laboratory conditions, while the VIRAD dataset [119] mainly contains simple person-car interaction. Updated recently by [115] the *Multicamera Human Action Video* (MuHAVi) [151] contains several everyday activities, but again in a total controlled laboratory environment.

Recently, triggered by the availability of cheap cameras for ego-centric perspective videos, a new field of activity recognition research has emerged. Therefore, ego-centric activity recognition [86] and novelty detection in daily activities [5] became of great interest to the research community. The *Georgia Tech Egocentric Activity Datasets* with different version from 2011 [46] to 2015 [100] can be seen as the most active one in this area.

2.1.3 Human Interaction Recognition

The previously-mentioned datasets and related approaches are focused on single activities performed by individuals. Further research aims to classify interactions between individuals, between individuals and objects or between individuals and groups. Datasets such as VideoWeb [37] or UT-Interaction [139] provide more or less constrained data for such interactions and also introduce concurrent activities. The Collective Activity Dataset [32] focuses less on interaction but more on group activities, where the classification of individual activities is supported by similar activities in a local spatial-temporal context. Very recently, a new data set for large-scale group activity recognition has been proposed by [8], where the authors are aiming to solve such a large problem by combining individual and group activities in a top-down / bottom-up hierarchical approach. Less controlled but still scripted interaction scenes have been captured for the already mentioned BEHAVE dataset by [15]. With the rise of Kinect$^{\mathrm{TM}}$sensors first interaction datasets appears for that kind of sensor too, like the SBU-Kinect interaction database in 2012 by Kiwon etal. [188], a set containing two person interaction classification under controlled laboratory conditions with Kinect$^{\mathrm{TM}}$-based RGB-D data. Kong etal. [89] published a benchmark with 8 different interaction in outdoor scenarios with various background within 400 videos. Very recently, a sports related dataset has been proposed to show the benefits of aggregating individual actions into group activity recognition on volleyball scenes by Ibrahim et al. [68].

2.1.4 Activities in Temporal Context

The recognition of daily routines and activities in, e.g., kitchen or living room generated an alternative branch of activity recognition research. Benchmarks like the ICPR 2012 Kitchen Scene Context based Gesture Recognition dataset (KSCGR[12]) or the *MPII Fine-grained Kitchen Activity Dataset* by [134] with 65 cooking activities (see Figure 2.9) define

[12] http://www.murase.m.is.nagoya-u.ac.jp/KSCGR

a separate research field.

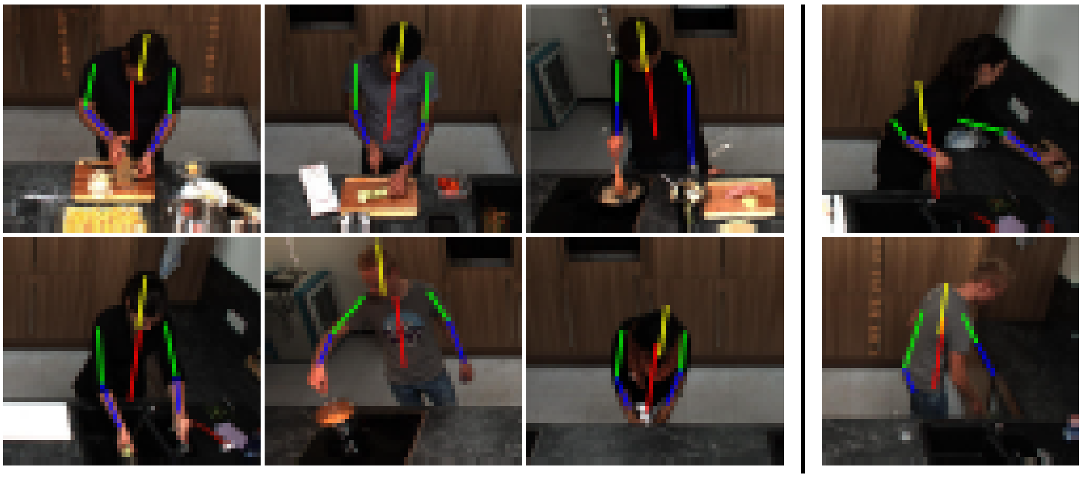


Figure 2.9: Example of typical kitchen dataset scenario with overlayed pose-estimation results. Images taken from [134].

The University of Dundee ”50 Salads” by Stein and McKenna [156], contains 3 high-level annotations like ”cut and mix ingredients”, ”prepare dressing” and ”serve salad”, and 17 low-level annotations. The authors supply video, depth and accelerometer data of devices attached to a knife, a mixing spoon, a small spoon, a peeler, a glass, an oil bottle, and a pepper dispenser.

The recently published Charades database [149] generated a completely new milestone in this area of daily living activity benchmarks. Instead of downloading YouTube videos or capturing videos under laboratory conditions, the authors used Amazon Mechanical Turk (AMT). In a three stage process AMT workers generated scripts, produced the corresponding video and finally the annotation. The Charades dataset contains over 9000 videos, with temporally annotated intervals for 157 action classes and labels for 46 objects. Besides activity recognition, automatic scripting generation for video description is an additional goal of the benchmark.

2.1.5 Alternative Sensors

The introduction of the Microsoft Kinect$^{\text{TM}}$has pushed this field into an application area within consumer electronics, and has also triggered further research on this very constrained topic of RGB-D indoor data, see the review paper by Zhang et al. [192] for state-or-the-art summary of existing methods and datasets. In general, the task of recognizing and detecting human centered activities in indoor scenes has drifted towards RGB-D data. Existing solutions like Kinect$^{\text{TM}}$, stereo-cameras, MOCAP (motion capturing) systems or different type of wearable sensors offers the possibilities to capture human pose information in a much more detailed resolution as compared to standard YouTube videos. On the other hand, capturing the data has to be done during training and testing with comparable settings and hardware. Nevertheless, such approaches safeguard their surplus value in constrained indoor or gesture control scenarios, but as mentioned before in 2.1.2, bridging the gap between video and MOCAP could deliver superior approaches in the future(see, e.g., results in [75]).

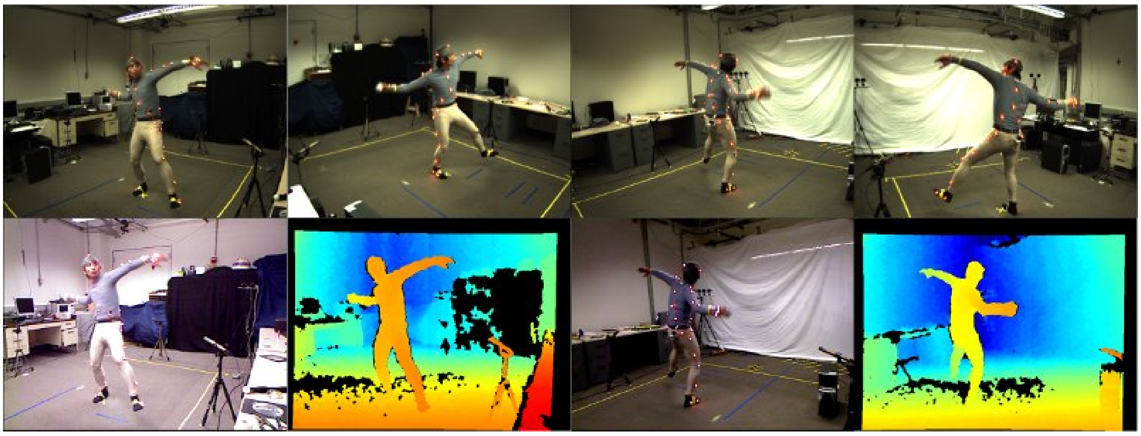


Figure 2.10: Comparison of sensor based motion capture (images with reflecting markers visible) and Kinect$^{\mathrm{TM}}$depth data (false-color visualization). Images taken from [118].

Before affordable sensors changed the scene, requiring MOCAP data was an complex and expensive task, requiring special hardware and setups(see Figure 2.10). First datasets like HDM05 [111] already contained 70 classes with up to 50 repetitions, captured with retro-reflective markers and 12 special cameras, where annotating the ground-truth could quite a challenge [112]. With the release of the Kinect$^{\mathrm{TM}}$sensor and the active support of research teams working with it, large group of new datasets appear, e.g., MSR-Action3D [98], MSRC-12 [50], MSR-DailyActivity-3D [169], UTKinect [178] or Florence-3D [143]. These datasets vary in number activtiy classes, repetition or number of actors, but all focus on indoor activity recognition. Targeting on computer games, the authors of the G3D-Gaming [13] created a dataset capturing persons playing Kinect$^{\mathrm{TM}}$games, later updated for two person examples in *G3Di* [14]. An interesting compilation of scenarios was collected for the ICPR-HARL 2012 competition by the authors of [177], summarized in the LIRIS human activities dataset.

A very impressive fusion of different sensors is available with the *Berkeley Multimodal Human Action Database* (MHAD) [118], see Figure 2.11. The authors simultaneously captured synchronized MOCAP data, stereo cameras, several Kinect$^{\mathrm{TM}}$sensors, accelerators and microphone recordings, as depicted in Figure 2.11.

While specialized depth sensors may simplify several vision tasks like detection and pose estimation, robustness to various view-points is still a challenging problem. Within the last years several dataset tackling this issue have been published and triggered further

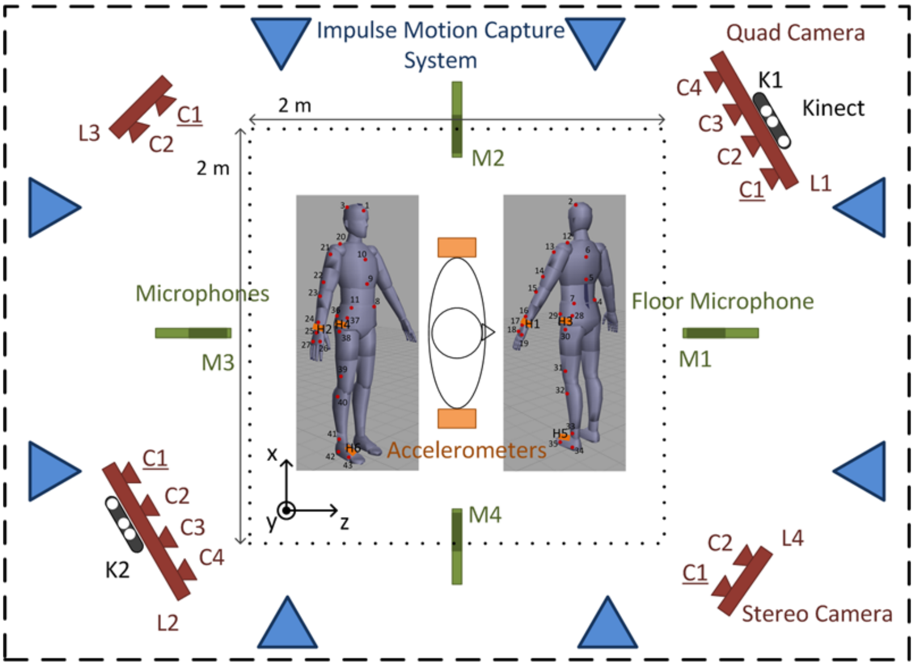


Figure 2.11: Different sensors applied in the MHAD dataset of [118].

research as the *ROSE NTU RGB-D* [144], the *Northwestern-UCLA Multiview Action3D* by [170] dataset with 3 different Kinect view-points for cross-view evaluations or the *UWA3D Multiview Activity* [126] and *UWA3D Multiview Activity II* [125] datasets.

2.2 Features for Activity Recognition

Feature description has always been inspired by the more prominent image classification and object detection research. Nevertheless, representative description of activities relies on some additional specific constrains on time and motion, not necessary for vanilla image classification tasks. In the following, the most important approaches are revised to give an overview of the related work.

2.2.1 Interest-Point Detection for Activity Recognition

Detecting points of interest has a long tradition in computer vision. The basic idea is based on detecting textured regions, meaning area in images with many edges, corners and structure. Homogenous and plain regions does not allow for re-identification due to ambiguities, while corners and boundaries emphasize structural breaks and retrievable points. Within videos, such structural breaks do not only appear in the spatial 2D image coordinates but also in the temporal domain. Therefore, we are speaking of spatio-temporal interest points or STIP in general. The purpose of STIP detections is on one hand the highlighting of moving parts of a video, therefore likely to be part of an activity, and on the other to minimize the computational effort of densely processing the whole video space. In the following we would like to highlight three very prominent version of STIP detectors, namely Harris3D by Laptev [94], Cuboid-Detector [38] by Dollar et al.and V-Fast by Yu et al. [186].

Harris3D as proposed by Latepv et al. [94], is an extension of the Harris corner detector [61] in the time domain. Derivatives on input Image I_t are done for all three dimensions (x,y,t), see Equ. 2.1 after Gaussian smoothing with $g(x, y, t; \sigma^2, \tau^2)$, where σ and τ represents independent spatial and temporal variance, respectively. The 3×3 second moment matrix U is build following Equ. 2.2, where '$*$' is the convolution operator and L_x, L_y and L_t are the scale space derivatives of the video volume $I(x, y, t)$. The variance values for calculating U differ from calculating the derivatives by a constant s. The Harris3D interest point detector is capable of operating on different scales via parametrization.

$$\begin{aligned} L_x &= \partial_x \left(g(\cdot; \sigma^2, \tau^2) * I\right) \\ L_y &= \partial_y \left(g(\cdot; \sigma^2, \tau^2) * I\right) \\ L_t &= \partial_t \left(g(\cdot; \sigma^2, \tau^2) * I\right) \end{aligned} \tag{2.1}$$

$$U = g(\cdot; s\,\sigma^2, s\,\tau^2) * \begin{bmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{bmatrix} \tag{2.2}$$

An spatio-temporal interest point is detected at locations with high intensity variations in all three dimensions, resulting in three large eigenvalues of U. These eigenvalues are evaluated as in Equ. 2.3, where the parameter k specifies how much the three eigenvalues are allowed to differ from each other. After applying a non-maxima suppression strategy on H locations with values above a certain threshold indicate interest points.

$$H = det\,(U) - k\,\mathrm{trace}^3(U) \tag{2.3}$$

Cuboid detector as presented by Dollar et al. [38] evaluates the response function R as depicted in Equ. 2.4 generated by a 2D Gaussian filter $g(x, y; \sigma^2)$ in the spatial domain and two 1D Gabor filters $h_{ev}(t; \tau, \omega)$ and $h_{od}(t; \tau, \omega)$ on the temporal domain. Final key-point detection is performed via non-maxima suppression on R. Different spatial and temporal scales can be evaluated by varying σ and τ, while ω is usually set to $4/\tau$. The Cuboid detector reacts very sensitive to periodic motions and in general generates more STIP detections compared to Haris3D.

$$\begin{aligned} R &= (I * g * h_{ev})^2 + (I * g * h_{od})^2 \\ h_{ev}(t; \tau, \omega) &= -\cos(2\pi t\omega)e^{-t^2/\tau^2} \\ h_{od}(t; \tau, \omega) &= -\sin(2\pi t\omega)e^{-t^2/\tau^2} \end{aligned} \tag{2.4}$$

V-Fast or Video-FAST detector by Yu et al. [186] extends the idea of the 2D Features from Accelerated Segment Test (FAST) [135] detector to the spatio-temporal domain. In contrast to previous detectors it is not based on convolution, but defines pixel comparison tests called *segment test* applied on each voxel in video volume $I(x, y, t)$. The test uses three Bresenham circles centered at (x_C, y_C, t_C), each having a radius of three voxels and a circumference of 16 voxels, as shown in Figure 2.12.

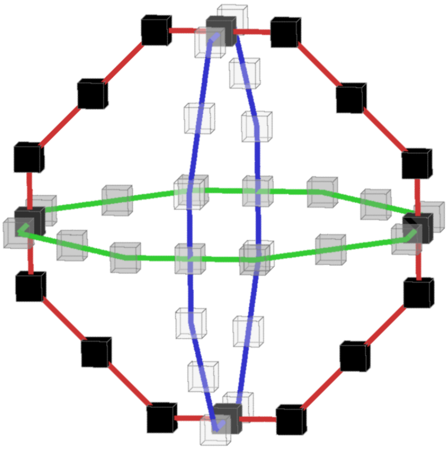


Figure 2.12: Illustration of the XY, XT and YT plane Bresenham circles used for the segment test of V-Fast STIP detector. Image taken from [186].

The center voxel $I(x_C, y_C, t_C)$ is compared with the surrounding voxels on the Bresenham circles $I(x_B, y_B, t_B)$. Following Equ. 2.5, a voxel is salient if n contiguous surrounding voxel are brighter or darker by an factor defined by threshold th. Finally, a non-maxima suppression step is filtering out local detections. V-Fast can process different spatial and temporal scales by changing video resolution and sub-sampling temporal frames.

$$\begin{array}{rclcll} & & I(x_B, y_B, t_B) & \leq & I(x_C, y_C, t_C) - th & (\mathit{darker}) \\ I(x_C, y_C, t_C) - th & < & I(x_B, y_B, t_B) & < & I(x_C, y_C, t_C) + th & (\mathit{similar}) \\ I(x_C, y_C, t_C) + th & \leq & I(x_B, y_B, t_B) & & & (\mathit{brighter}) \end{array} \tag{2.5}$$

2.2.2 Motion Estimation for Activity Recognition

Spatio-temporal interest points highlights moving regions of interest, but do not describe motion in any sense. But motion estimation and description defines a vital part of activities in videos. Given two consecutive image frames I_t and I_{t+1}, the motion field $\omega_t = (u_t, v_t)$ defines the displacement of all pixels from t to $t + 1$. The motion field with its horizontal u_t and vertical v_t components can be either found by optimization or matching. Several approaches have been developed over time and can be roughly divided into sparse and dense motion estimation ones. Dense optical flow methods adds an regularization term

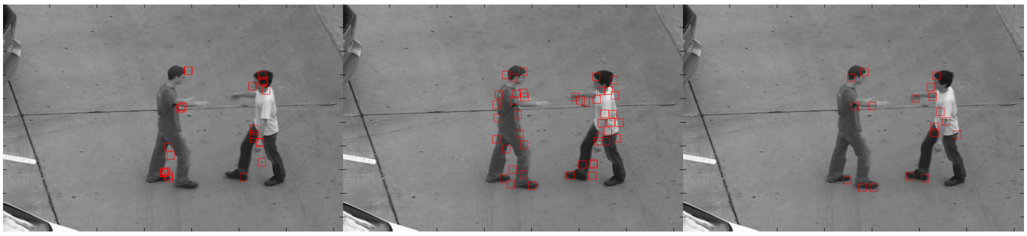


Figure 2.13: Visualization of different STIP detections indicated by red squares (no scaling is visualized) on UT-Interaction video dataset. From left to right: Harris3D, Cuboid and V-FAST interest point detectors. In general Cuboid detector generates more responses, often along edges, while Harris3D and V-Fast focus more on moving corners.

which allows for estimating u and v for pixel without texture or going into occlusion in frame $t+1$, based on neighboring pixels. Example visualizations of optical flow by [45], [63] and [176] are shown in Figure 2.14. With increasing regularization capabilities resulting flow fields get smoother and less noisy. Nevertheless, experimental results by [167] and [75] demonstrated that optical methods superior on specific benchmarks, e.g., Middlebury[13] must not necessarily be the best choice for activity recognition.

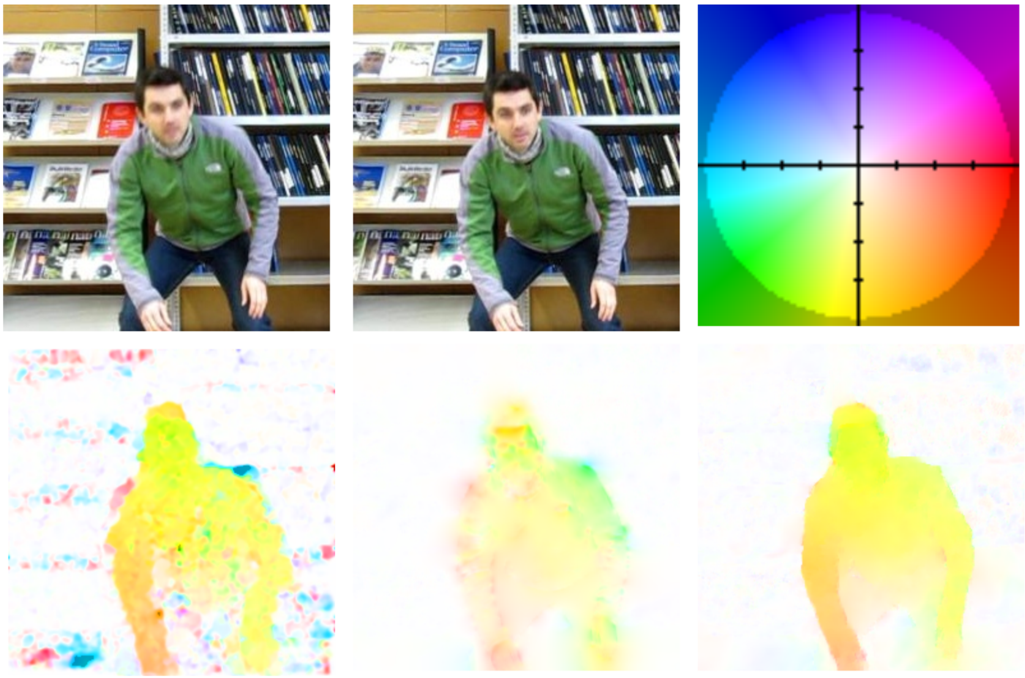
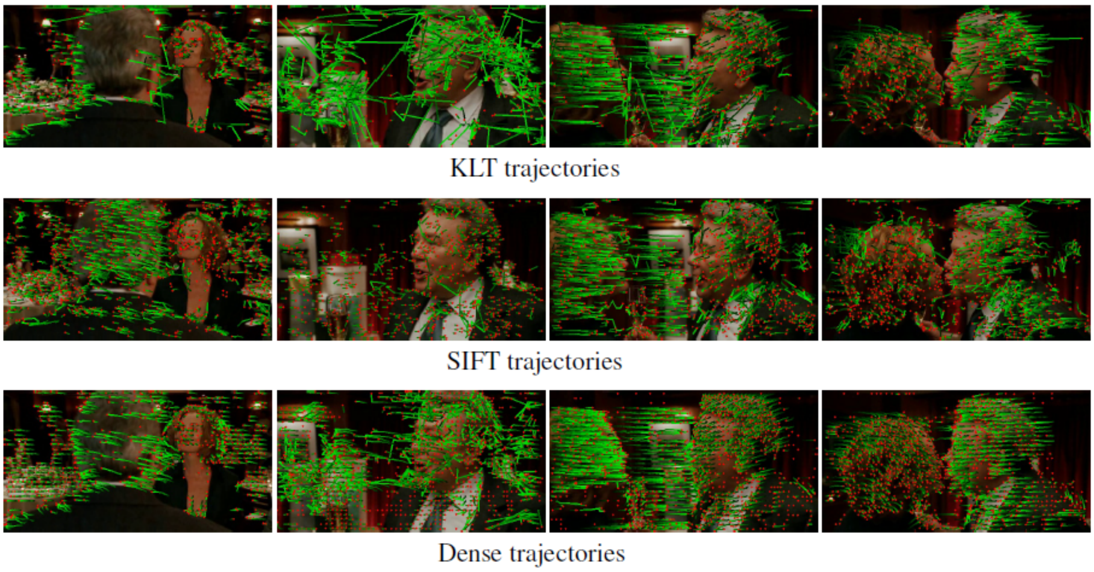


Figure 2.14: Comparing different algorithms for optical flow estimation on two sequential image frames. Bottom row from left to right: Farnebäck [45], Horn and Schunck [63] and Werlberger et al.[176].

In contrast to dense computation by regularization sparse motion estimation is com-

[13] http://vision.middlebury.edu/flow/

puted locally, either in the surrounding neighborhood like in KLT [105] or based on key-point matching between frames as proposed by [158]. Both approaches extend the motion estimation to trajectories over several frames. Therefore, generating more long-term information compared to motion estimation between only two consecutive frames. As a drawback, sparse local motion estimations may be easily confused by sudden motion changes, see Figure 2.15. Furthermore, previous experiments by [168] have shown the superior performance of dense sampled features over spare key-point detections. As consequence, Wang et al. [167] developed the concept of *improved dense trajectories*, a combination of trajectory tracking based on dense optical flow (see Figure 2.15).



KLT trajectories

SIFT trajectories

Dense trajectories

Figure 2.15: Sparse long-term motion estimation by various trajectory generating methods. Image taken from [167].

2.2.3 Feature Description for Appearance and Motion

Given key-point detections or dense sampled positions generated by, e.g., *dense trajectories* we need a discriminative features representation. Within all approaches proposed over the years HOG (histogram of gradients) descriptor [35] for appearance and HOF (histogram of optical flow) [95] as well as MBH (motion boundary histogram) [36] representation for motion have become standard. As they form the common feature basis of the following chapters, we summarize the main concepts.

HOG As first step in the HOG calculation, we have to estimate the gradients $g_x(x, y)$ and $g_y(x, y)$. For each position (x, y) the image $\mathbf{I_t}$ is filtered by 1-dimensional masks $[-1, 0, 1]$ in x and y direction [35]. Then, we calculate the magnitude $m(x, y)$ and the signed orientation $\Theta_S(x, y)$:

$$m(x,y) = \sqrt{g_x(x,y)^2 + g_y(x,y)^2} \tag{2.6}$$

$$\Theta_S(x,y) = \tan^{-1}\left(g_y(x,y)/g_x(x,y)\right) \ . \tag{2.7}$$

To avoid problems due to intensity changes and to make the descriptor more robust, we transform the signed orientation Θ_S into an unsigned orientation:

$$\Theta_U(x,y) = \begin{cases} \Theta_S(x,y) + \pi & \theta_S(x,y) < 0 \\ \Theta_S(x,y) & \text{otherwise} \end{cases} \ . \tag{2.8}$$

To estimate the HOG descriptor, we divide the image $\mathbf{I_t}$ into non-overlapping cells. For each cell, the orientations Θ_U are quantized into 8 bins and weighted by their magnitude m (see Figure 2.7).

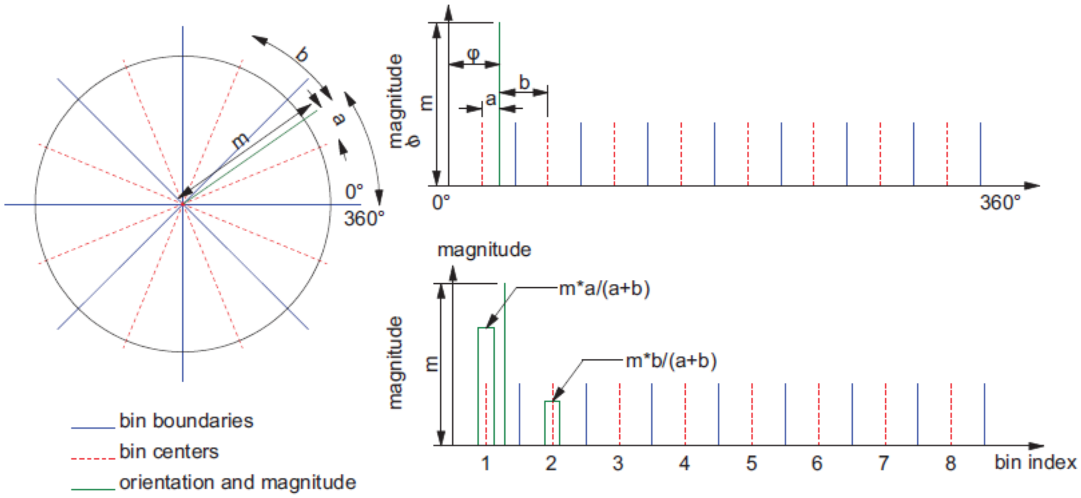


Figure 2.16: Aggregation of oriented magnitude information into neighboring HOG histogram bins. Images taken from [113].

Groups of 2×2 cells are combined in overlapping blocks and the histogram of each cell is normalized using the L2-norm of the block. The final descriptor is built by concatenation of all normalized blocks (see Figure 2.17).

Working in video volumes, the HOG computation can be extended by aggregation over the volume and definition of several spatio-temporal partitions as described by [95]. A further extension to the 3D domain was proposed by Kläser et al. [87].

HOF In contrast to HOG, which is estimated from one frame only, HOF describes the motion between two consecutive frames. Based on a dense optical flow field, as depicted

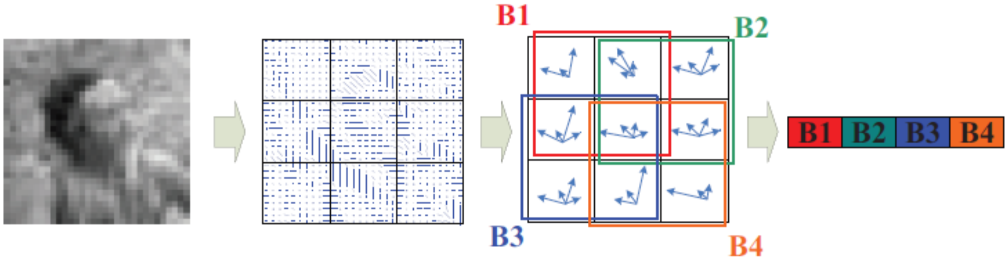


Figure 2.17: Forming HOG feature vectors by concatenation of gradient histograms per block. Images taken from [113].

in Figure 2.14, given $\mathbf{I_t}, \mathbf{I_{t+1}}$, the optical flow describes the shift from frame t to $t+1$ with the disparity ω_t, where $u_t(x,y)$ and $v_t(x,y)$ denote the disparity components in x and y direction at location (x,y). We then compute the HOF descriptor similar as described above by applying Equ. (2.7) and (2.8). However, the gradients $g_x(x,y)$ and $g_y(x,y)$ are replaced by the disparity components $u_t(x,y)$ and $v_tx(x,y)$. Moreover, to capture different motion directions for same poses, we use the signed orientation Θ_S and quantize the orientation into 9 bins (one additional for accounting zero motion).

MBH As dense optical flow represents the complete motion between two consecutive frames, the motion origin may be related to background or camera motion too. This can confuse activity recognition approaches or lead to bias towards learning specific camera motions instead of the activities themselves. Dalal et al. [36] proposed the MBH, by computing the derivatives on the horizontal u_t and vertical v_t flow-field separately. This gradient of the optical flow removes constant camera motion while information about borders and disruptions in flow-field is kept. MBH are more robust to camera motion than optical flow, and thus more discriminative for action recognition. Figure 2.18 summarizes the underlying information for HOG, HOF and MBH.

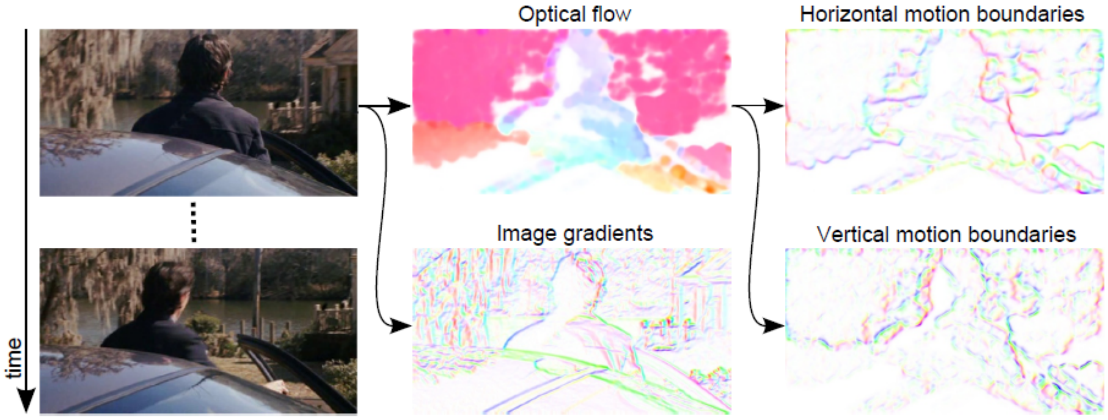


Figure 2.18: Input modalities for different features. Give two consecutive images (left), we generate dense optical flow for HOF and gradients for HOG (middle), while derivatives of the optical flow form the basis of horizontal and vertical MBH (right). Images taken from [167].

3 Temporal Information for Improved Activity Classification and Detection

Activities are, in general, defined by the chronology and temporal relation between specific events or characteristic motion. Often, such information is either not represented in the description and classification of activities, or not exploited during training of related classifiers and detectors. In this chapter, we aim at integrating temporal and spatial information and relations as the principal idea for improving classification results, and allowing unsupervised retraining of detectors for adaption to new backgrounds and surroundings. Especially the absence of negative or background training data can be a limiting factor. With the proposed exploitation of temporal and spatial relations within activities, training of detectors from small training sets and adaption to new scenarios can be supported. Furthermore, improved classification performance can be achieved by local temporal weighting determined by temporal relations between various feature cues. Two different approaches for discrete and continuous representation of temporal relations are presented, namely prototype based hierarchical structures and augmented canonical component analysis. Both showing improved performance compared to related approaches and reveal new application areas for adaptive activity recognition systems.

Contents

3.1 Introduction and Related Work

Human activity classification combines several fundamental vision problems such as detection, tracking and recognition. In contrast to global activity classification on video level, this typically requires a linking of consecutive action detections or, during an online procedure, the tracking of activity detections over time. Within this chapter, two related problems are tackled. First, the need for learning temporal displacement relationships for supporting robust activity detection training. And second, the incorporation of temporal information into classification, assuming that individual activity detections are already linked.

3.1.1 Spatial Temporal Relations for Training

Accurate and correct detections are often considered as given or combined with off-the-shelf-trackers, without integrating knowledge about activity behavior over time. With the growing interest in human activity recognition in computer vision, a variety of approaches have been proposed introducing new features, representations, or classification methods.

For the problem of activity classification in videos, global representations are a common choice [90, 168]. Different spatio-temporal interest points (STIP) and descriptors were evaluated in [168]; the relationship between spatial and temporal STIPs was additionally used in [90]. This applies for early datasets with homogeneous backgrounds and single moving objects like in [57] as well as for recently collected datasets like in[133], showing a larger variety of the activities. The problem even gets harder if additionally to classification also detection is considered. This is of relevance if, e.g., several simultaneous activities, background motion or high inner class variability of backgrounds have to be considered. Previous methods either assumed given tracks from background subtraction or general tracking approaches like [22, 102] just analyzed activities per track or per frames detections as [85, 182]. For linking the latter one in an online manner applying particle filter is a favored choice, see [102, 182]. Recently an offline linking method via graph-based representation has been proposed by [85]. But in general, all these approaches do not take into account the prior knowledge about the correlation of motion and specific activities.

This chapter presents a novel strategy for learning activity specific motion models by feature-to-temporal-displacement relationships. A new method based on an augmented version of canonical correlation analysis (AuCCA) is proposed for linking high-dimensional features to activity-specific spatial displacements over time. This continuous and discriminative approach is compared to other well established methods in the field of activity recognition and detection, namely k-means hierarchies [117] and Hough forests [53]. In particular,a improved activity detection by incorporating temporal forward and backward mappings for regularization of detections is presented. Second, a particle filter framework is extended by using activity-specific motion proposals, allowing for reducing the search space drastically. To show these improvements, detailed evaluations on several benchmark datasets are performed, clearly showing the advantages of the activity-specific motion models.

The main contributions concerning activity motion estimation of this chapter are:

- an efficient classification and voting framework via AuCCA,
- correction of activity detections via consistency checks and
- particle filter tracking with activity specific motion proposals.

3.1.2 Prototype based Activity Recognition

Besides the correlation of extracted features to motion behavior, another important source of information is the relation of features over time. Prototype-based classification methods became popular for activity recognition, as temporal relation information is coded easily within prototype sequences. However, such methods, even showing competitive classification results, are often limited due to too simple and thus insufficient representations and require a long-term analysis. To compensate these problems, in the following, we present and an efficient prototype-based representation based on sophisticated features, allowing for a single-frame activity recognition as well as sequential information exploitation.

Thus, a variety of approaches have been proposed introducing new features, representations or classification methods. Since actions can be described as chronological sequences, special attention has been paid on how to incorporate temporal information. In general, this can be realized either by keeping a very strict spatial-temporal relation on the features (e.g., by spatio-temporal volumes [12, 84] or descriptors [38, 141, 142]) or on the representation level [102, 162]. For such approaches the classification is typically performed on single-frame basis and the analysis of longer sequences is based on a simply majority voting or on averaging over multiple frames. If spatial-temporal consistency is totally ignored [173] only whole sequences can be analyzed. One effective way for directly describing temporal information, that showed great success in the past, is the usage of prototypes, e.g., [41, 102, 162, 173].

In general, prototype-based learning methods can be described by a prototype space $\mathbf{X} = \{x_1, \ldots, x_n\}$, which is defined as a set of representative samples x_j describing the data (prototypes), and a distance function ρ [41]. In particular, for action recognition the data is split into a smaller set of reference templates referred to as prototypes [102], key-poses [173] or pose-primitives [162].

Weiland and Boyer [173] used foreground segmentation to create a set of silhouette exemplars, so called key-poses, using a forward selection process. The final action description is achieved by comparing the occurrence frequency of key-poses in a video. Although they presented excellent results, they completely neglected the temporal ordering of prototypes within a sequence and showed only recognition results on complete videos. Similarly, Elgammal et. al. [41] modeled an action as a sequence of silhouette prototypes using an hidden markov model (HMM) for incorporating temporal constrains and being more robust to small deviations. To incorporate temporal context in a prototype-based representation, Thurau and Hlavac [162] introduced n-grams models. They define subsequences of n frames and describe the transitions between prototypes by n-dimensional histograms. However, the required number of samples to fill the n-dimensional histograms is high and the temporal ordering is very strict. Furthermore, the representation of n-grams is getting difficult if $n > 3$. Experimentally, they showed state-of-the-art results on sequences with lengths of around 30 frames. Since shape information clearly gives only limited information on single-frame basis Linet al. [102] proposed to create prototypes in a joined shape-motion space using binary foreground silhouettes for shape and flow features as introduced in [40]. The prototypes are trained and represented efficiently using hierarchical k-means, leading to real-time evaluation performance. Finally, the temporal information is incorporated using Dynamic Time Warping (DTW). Although DTW is a powerful method for aligning temporal sequences, as a drawback it only compares one sequence to another and cannot handle transitions between different actions. Again only results on sequence level are shown.

Even though showing competitive recognition results, existing prototype-based action recognition methods are limited due to a required long-term analysis (i.e., on a whole sequence) and mainly rely on accurate segmentations – at least for learning the shape prototypes (e.g., [102, 162, 173]). In practice, however, perfect segmentations are often not available and short and fast actions such as in sports analysis should be recognized. Hence, the goal for human action recognition should be to robustly classify on a short sequence length. Schindler and van Gool [141] showed that if a more sophisticated representation is used human action recognition can also be performed from very short sequences (snippets). They use motion and appearance information in parallel, where both are processed in similar pipelines using scale and orientation filters. The thus obtained features are then learned by using a Support Vector Machine. Their approach showed impressive results, reaching state-of-the-art results even though only short sequences of $5-7$ frames are used.

Hence, in addition to the correlation approach already motivated in 3.1.1, a further goal of this chapter is to introduce an efficient action recognition approach working on short-frame level that takes advantage of prototype-based representations such as fast evaluation, multi-class capability and sequential information gain. In particular, several feature cues are processed in parallel (two for appearance and two for motion) and hierarchical k-means tree [117] are trained, where the obtained leaf nodes represent the prototypes. To increase the classification power, a temporal weighting scheme for the co-occurrences of prototypes between different feature cues is learned.

Thus, in contrast to existing methods (e.g., [69, 141]), which typically use global weighting strategies (i.e., the same weights are applied for all data) the reliability weights of individual features are continuously estimated for each specific point in time. Hence, in contrast to existing methods using different cues , the proposed approach does not estimate global weights, which allows us to temporally adapt the importance of the used feature cues. Moreover, even using temporal context we can still run a frame-wise classification! The approach is demonstrated on standard benchmark datasets showing excellent classification results. In particular, a detailed study on the applied features, the hierarchical tree representation, and the influence of temporal weighting as well as a competitive evaluation to comparable state-of-the-art methods are shown.

The reminder of the chapter is organized as follows. First, in Section 3.2 let us introduce the new action recognition approach consisting of an efficient prototype-based representation and a temporal feature weighting scheme (TPFW). Next, within 3.3 we first define the problem of learning combined classification and temporal displacement voting. The proposed AuCCA approach is presented, and the idea of temporal displacement voting is shown to be incorporated into other standard learning frameworks for comparison. The section further depicts the detection and tracking framework and shows how to incorporate the learned temporal motion models. Experimental results are shown in Section 3.4, where competitive results for activity classification using weighted prototypes are presented, and the advantages of learned motion models for unsupervised detection correction and tracking vs. random-walk models or optical flow based motion models are evaluated. Finally, Section 3.5 summarizes the proposed approaches and gives an outlook on how to incorporate AuCCA into other hierarchical concepts for temporal activity detection and linking.

3.2 Temporal Relations for Weighted Classification

In the following, the temporal prototype-based feature weighting approach (TPFW) for action recognition, as illustrated in Fig. 3.1, is defined. To gain different kind of information, we apply four feature cues in parallel, two for appearance (Section 3.2.1) and two for motion (Section 3.2.2). For these cues we independently train hierarchical k-means trees [117], which provide several benefits such as very efficient frame-to-prototype matching and an inherent multi-class classification capability (Section 3.2.3). To incorporate temporal information, we further estimate temporal weights for the different feature cues. In particular, from temporal co-occurrences of prototypes we learn a temporal reliability measure providing an adaptive weight prior for the evaluation step (Section 3.2.4). In this way during evaluation at a specific point in time the most valuable representations get higher temporal weights increasing the overall classification power (Section 3.2.5).

3.2.1 Appearance Features

In contrast to existing prototype-based action recognition methods, which mainly use segmentation results to represent the data, we apply four more sophisticated feature cues in parallel, two describing the appearance and two describing the motion, respectively. In particular, for appearance these are the Histogram of Gradients (HOG) descriptor [35] and the Locally Binary Patterns (LBP) descriptor [120]. HOG estimates a robust local shape

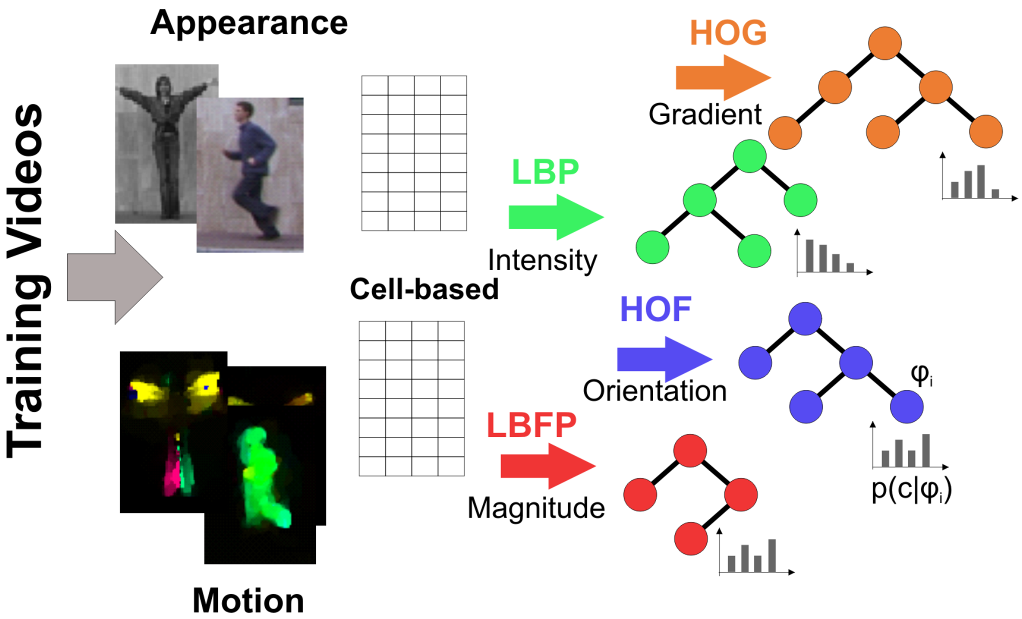


Figure 3.1: Prototype-based action recognition: For each feature cue f a hierarchical k-means tree T^f is estimated, where the leaf nodes of T^f are treated as prototypes φ. In addition, to allow for a prototype-based classification, for each prototype φ the probabilities $p(c|\varphi)$ are estimated, where c is the corresponding action class.

description using a histogram binning over the gradient orientation a local normalization whereas LBPs, originally introduced for texture description, are valuable due to invariance to monotonic gray level changes and robustness to noise. Thus, both have shown to be very valuable for human detection as well as for action recognition. To describe the motion information, we adapted both methods to describe a motion field obtained from an optical flow estimation: Histogram of Flow (HOF) and Locally Binary Flow Patterns (LBFP). In the following, we give the details on these descriptors given the image $\mathbf{I_t} \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$ at time t.

HOG descriptor is applied as described in Section 2.2.3. We apply the standard configuration of 8 orientation bins, using unsigned gradient orientation. Groups of 2×2 cells are combined in overlapping blocks and the histogram of each cell is normalized using the L2-norm of the block. The final descriptor is built by concatenation of all normalized blocks. For speed issues we avoid the tri-linear interpolation, but instead distribute weights over neighboring bins as depicted in Figure 2.7.

LBP An LBP pattern p is constructed by binarization of intensity differences between a center pixel and a number of n sampling points with radius r. The pattern p is assigned 1 if the intensity of a sampling point has a higher intensity than the center pixel and 0 otherwise. The final pattern is formed by the $0 - -1$ transitions of the sampling points in a given rotation order. To avoid ambiguities due to rotation and noise we restrict the number of allowed $0 - -1$ transitions to a maximum u, hence, defining uniform patterns $LBP^{u}_{n,r}$. For our final description we build $LBP^{4}_{8,1}$ pattern histograms for each cell and sum up the nonuniform patterns to one bin (see [120] for more details). To finally estimate the LBP descriptors, similar to [171], we keep the cell-based splitting of the HOGs and extract pattern histograms as described before for each cell.

3.2.2 Motion Features

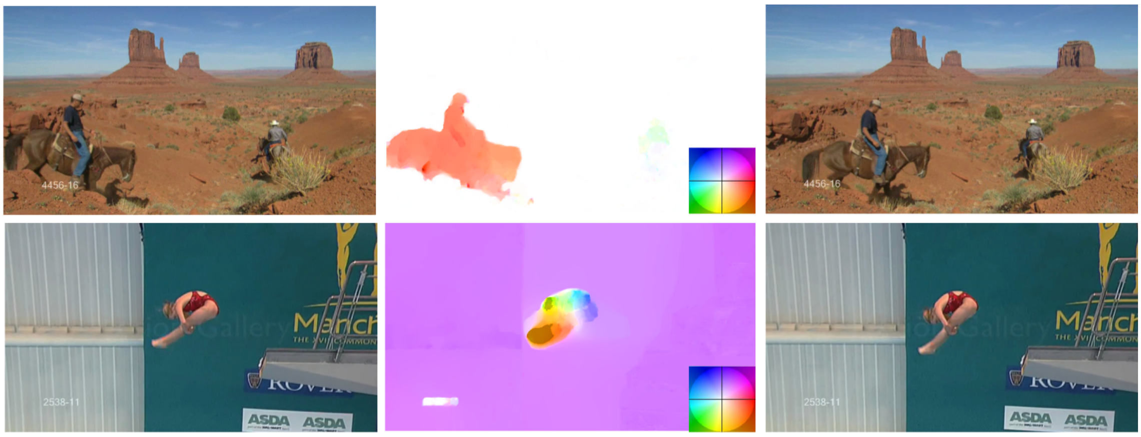


Figure 3.2: Optical flow visualization with related consecutive video frames from UCF-sports dataset[133].

HOF As described in Section 2.2.3 we compute the HOF descriptor similar to the HOG computation. Moreover, to capture different motion directions for same poses, we use the signed orientation Θ_S and quantize the orientation into 9 bins. The other parameters such as cell/block combination are the same as described above for HOG.

LBFP Motivated by the relation between HOG and HOF we directly apply LBPs on optical flow as well. Integrating direction information into LBP descriptors is quite difficult since due to noise in the dense optical flow field the orientation information is sometimes misleading. However, LBPs are known to be robust against such clutter, also appearing in texture, and therefore they are a considerable choice for an additional complementary flow feature. In particular, we keep the cell structure of the appearance LBP and compute the LBFP histograms on the flow magnitude $m(x, y)$, which is computed using Eq. (2.6) from $d_y(x, y)$, $d_x(x, y)$. Although the description is slightly simpler compared to HOF, it is more robust in presence of noise. In general, the same parametrization as for LBP is used. Please note that LBFP are not related to Local Trinity Patterns [184], which are computed on video volumes.

3.2.3 Learning a Prototype-based Representation

Two main issues have to be considered with respect to a prototype-based representation after having the feature descriptions discussed in Section 3.2.1 and 3.2.2. First, how to select a representative set of prototypes. Second, if the number of prototypes is increasing, simple nearest neighbor matching gets infeasible and a more efficient method is required. In particular, we solve both problems by applying a hierarchical k-means clustering, which is also known as Vocabulary Tree [117].

Given the training set S, we first perform a k-means clustering on all training samples s. According to the cluster indices, the data S is then split into subsets (branches of the vocabulary tree), and each subset is clustered again using k-means clustering. This process is repeated recursively until no samples are left in a branch of the tree or if the maximum depth is reached. The thus obtained leaf nodes of the tree are then treated as prototypes φ. Hence, only two parameters are required: the split number k and a maximum hierarchy depth L, allowing to generate a maximum number of k^L prototypes. During evaluation, a test sample is matched to a prototype by traversing down the tree, using depth-first-search, until it reaches a leaf node.

As illustrated in Fig. 3.1, we independently build a vocabulary tree T^f for each feature cue f. Thus, for each cue f we obtain prototypes φ_j^f, i.e., the leaf nodes of tree T^f, from which we can build the prototype sets $\Phi^f = \left\{\varphi_1^f, \ldots, \varphi_N^f\right\}$. To enable a multi-class classification (i.e., one class per action), we have to estimate the class probability distribution $p(c|\varphi)$ for all prototypes φ. Let $S_c \subset S$ be the set of all samples belonging to class c and $S_{\varphi,c} \subset S_c$ be the set of all samples belonging to class c matching the prototype φ. Then the probability that a sample s matching the prototype φ belongs to class c can be estimated as $p(c|\varphi) = \frac{|S_{\varphi,c}|}{|S_c|}$. If no samples from class c reached the prototype φ, i.e., $|S_{\varphi,c}| = 0$, the probability is set to $p(c|\varphi) = 0$.

Illustrative thus obtained classification results are shown in Fig. 3.3. The first row gives a color-coded view of different actions, whereas in the second row (a) the corresponding prototypes and (b) the correct classifications are visualized. It can be seen that for correct classifications over time different prototypes are matched, leading to representative prototype sequences. This clearly shows that the variety in the data can be handled well by using our prototype-based description.

3.2.4 Learning Temporal Weights

However, from Fig. 3.3(b) it also can be recognized that the classification results for the single cues are very weak. Hence, the goal would be to combine these results to improve the classification results. The naive approach to fuse the results from different information cues would be to use majority voting or to estimate a mean over all decisions. Such approaches, however, totally neglect the information provided by temporal constraints. Thus, in the following we introduce a more sophisticated information fusion strategy based on temporal

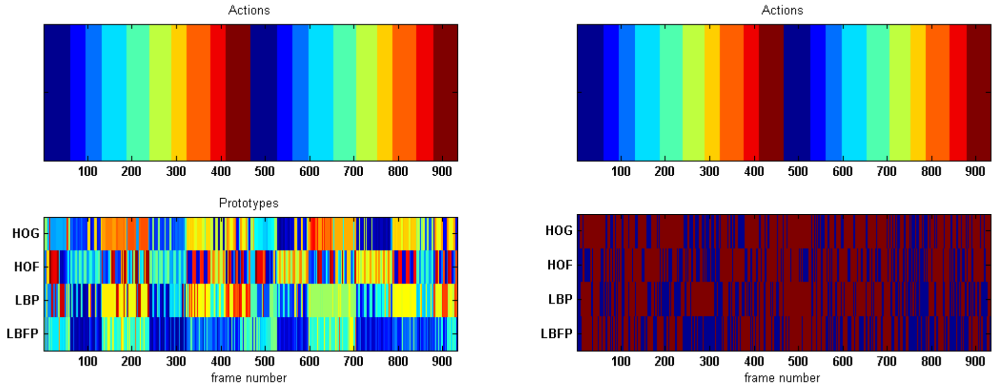


Figure 3.3: Single cue prototype-based classification: (a) sequences (color-coded prototype numbers) of matched prototypes for each feature cue and (b) classification results, where red indicates a correct classification (second row). The actions (first row) are color-coded in the range $1-10$, respectively.

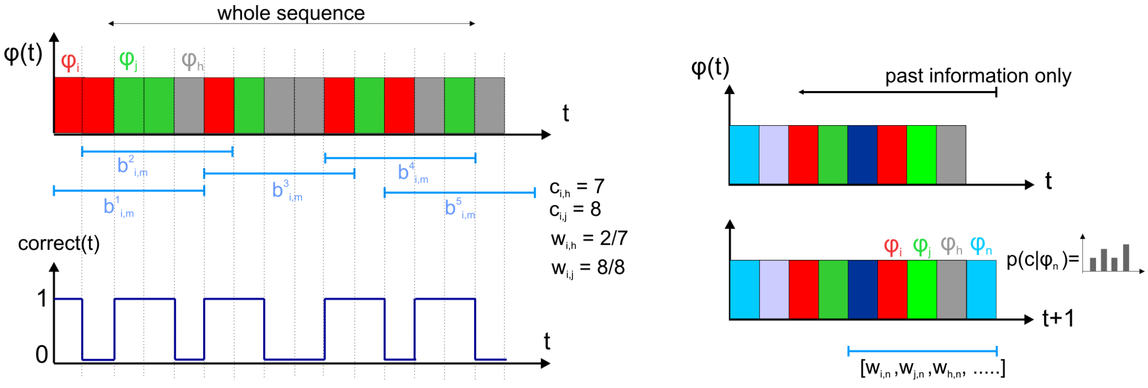
weighting. The main idea is to exploit the knowledge which cues provided reliable results during training to assign temporal adaptive weights to the feature cues during evaluation.

Given the prototype sets Φ^f the key idea is to estimate the reliability of a feature cue for the prototype transitions $\varphi_i^f \rightarrow \varphi_j^f$. This is similar to prototype frequencies or transition statistics as used in [162, 173], which, however, require long sequences to get sufficient data to estimate the discriminative votes. Instead, we consider these transitions only in a short time frame introducing temporal bags, which is illustrated in Fig. 3.4(a).

A temporal bag[1] $b_{i,m}^t$ is defined as set of m prototypes φ_j, which followed the prototype φ_i at time t: $b_{i,m}^t = \left\{\varphi^{t+1}, \ldots, \varphi^{t+m}\right\}$. Once all bags $b_{i,m}^t$ were estimated (i.e., for each occurrence of φ_i) these are combined to a global bag $B_i = \left\{b_{i,m}^1, \ldots, b_{i,m}^T\right\}$, where T is the number of temporal bags $b_{i,m}^t$. Then from B_i we can estimate the temporal co-occurrences of φ_i and φ_j. In particular, we calculate a co-occurrence matrix $\mathbf{C}$, where $c_{i,j}$ integrates all cases within B_i where a prototype φ_i was followed by φ_j: $c_{i,j} = \sum_{t=1}^{T} |\varphi_j \in b_{i,m}^t|$. Having estimated the co-occurrence matrix $\mathbf{C}$, we now can compute a temporal reliability measure $w_{i,j}$. Let $n_{i,j}$ be the number of samples that were classified correctly by prototype $\varphi_j \in B_i$, then we set the reliability weight to $w_{i,j} = \frac{n_{i,j}}{c_{i,j}}$.

This is illustrated in Fig. 3.4(a). The bag B_i contains 7 instances of φ_h and 8 instances of φ_j. While prototype φ_j classified all 8 frames correctly, φ_h provided the correct class for only two samples. Thus, yielding reliability weights of $w_{i,h} = 2/7$ and $w_{i,j} = 1$. If this procedure is repeated for all prototypes in all feature cues this finally yields to four co-occurrence matrices $\mathbf{C}^f$ and four reliability matrices $\mathbf{W}^f$, which can then be used during the test stage as illustrated in Fig. 3.4(b).

[1] Since these calculations are performed for each cue f, in the following for reasons of readability we skip the superfix f in the notation.



(a) (b)

Figure 3.4: Temporal weighting for feature cues: (a) during training the weights $w_{i,j}$ for temporal co-occurrences of prototypes φ_i and φ_j of a feature cue are estimated; (b) during evaluation these weights are used to temporally change the importance of that features cue.

3.2.5 Recognition Using Temporal Weights

Once we have estimated the hierarchical trees T^f and the prototype reliability matrices $\mathbf{W}^f$ as introduced in Sections 3.2.3 and 3.2.4, we can perform action recognition using the following classification problem:

$$p(c|t) = \sum_{f=1}^{4} w_t^f p\left(c|\varphi_t^f\right) \ , \tag{3.1}$$

where w_t^f is the weight of the feature cue f and φ_t^f the identified prototype for cue f at time t. The crucial step now is to estimate the weights w_t^f, which is illustrated in Fig. 3.4(b).

For that purpose, we use the information given by the past, i.e., the identified prototypes per cue, to estimate temporal weights. In particular, considering a temporal bag of size m we estimate the prototype transitions $\varphi_i^f \to \varphi_j^f$, where $i = t-m, \ldots, t-1$ and $j = t$. Based on these selections using the reliability matrices $\mathbf{W}^f$ we can estimate the m corresponding weights $w_{i,j}$. Finally, the weight w_t^t is estimated by averaging the m weights $w_{i,j}$ over the temporal bag.

This recognition process is demonstrated in Fig. 3.5, where the first row illustrates three actions, the second row the identified prototypes, and the last row the corresponding weights. It clearly can be seen that the same action is characterized by different prototypes and also that the weights are changing over time.

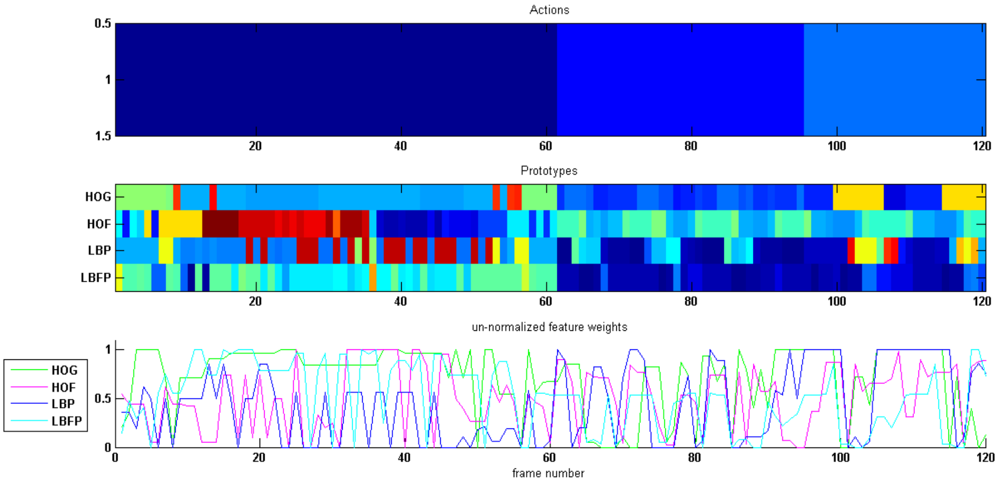


Figure 3.5: On-line adapted feature weights obtained from our temporal reliability measure: color-coded actions (first row), matched prototypes of each feature cue (second row), and estimated weights (third row).

3.3 Temporal Relations for Robust Detection

The temporal prototype-based feature weighting approach (TPFW) described so far assumes detection which are linked over time. Learning this spatio-temporal linking between consecutive points of an activity is the focus of this section. In the following we determine three approaches that can be adopted for that purpose. We start with our proposed AuCCA to exploit the correlation between the feature vectors and the displacement vectors, and additionally extend k-means already explained in 3.2 and RF to learn displacement values for comparison. Canonical Correlation Analysis (CCA) additionally allows for maximizing the correlation between input feature vectors, class labels and displacement vectors. Hierarchical K-means [117] is an unsupervised generative method not exploiting class labels during training; however, class labels and displacement values can be stored according to the obtained clusters. In contrast, Random Forest (RF) use a discriminative splitting function and can also be extended to train according to displacement values [53].

First of all, we describe the general idea of learning motion models via temporal displacements. To allow for combined activity detection/classification we describe an action by a d-dimensional feature vector $x_i \in \mathbb{R}^d$ in a temporal context. Thus, in addition to the corresponding class labels $y_i \in \{1, 2, ..., c\}$ also the temporal offsets for the activity center within in $t + \tau$ frames are given: $\mathbf{d}_i{}^\tau \in \mathbb{R}^2$. Thus, during training we have a prior knowledge on the object's movement within τ frames.

3.3.1 Augmented Canonical Correlation Analysis

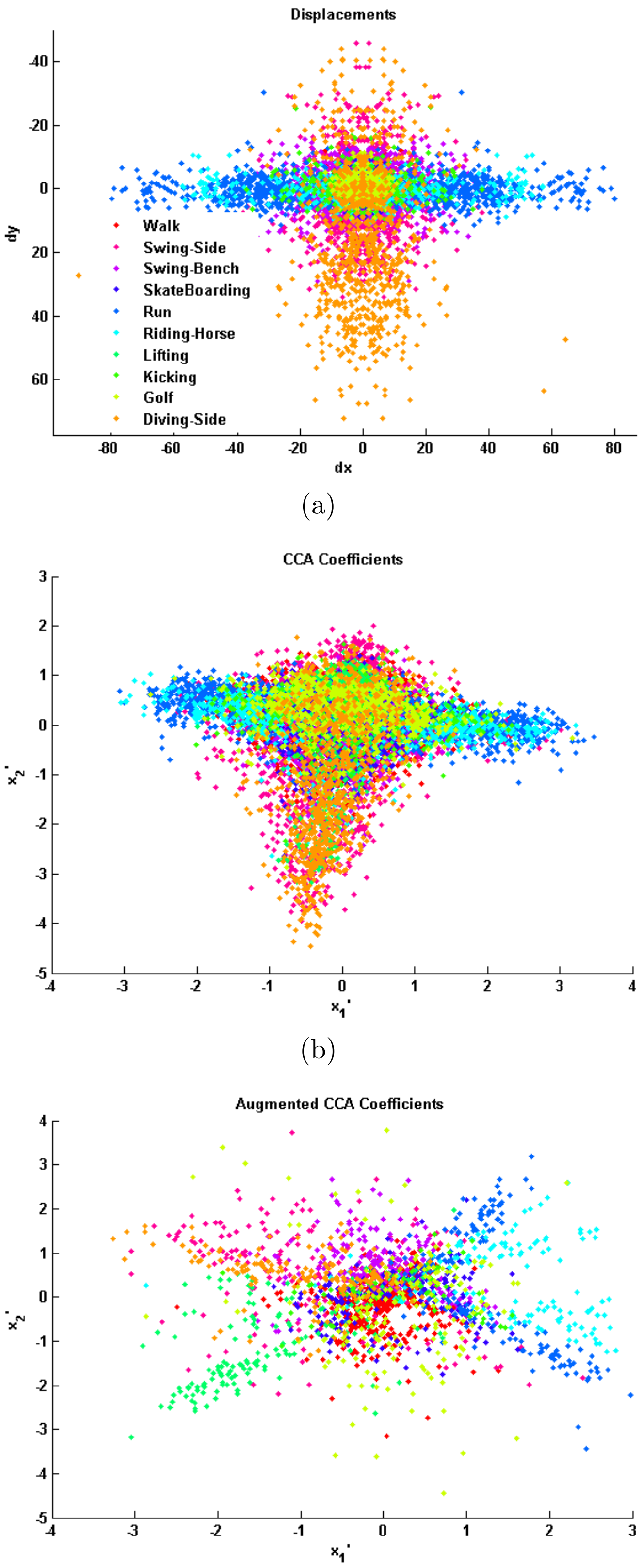
In general, the goal of CCA is to find pairs of directions that maximize the correlation between two random variables [65]. Formally, given two mean normalized random variables $\mathbf{x}$ and $\mathbf{d}$, CCA is defined as the problem of finding a set of two basis vectors $\mathbf{w}_x$ and $\mathbf{w}_d$ such that the correlation between the projections $x' = \mathbf{W}_x^\top \mathbf{x}$ and $d' = \mathbf{W}_d^\top \mathbf{d}$ is maximized. These are obtained by maximizing the correlation coefficient

$$\rho = \frac{\mathbf{w}_x^\top \mathbf{C}_{xd} \mathbf{w}_d}{\mathbf{w}_x^\top \mathbf{C}_{xx} \mathbf{w}_x \mathbf{w}_d^\top \mathbf{C}_{dd} \mathbf{w}_d} \; , \tag{3.2}$$

where $\mathbf{C}_{xx}$ and $\mathbf{C}_{dd}$ are the within-class covariance matrices and $\mathbf{C}_{xd}$ is the between-class covariance. The projections onto $\mathbf{w}_x$ and $\mathbf{w}_d$, i.e., x' and d' are called *canonical factors*. In our case, we build on a more efficient and numerical more stable formulation building on an SVD decomposition [109].

Assuming that all input features in $\mathbf{X} \in \mathbb{R}^{d \times n}$ and the corresponding displacement values in $\mathbf{D} \in \mathbb{R}^{2 \times n}$ are given, we can project the training data onto their canonical correlation coefficients: $\mathbf{X}' = \mathbf{W}_x^\top \mathbf{X}$ with $\mathbf{X}' \in \mathbb{R}^{2 \times n}$. We estimate a linear mapping $\mathbf{F}$ from the canonical correlation coefficients to the corresponding $\mathbf{D}$ using the least square solution of $\mathbf{F} = \mathbf{D}\mathbf{X}'^\top$. This yields a very efficient representation as only $\mathbf{W}_x \in \mathbb{R}^{2 \times d}$ and $\mathbf{F} \in \mathbb{R}^{2 \times 2}$ have to be stored for later evaluations. During testing we estimate the displacements for a given sample vector $\hat{\mathbf{x}}$ by projection onto the canonical space by $\hat{\mathbf{x}}' = \mathbf{W}_x \hat{x}$ and mapping $\hat{\mathbf{d}} = \mathbf{F}\hat{\mathbf{x}}'$.

As the dimensionality of the canonical correlation coefficients is limited by the smaller dimensionality of the correlated training sets in $\mathbf{X}$ and $\mathbf{D}$ too much discriminative information could be lost due to reduced dimensionality. To overcome this problem, Kernel-CCA could be used, which, however, would be computationally much more expensive. Instead, we exploit the additional information given by a multi-class problem and augment the displacement values in each $\mathbf{d}_i$ with a binary label vector $\mathbf{y}_i$ with $\mathbf{y}_i(j) = 1$ for the correct class and 0 otherwise. The incorporated class information strengthens the correlation between points and the higher dimensional correlation space allows for better representation of the mapping from feature to offset space. Moreover, we get a classifier for free. Fig. 3.6 depicts the distribution of temporal offsets over all activities in the UCF sports dataset. Coefficients of standard CCA shows directly the distribution in $d = 2$ dimensional space before multiplying with $\mathbf{F}$. The $d = (c + 2)$ dimensional coefficients of the AuCCA are visualized using a metric multidimensional down-scaling to a dimensionality of 2 using *Matlab* function *mdscale.m*. Due to the high dimensionality of the coefficient space, spatial offsets cannot been seen directly, but the class specific grouping of AuCCA is demonstrated. Results for displacement vectors can be seen in Figure 3.7.



(a)

(b)

(c)

Figure 3.6: Effects of correlating motion information augmented with class labels. a) Original displacement values for UCF sports dataset, color-coded for individual actions. b) Standard 2-dimensional CCA coefficients solely based on motion. c) Coefficients of AuCCA show nicely the separated classes. See text for more detail.

3.3.2 Hierarchical K-means Clustering for Temporal Voting

The key idea of hierarchical k-means clustering, also known as Vocabulary Tree [117], is to recursively split the training data via k-means clustering. This process is repeated recursively until no samples are left in a branch of the hierarchy or the maximum depth L is reached. For each of the thus obtained cluster centers φ (i.e., leaf nodes) we can then estimate a class probability distribution $p(c|\varphi)$.

To learn actions specific motion models, we additionally store temporal displacement vectors of each training sample in the leaf node φ. Therefore, each leaf contains an set of displacement estimation. During evaluation, a test sample is traversed down the tree, using depth-first-search, until it reaches a leaf node. Hence, for each sample we get the class probabilities and the temporal displacement values stored in the leaf node. For simplicity we average the displacement values and report the mean values for d_x, d_y. Certainly one could think about weighting the offset values by the class probabilities or reporting per class displacements, but this has not shown any significant performance gain in our experiments.

An important factor for good generalization and robustness is the use of k-means ensembles, i.e., we split the training set randomly into T subsets and train T individual k-means hierarchies. Similar to the idea of Random Forests [20, 53], this avoids overfitting on the training data and shows better generalization capabilities. During evaluation, a test sample is traversing down all hierarchies and the results of all reached leaves are averaged.

3.3.3 Random Forests for Temporal Voting

Hierarchical k-means clustering can be adopted for the intended task by additionally storing the displacement vectors in the leaves nodes. However, the discriminative information given by the class labels is ignored. To exploit this information, Random Forests can be used. In general, a forest consists of an ensemble of T binary decision trees, which are constructed recursively starting from the root node. For each node, binary tests are computed to split the data $\mathbf{X}$ into subsets going to the left or the right branch [19].

$$\mathbf{w}^T\mathbf{x}_i + b \begin{cases} \geq 0, & \textit{left branch } \mathbf{X}_l \\ < 0, & \textit{right branch } \mathbf{X}_r \quad ; \end{cases} \tag{3.3}$$

with $\mathbf{w} \in 1 \times d$ being a vector where $n_f < d$ values are randomly chosen within the range of $[-1, 1]$, while remaining entries are zero. The bias value b is obtained randomly as well.

Two prominent ways to build a Random Forest are to maximize the information gain ΔE [19, 20, 53] or in addition to minimize the displacement uncertainty U (as proposed for the Hough Forest [53]).

$$\Delta E = -\sum_{i:[l,r]} \frac{|\mathbf{X}_i|}{|\mathbf{X}|} E(\mathbf{X}) \quad , \qquad U = \sum_{i:[l,r]} \frac{|\mathbf{X}_i|}{|\mathbf{X}|} \sum_{j \in \mathbf{X}_i} (\mathbf{d}_j - \bar{\mathbf{d}}_i) \; , \tag{3.4}$$

where $|.|$ denotes the number of samples, respectively, $E(\mathbf{X})$ is the entropy of set $\mathbf{X}$, and $\bar{\mathbf{d}}_i$ is the mean displacement vector in $\mathbf{X}_i$.

Maximizing the information gain is equivalent to minimizing the class-label uncertainty. In the second case, samples with similar displacements are grouped together. The decision whether the class-label or displacement uncertainty should be minimized is selected randomly during training. In our case the displacements are not given in the spatial domain, but instead depicts the motion within the next frames. For evaluation, a test sample is traversing each tree and results of all reached leaves are averaged. Thus, in contrast to the previous mentioned k-means we have a supervised and discriminative discretization of our feature space.

3.3.4 Detection Correction with Forward-Backward Consistency Regularization

This section delineates how the proposed activity-specific motion models learned with AuCCA may be exploited for task specific detection and tracking. The concept of forward-backward mapping is known from key-point tracking or optical flow estimation. Estimating the forward displacement of an small and local image patch $t \to t+1$, should deliver the same result as tracking backward $t+1 \to t$. Section 3.3 formalized how temporal voting information can be learned for complete object patches, where we are not limited to $t+1$ displacements but can train arbitrary mapping for τ and even $-\tau$frames.

Now we exploit this knowledge to filter false-positive detections by applying AuCCA displacement mapping for positive detections at t making forward mapping of τ frames, run the detector at those proposed positions at $t+\tau$ and map backward positive detections with AuCCA learned for $-\tau$ displacements. To show robustness and generalization capabilities, results from tracking on Weizmann robust data are shown in Fig. 3.7. Green and red points visualize positive detections for t and $t+\tau$ respectively, and colored lines temporal votings in $\tau = 5$ frames. We can see a accurate voting to the objects center in $+\tau$ frames, and how false positive detections are marginalized out by missing backward mappings. We derive a conservative learning framework from this observation, for unsupervised mining of new positive and negative samples. We are regularizing over the spatial and temporal neighborhood of positive detections. Detections with no forward-backward consistent detection in their local neighborhood are defined as new negative samples. Detections with a majority of consistent neighbors are defined as new positives while others are seen as neutral and not used for updating the detector, see Fig. 3.18.

3.3.5 Integrating Intelligent Motion Models

Particle filtering for tracking [9] provides a probabilistic framework, which maintains multiple hypotheses of the current object state and has proven to yield impressively robust tracking results. The probability distribution of the hidden target state $\mathbf{s}_t$ of the tracked object at time step t is estimated using a set of N_P weighted particles $S_t = \{s_t^i, w_t^i\}$ with

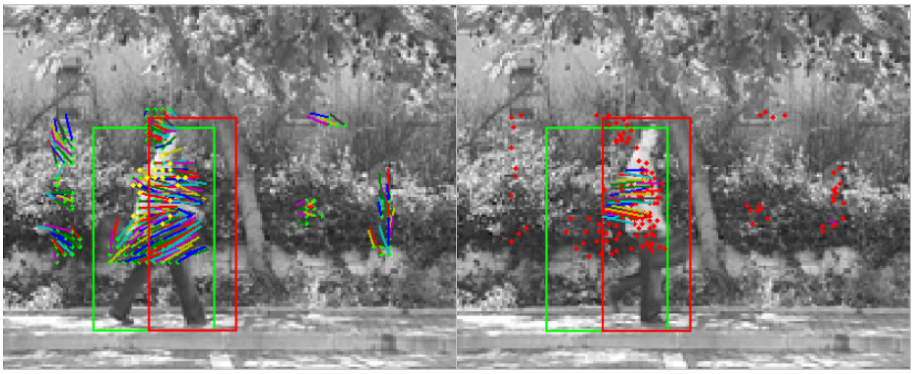


Figure 3.7: Left: Positive detections in frame t with AuCCA forward displacement vectors for $t + \tau$. Right: Detector evaluation of proposed positions at $t + \tau$ and backward displacements for positive detections with consistent backward mapping. Green and red bounding boxes show ground-truth for time t and $t + \tau$, respectively.

$i = 1...N_P$ at time-step t, and associated measurements z_t^i. Each particle x_t^i simulates the real hidden state of the object, using the dynamic model $p(s_t^i|s_{t-1}^i)$ and the observation likelihood $p(z_t^i|s_t^i)$. The object state is approximated by a weighted average over this finite set of particles. To avoid degeneration of the particle set, re-sampling of particles is necessary according to their particle weights $w_t^i \approx p(z_t^i|s_t^i)$, see [9] for more detail.

The most important parameters are the number of particles N_P and the choice of the dynamic model $p(s_t^i|s_{t-1}^i)$. The better the motion model the less particles are needed and a small re-sampling rate would indicate a good fit. In addition, the efficiency is increased, as the runtime is linear according to the number of particles. Classical choices for the dynamic model are a random walk, e.g., used for activity tracking in [102], where motion is modeled by Gaussian noise $s_{t-1}^i + \mathcal{N}(0, \sigma_t)$. Obviously this is very inefficient as no information about current object motion is incorporated, but a general choice if nothing about the objects behavior is known. A slightly advanced version would be a constant velocity model, where in addition to Gaussian noise (u, v) as velocity at $t - 1$ are added to move particles to s_t [182].

None of the previous mentioned approaches incorporates the circumstance, that during tracking of activities prior knowledge about activity specific motion is given. In addition, the particle filter approach is perfectly suited for incorporation an individual motion model per s_{t-1}^i. Given the learning frameworks presented in Section 3.3, we can directly integrate their temporal displacement proposals into the motion model. For each particle state s_{t-1}^i, describing a potential position, we can extract a feature vector $\mathbf{x}_{t-1}^i$. Evaluate this feature with the proposed AuCCA model yields a displacement suggestions $\hat{\mathbf{d}}_i$, and if needed together with estimated class labels $\hat{\mathbf{y}}_i$. The transition to s_t^i is modeled by $\hat{\mathbf{d}}_i + \mathcal{N}(0, \sigma_d)$, where in case $\sigma_d << \sigma_t$ and is mainly needed for diffusing particle positions after re-sampling. Experiments show a evaluation and comparison to optical flow based and random-walk motion models for displacements proposed over several frames.

3.4 Experiments

The following section we give a detailed analysis of the temporal prototype feature weighting (**TPFW**) and the augmented canonical correlation analysis (**AUCCA** and compare to related approaches, considering own previous publications and state-of-the-art results at the time of publications underlying to this chapter.

The selected datasets where trend-setting during the time the experiments where performed, namely the Weizmann and the KTH human action dataset for the TPFW, and additionally the UCF sports dataset for the AUCCA. The capabilities of training and testing between different datasets with variable background structure was tested using the Weizmann robust dataset.

3.4.1 Benchmark datasets

Weizmann human action dataset [12] is a publicly available dataset, that originally contains 81 low resolution videos (180 × 144) of nine subjects performing nine different actions: running, jumping in place, bending, waving with one hand, jumping jack, jumping sideways, jumping forward, walking, and waving with two hands. Subsequently, a tenth action, jumping on one leg (*skip*), was added [57]. Illustrative examples for each of these actions are shown in Figure 3.8.

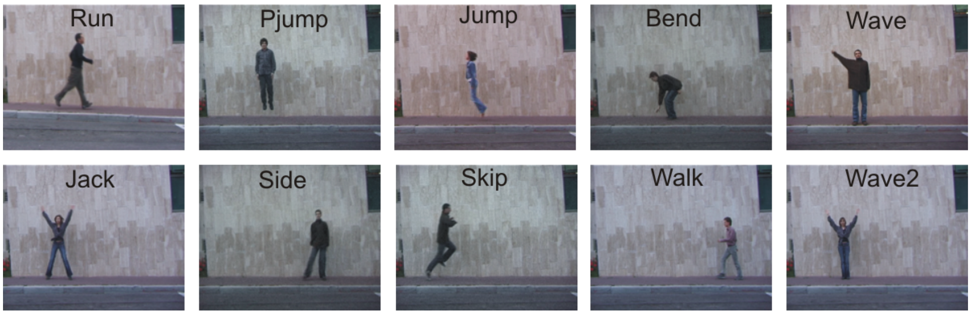


Figure 3.8: Examples from the Weizmann human activity dataset.

Alternatively, the Weizmann-Robustness dataset[2], a less well known dataset with cluttered backgrounds and high irregularities within the performed activities, see Figure 3.9. This dataset will be applied to analyze capabilities of the proposed AUCCA approach for transferring activity detectors to new environments.

KTH human action dataset, originally created by [142], consists of 600 videos (160 × 120), with 25 persons performing six human action in four different scenarios: outdoors (*s1*), outdoors with scale variation (*s2*), outdoors with different clothes (*s3*), and indoors (*s4*). Illustrative examples for each of these actions are shown in Figure 3.10.

[2] http://www.wisdom.weizmann.ac.il/vision/VideoAnalysis/Demos/SpaceTimeActions/DB/robust-deform.zip

Figure 3.9: Examples from the Weizmann ”robust” activity dataset.

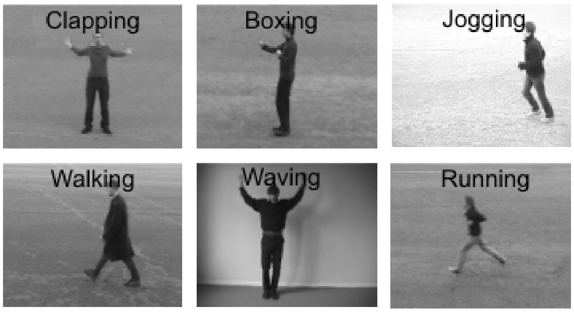


Figure 3.10: Examples from the KTH activity dataset.

UCF sports dataset is a collection of low-quality television broadcasts, containing 150 videos of various sports. This dataset depicts challenging scenarios including camera motion, cluttered backgrounds, and non-rigid object deformations. Furthermore, it provides ground truth bounding box annotations for all activities. Therefore, it is highly applicable to evaluate motion estimation and activity tracking.

Figure 3.11: Examples from the UCF sports activity dataset.

In the first part of the experiments the advantage of a flexible feature weighting over time in comparison to other weighting schemes is proven. Results on short and long-term sequences perform favorable in comparison to state-of-the-art and own previous publications. The second part discusses the possibilities of learning motion vectors, and compares to basic methods for motion propagation like optical flow. As an application example the learned motion models are applied for adopting detectors to new scenes and backgrounds.

3.4.2 Analysis of Prototype-based Learning

In the following, the prototype-based action recognition approach TPFW is evaluated, where several experiments on publicly available action recognition benchmark datasets, *i.e.* Weizmann and KTH are performed. First a detailed analysis of prototype-based action recognition is given and shows that temporal information can be useful to improve the classification results. Next, a comparison against own previous research [136] on learning global weights for features is done, showing the advantage of learning weights per prototypes. Finally, a detailed comparison to published state-of-the-art action recognition approaches is given. In both cases, the given results were obtained by a leave-one-out cross-evaluation [141, 162] (*i.e.* we used all but one individuals for training and evaluated the learned model for the missing one).

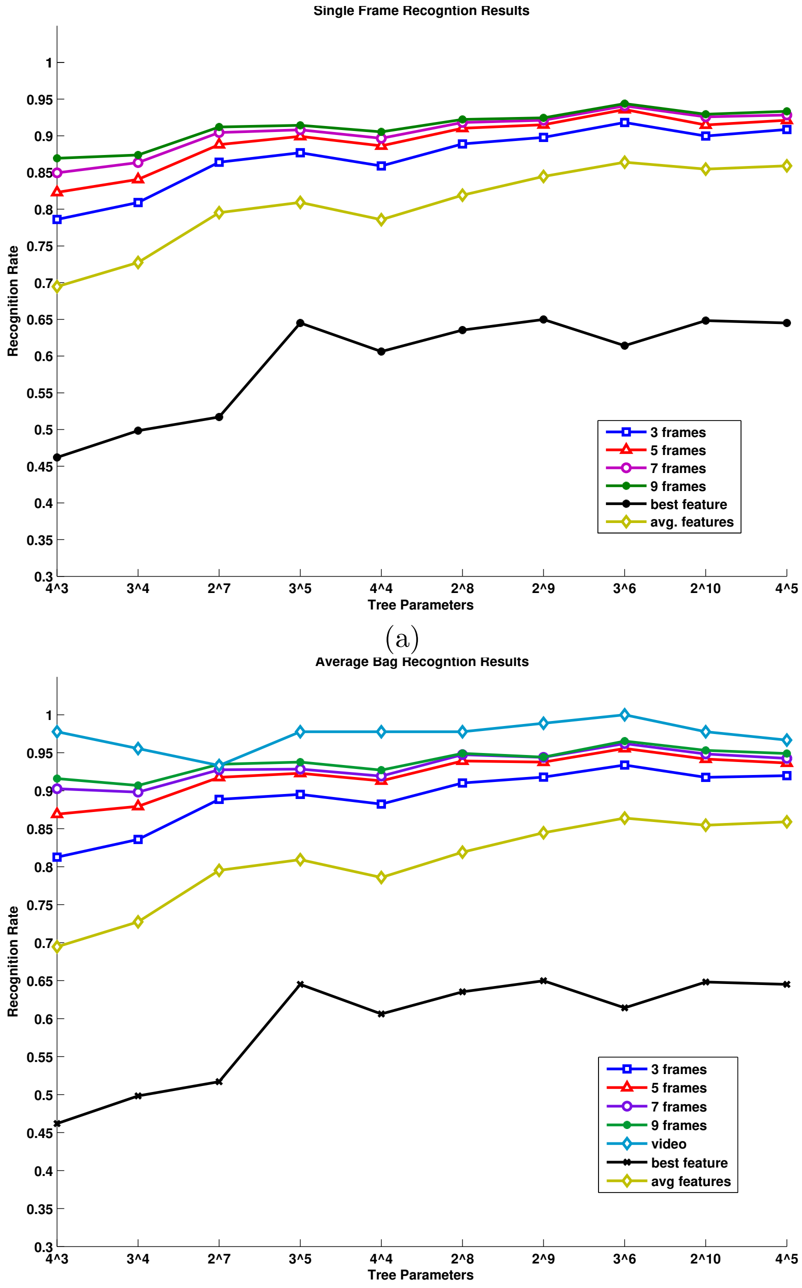
First of all, a detailed analysis of the TPFW action recognition approach is carried out, by analyzing the influence of the parameters to the hierarchical k-means tree and the bag size for the temporal weighting. For that purpose, several experiments varying these parameters on the Weizmann dataset are performed. The corresponding results are given in Fig. 3.12.

Three main trends can be recognized. First, increasing the temporal bag size, which was varied between 3 and 9 increases the classification accuracy. However, using a bag size greater than 5 has only little influence on the classification performance. The bag size defines the temporal memory of prototypes, meaning the influence, importance or discriminative information of successive prototype sequences. If longer sequence lengths are analyzed the ambiguities may be stronger than the additionally gained information. Further, the datasets often contains repetitive activities where individual motion patterns are only a few frames long and no additional discriminative information can be gathered by adding more frames. Nevertheless the results show that TPFW is capable and flexible enough to adopt to different demands. Furthermore, TPFW can describe longer sequences than, e.g., the n-gram approach by [162] and therefore leading to better results, as discussed in section 3.4.4, and still more flexible than global weighting approaches as [136], described in section3.4.3.

Second, increasing the number of prototypes (using different tree parameters, i.e., split criteria and depth) increases the classification power. However, if the number of possible prototypes gets too large, i.e., too many leaf nodes are weakly populated, the classification power is decreased. Experiments show that the optimum number is around 2^8 for the tested datasets. While splitting the training data into more prototypes may be beneficial for getting discriminative hierarchies, it is contradictory for the statistical information of prototype sequences. Fortunately, this effect can be handled during training by using validation data and iteratively grow the hierarchies and train the temporal weights.

Third, it can be seen that using the proposed weighting scheme the single cue classification results as well as a naive combination can clearly be outperformed. In addition, Fig. 3.12(b) shows that averaging the single cue classification results over the temporal bags almost reaches the classification result if the whole sequences are analyzed.

Next, a comparison of different evaluation strategies for TPFW is done in detail for the Weizmann as well as for the KTH dataset. In addition, we show results of the underlying features (on single frame basis) without using the temporal weighting:



(a)

(b)

Figure 3.12: Classification results on the Weizmann dataset with different numbers of prototypes by varying parameters for hierarchical k-means tree and temporal bags: (a) single frame results and (b) bag averaged results.

1. TPFW:

   (a) Single frame evaluation

   (b) Averaging the single frame results over temporal bags

   (c) Analyzing the whole sequence (using a majority voting)

2. Feature combination:

   (a) Analyzing the best single feature cue

   (b) Naive feature combination (majority voting)

Based on the results shown in Fig. 3.12 for the remaining experiments we set the bag size to 5 and used a binary split criterion. The thus obtained results are summarized in Table 3.1 for the Weizmann dataset and in Table 3.2 for the KTH dataset.

	TPFW					
	single frame	bag average	all video	best feature	comb. features	#proto.
2^8	92.4%	94.5%	97.8%	60.1%	84.6%	122
2^{12}	92.4%	94.2%	100.0%	70.2%	89.9%	614

Table 3.1: Overview of recognition results on the Weizmann-10 dataset using 2-means clustering on a maximal depth of 8 and 12, respectively. TPFW is able to perform nearly equally and compensate the weak representation, even with a strong variation of the hierarchies, while the underlying features are improving significantly with larger prototype sets.

From Table 3.1 the benefits of the proposed method clearly visible. In fact, considering a tree-size of 2^8 the best single feature cue provides a classification result of approximative 60%. If the four cues are naively combined the overall classification result can improve to 85%. In contrast, using the proposed temporal weighting (with a bag size of 5 frames) an improvement of the classification rate by 7%, by further averaging over the bag even 9%, can be obtained. If the whole sequence is analyzed, we finally get a correct classification rate of 98%, which can further be improved to 100% if the tree depth is increased.

The same trend can be recognized for the KTH dataset in Table 3.2, where we split the results for the four sub-sets. In particular, there is a significant improvement using the proposed method compared to the best single feature cue and the naive combination. However, as can be seen, the single frame classification result is less improved by averaging over the temporal bag. This can be explained by the difference between KTH and Weizmann datasets, where the former contains solely repetitive activities with high frequency, in which case the temporal context is less informative. But if the whole sequences are analyzed, still a considerable improvement can be recognized.

3.4.3 Evaluation of global feature weighting

During research previous to the results within this chapter, the effect of learning global weights for different features, depending on the activities, has been analyzed and published together with several colleagues in [136]. Following previous action recognition methods,

	TPFW					
	single frame	bag average	all video	best feature	comb. features	#proto.
s1	92.7%	92.7%	97.3%	69.4%	90.1%	113
s2	89.1%	90.6%	94.7%	59.0%	82.0%	118
s3	93.4%	94.5%	98.7%	71.0%	88.0%	116
s4	91.6%	91.7%	98.7%	63.0%	91.9%	109

Table 3.2: Overview of recognition results on KTH dataset using 2-means clustering on a maximal depth of 8. TPFW clearly outperforms the pure feature based results.

which applied different cues such as motion and appearance (e.g., [69, 141]) and also explored the importance of weighting these different cues, we enhanced those ideas and analyzed the possibilities of learning global weights per individual activity. Their results showed that selecting the weights in the range $0.4-0.6$ provide the best results, where only a global weighting was considered. However, considering different weights for different actions might be meaningful. For instance an action like “running” may benefit more from motion compared to an action like “waving with one hand”, where shape is ore beneficial. For that purpose, given specific actions, we applied Multi-Kernel-Learning (MKL) to estimate the optimal weights for the different information cues. Thus, in the following a summary of the outcomes in Roth et al. [136] is given, with a detailed evaluation on the importance and the influence of weighting for action recognition.

3.4.3.1 Multi Kernel Learning

Based on Multiple Kernel Learning (MKL) [93, 129, 153] a method to combine data from multiple information sources, the main idea is to create a weighted linear combination of the kernels obtained from each information source. Moreover, in Rakotomamonjy et al. [129] it was shown that by using multiple kernels instead of one a more effective decision function can be obtained. In particular, given two feature vectors x and x', the kernel $K(\mathbf{x}, \mathbf{x}')$ can be considered a convex combination of M basis kernels $K_j(\mathbf{x}, \mathbf{x}')$:

$$K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{M} d_j K_j(\mathbf{x}, \mathbf{x}'), \tag{3.5}$$

where $d_j \geq 0$ are the weights of the kernels K_j and $\sum_{j=1}^{M} d_j = 1$. Thus, the decision function $g(x)$ of an SVM with multiple kernels can be represented as

$$\begin{aligned} g(x) &= \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) - b \\ &= \sum_{i=1}^{N} \alpha_i y_i \sum_{j=1}^{M} d_j K_j(\mathbf{x}_i, \mathbf{x}') - b, \end{aligned} \tag{3.6}$$

where $\mathbf{x}_i$ are the training samples and $y_i \in \{-1, +1\}$ are the corresponding class labels. Hence, when training an MKL model the goal is to learn both, the coefficients α_i and the weights d_m, in parallel.

3.4.3.2 Results on Global Feature Weighting with MKL

Within the experiments in [136] we analyzed the importance of each information source for the classification task considering the MKL weights, which were estimated for linear kernels using the MKL method introduced in [129]. Having only two cues, the convex combination of the basis kernels in Eq. (3.5) can be simplified to

$$K(\mathbf{x}, \mathbf{x}') = d_{mot} K_{mot}(\mathbf{x}, \mathbf{x}') + d_{app} K_{app}(\mathbf{x}, \mathbf{x}') \; , \tag{3.7}$$

where the subscripts *mot* and *app* indicate the motion and the appearance components, respectively. The thus obtained results obtained for the standard Weizmann and KTH benchmark datasets described in Section 3.4.1. The advantage of applying non-negative matrix factorization (NMF) has already been shown within an earlier publication [159]. Results in [136] are based on those findings, therefore results for 10 and 100 NMF basis vectors are illustrated in Figure 3.13 and Figure 3.14, respectively. A more detailed description on NMF is given in chapter 4 of this thesis.

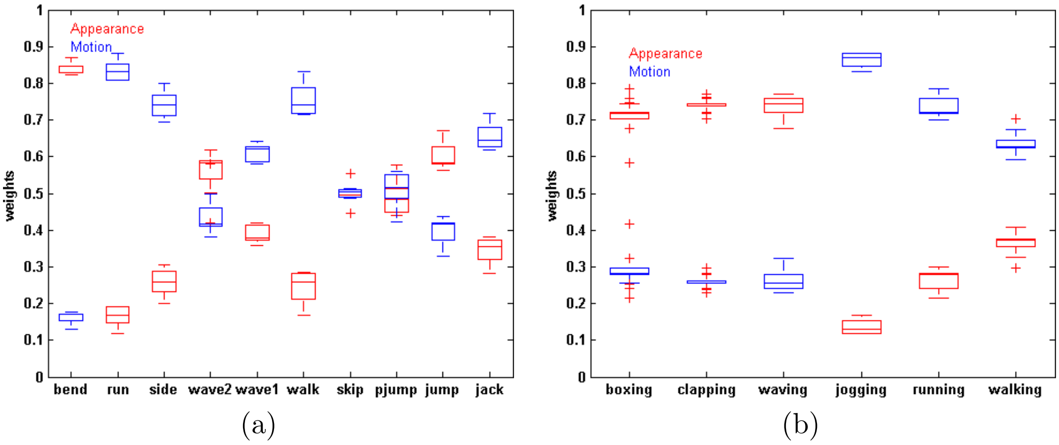


Figure 3.13: MKL weights using linear kernels for the Weizmann (a) and the KTH (b) dataset for an NMF representation of 10 modes. As the representation with 10 basis vectors only is not sufficient, feature weights play an important role. Static activities get more bias towards appearance and shape, while dynamic activities favour the motion features.

These results clearly show that the different cues have different importance for different data; especially, if the representation size is quite small (see Figure 3.13). However, they also show that increasing the representation size (such that sufficient classification results can be obtained) the importance of weights is decreasing (see Figure 3.14). If the representation size is further increased, all weights are reaching approx. 0.5. Here, one

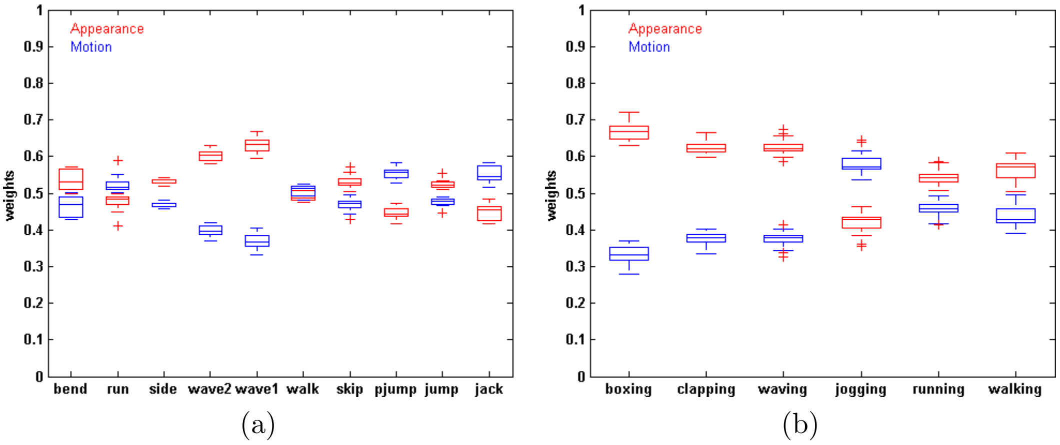


Figure 3.14: MKL weights using linear kernels for the Weizmann (a) and the KTH (b) dataset for an NMF representation of 100 modes. Where there is still a small difference in weighting between static and dynamic activities in both datasets, it is getting less significant with more a discriminative representation using more basis vectors.

should remember the representation size of 600 prototypes for the TPFW experiments. Moreover, for those actions where the classifications "fail" the weights are very similar and can thus not help to increase the classification power! Again, our results within [136] showed that global weighting of features has limitations. In contrast we will see in the following comparison to state of the art the advantage of local temporal weightings of TPFW.

3.4.4 Comparison of TPFW to the State-of-the-Art

Finally, a comparative study of TPFW approach compared to state-of-the-art action recognition methods on the Weizmann and the KTH dataset is performed. Since different authors used different versions of the Weizmann dataset, i.e., 9 vs. 10 actions, we split the Weizmann experiment into two parts. In particular, we compared our approach to Schindler & van Gool [141] and to Thurau & Hlaváč [162], which are most similar to our method – also providing an analysis on short frame basis – and to recent methods reported the highest classification results. The thus obtained results are given in Tables 3.3–3.4. The best classification results when analyzing the whole sequence are set boldface, respectively.

It can be seen from Table 3.3 that TPFW obtains competitive results on short frame basis and also when analyzing the whole sequence. In fact, it can be seen that one obtains comparable results to Schindler & v. Gool and that TPFW clearly can outperforms the approach of Thurau & Hlaváč on short frame basis. Moreover, when analyzing the whole sequence for both data sets we obtain classification results of 100%. Finally, the same experiments are carried out for the KTH dataset, showing the results in Table 3.4. Again, it can be seen that TPFW obtains competitive results on short frame basis as well as

method	rec.-rate	#frames
proposed	94.9% **100.0%**	1/5 all
Roth et al. [136]	97.0%	2
Schindler [141] & v. Gool	93.5% 96.6% 99.6%	1/2 3/3 10/10
Blank et al. [12]	99.6%	all
Jhuang et al. [76]	98.8%	all

(a) Weizmann-09

method	rec.-rate	#frames
TPFW	92.4% 94.2% **100.0%**	1/5 5/5 all
Roth et al. [136]	94.2%	2
Thurau & Hlaváč [162]	70.4% 94.4%	1 30/30
Gorelick et al. [57]	98.0%	all
Lin et al. [102]	**100.0%**	all
Fathi & Mori [47]	**100.0%**	all

(b) Weizmann-10

Table 3.3: Recognition rates and number of frames used for different approaches reported for the *Weizmann* dataset. The best results are shown in bold-face, respectively.

when analyzing the whole sequence, even (significantly) outperforming most state-of-the-art methods an all four data sets. In particular, also on this dataset TPFW outperforms the approach of Schindler & van Gool on short frame basis and yield the best overall performance for the full sequence analysis!

In summary, we see that in comparison to related approaches also exploiting temporal relations or feature weighting, TPFW performs favorable. This supports the original idea of this work that local temporal weighting and relations between individual prototypes are beneficial and that TPFW is capable of exploiting this information to generate superior results.

3.4.5 Learning Spatio-Temporal Relations with AUCCA

The above discussed TPFW approach assumes already annotated training data to create prototype representation and temporal weighting functions. Generating this annotation

	s1	s2	s3	s4	average	# frames
TPFW	92.7%	89.1%	86.1%	91.3%	89.8%	1/5
	92.6%	90.6%	94.5%	91.7%	92.4%	5/5
	97.3%	**94.7%**	**98.7%**	**98.7%**	**97.4%**	all
Roth et al.	88.1%	84.1%	-	88.4%	-	2
Schindler & v. Gool [141]	90.9%	78.1%	88.5%	92.2%	87.4%	1/2
	93.0%	81.1%	92.1%	96.7%	90.2%	7/7
Lin et al. [102]	**98.8%**	94.0%	94.8%	95.5%	95.8%	all (NN)
	97.5%	86.2%	91.1%	90.3%	91.3%	all (proto.)
Yao and Zhu[183]	90.1%	84.5%	86.1%	91.3%	88.0%	all
Jhuang et al. [76]	96.0%	87.2%	91.7%	95.7%	92.7%	all

Table 3.4: Recognition rates and number of required frames for different approaches reported for the *KTH* dataset. The best results are shown in bold-face, respectively.

can be an expensive task and may be repeated if the dataset if enlarged by new activity, repetitions, backgrounds or environments. Further, for many applications the focus lies more on detection, tracking and linking of activities. The proposed AUCCA method presented in Section 3.3 tackles exactly this problem of learning correct spatio-temporal motion predictions. Within the the following experimental section an evaluation on the two main fields of application is done: first, verification of motion prediction accuracy and comparison to alternative approaches and second supporting adoption of activity classification and detection to a new dataset.

We use Weizmann[3] dataset [57] and UCF-Sports[4] [133] for evaluation of the AUCCA framework. Both datasets allow for object centered recognition and contain activities with different or alternating motion directions. Please note that we do not tend towards breaking recognition scores. Especially for the UCF-Sports dataset global representations on video level show superior results [168], while such results may be somehow misleading. The activity defining the label of a video may only occupy a fraction of the video volume. Therefore, global representations are influenced more by the global motion and small background inner class variability of this dataset, than by the activity itself. They may fulfill their task in solving the dataset, but within this work the focus of interest lies on exploiting the objects motion related its current activity, and in generating detection ”tracklets” throughout videos. We compensate for global camera motion to estimate the real objects motion for training and testing. Otherwise the object centered camera motion gives a too strong prior to evaluate for displacement estimation. Note that for tracking our proposed motion estimation could be combined with global motion in the same way is shown in [182].

[3] http://www.wisdom.weizmann.ac.il/ vision/SpaceTimeActions.html
[4] http://server.cs.ucf.edu/ vision/data.html

3.4.6 Complexity Analysis

We start with a general evaluation of the compared motion estimation models, namely the AuCCA, Random Forest(RF) and the k-means hierarchies (Kmeans). We chose classical features for describing the activities by using HoG and HoF features for Weizmann, and in addition bag-of-words histograms describing the spatio-temporal interest points in a local surrounding of the objects for UCF-Sports. To emphasize the efficient representation using AuCCA, we compare training and evaluation time of a cross validation run on UCF-sports. On average training AuCCA takes $33.3sec$ respectively. In comparison on the same PC and the same training set RF, with 5 trees of maximum depth 10, trains $1400sec$ and Kmeans, 5 hierarchies with $k = 4$ and maximum depth 5, even trains $4044sec$. Evaluation takes $0.23sec$ for AuCCA, $9sec$ for RF and $20sec$ for Kmeans. Concerning efficiency, AUCCA is a light-weight and efficient approach, fast in training and evaluation. In contrast to the evaluated competitors no complex structure has to stored and loaded for evaluation.

3.4.7 Comparison of Motion Estimation and Particle Behavior

One motivation for AUCCA was to estimate spatio-temporal displacements from high-dimensional feature description within training data, and to show that this is superior compared to random sampling or optical flow driven motion models. First, the capabilities of the individual methods proposed in Section 3.3 are analyzed and compared to optical flow based displacement estimations. Average results for Weizmann and UCF are shown in Figure 3.15 and Figure 3.17.

Different displacements were trained, from $\tau = 1, ..., 5$ frames, and compared to optical flow with a constant velocity assumption, meaning that current flow at time t is propagated linear by multiplying with τ. We group together results for static activities like *wave* and *bend* for Weizmann and *golf* and *lifting*, and show individual results for the dynamic activities. When evaluating the prediction for different temporal ranges; e.g., $\tau = 1 - 5$ as in Figure 3.16, one can clearly see that optical flow is not a good choice for predicting object motions from current activity detections over several frames. The results of all evaluated methods are in the same range with minor deviations only, but in relation to the complexity and runtime of the methods evaluated before, AUCCA is the most efficient one. The spatial resolution of all compared methods is limited with the cell-size of the underlying descriptors. Therefore, this cell-size factor is visualized for comparison in Figure 3.15 and Figure 3.16.

Evaluating the applicability of the activity linking particle filter approach discussed in Section 3.3.5, we compared our activity specific displacements learned with AuCCA and a random walk model. We made two tracking runs on UCF sports, once with random walk motion model for a particle filter (see Section 3.3.5), and the other with AuCCA motion propositions per particle. To allow for fair comparison, the underlying detector was the same for both tracking runs. This guarantees that particle weights are not influenced by a better representation, solely using the displacement values of the AuCCA and not the

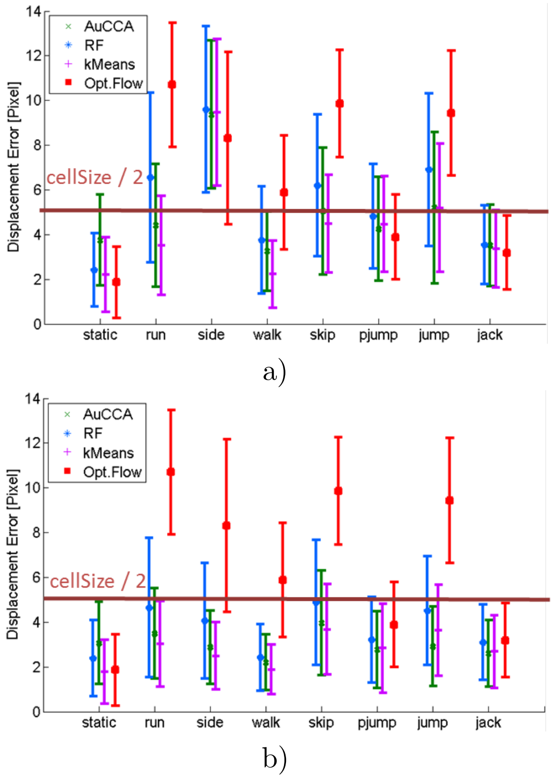


Figure 3.15: Results of object motion estimation for $\tau = 3$ frames on the Weizmann dataset for different methods (AUCCA, RF, kMeans): a) HOG features only, therefore solely appearance information is used to estimate motion. b) Additionally HoF features increase the estimation accuracy especially for non-smooth motions like ”skip”, ”side” and ”jump”. Propagation of optical-flow estimation is not effective, except static and solely vertical motions like ”pjump” and ”jack”.

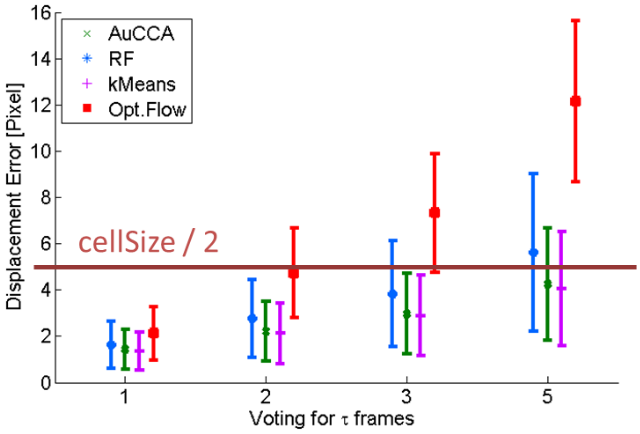


Figure 3.16: Results of object motion estimation for different τ. For temporal displacements longer than 2 frame, optical flow is not applicable anymore.

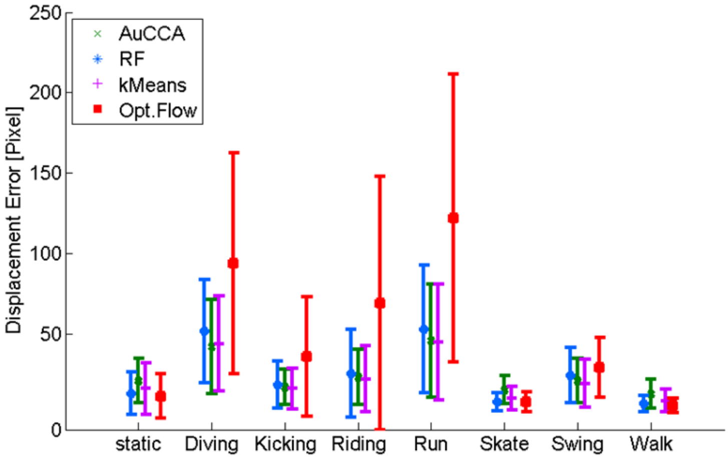


Figure 3.17: Results of object motion estimation for $\tau = 3$ frames on the UCF sports dataset. We see a wider distribution of optical-flow predictions in contrast to the learned predictions.

classification results nor the coefficient values. We made two tests with random walk (RW), first with set of $N_P = 100$ particles and $\sigma_t = 50$ and second with $N_P = 100$ $\sigma_t = 200$. The AuCCA particle filter used only $N_P = 30$ and $\sigma_t = 25$ to noise resampled particles. The average errors to target position for testing on *diving*, *riding*, *run* and *swing-side* where $15.86pix$ for the AuCCA, $12.64pix$ for RW100 and $7.83pix$ for RW200. This shows that our proposed voting scheme is accurate enough and allows for significant smaller particle sets. Nevertheless, the performance is sufficient but not superior compared to the more simple random-walk particle filter approach. The outcome of our investigation was that the videos within the UCF dataset are mostly centered around the object, and therefore the main tracking is already done by the camera man.

3.4.8 Unsupervised Adaption via Forward-Backward Regularization

Considering the results in Section 3.4.6 and Section 3.4.7, AUCCA is capable of performing the motion prediction task with the same accuracy as more heavy competitors within less time and lower resources. Besides the above discussed issue of motion prediction for tracking, another important application is the support of training activity detectors or adopt them to new dataset or background modalities. We tested our forward-backward consistency regularization, motivated in Section 3.3.4, within the task of unsupervised scene adaption of an activity detector. We trained a *walking* detector (SVM with linear kernel) using solely positive and negative samples from the Weizmann dataset, and in addition the motion prediction as depicted in the above sections. In contrast to tracking, two predictions were trained independently for temporal forward and backward prediction.

For testing the adaption capabilities, the Weizmann-Robustness dataset[5], a less well

[5] http://www.wisdom.weizmann.ac.il/vision/VideoAnalysis/Demos/SpaceTimeActions/DB/robust-deform.zip

known dataset with cluttered backgrounds and high irregularities within actions, see Figure 3.19 and Figure 3.9, offers a perfect set to test for robustness of our proposed approach.

In Figure 3.18 we compare the performance of the originally trained detector (baseline) with the results of forward-backward consistency checked detections (AuCCA regularized), which shows a clear reduction of false positives. The poor performance of the baseline classifier is induced by the lack of meaningful background samples during training. Consistency checks within the forward-backward consistency checks allow for automatic generation of new positive and especially negative training samples, which are incorporated into a update of the detector via boot-strapping, leading to further enhanced detection results (retrained). Results in Figure 3.19 show the capability of AuCCA for mining new training samples by exploiting activity specific motions.

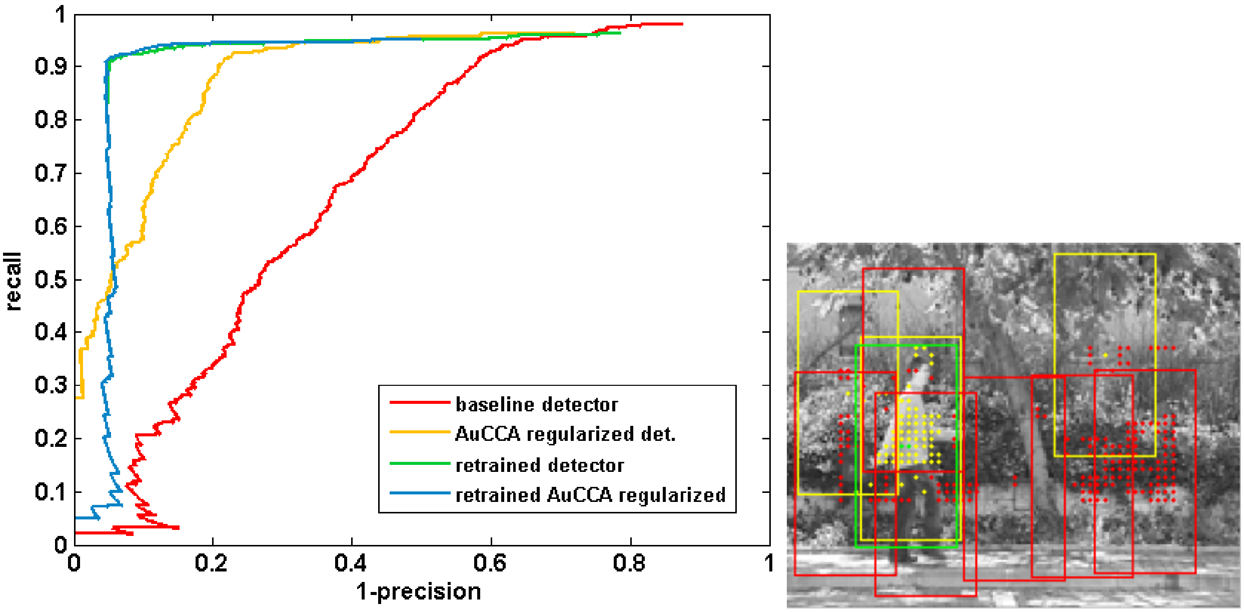


Figure 3.18: Left: Results of unsupervised activity detector training using AuCCA forward/backward mapping ($\tau = 5$). Right: Examples of positive(green), neutral(yellow) and negative (red) updates generated by the forward-backward regularization based on AUCCA motion predictions. The high amount of detections on background depicts the accuracy of the baseline detector.

3.5 Discussion of Results

Temporal relations are a vital source of information for many activity related tasks. At the time of the underlying publications for this chapter, related approaches were focused mainly on two subsequent frames to create motion features, or on the creation of global classifiers of complete activity videos. In addition, fully supervised training data was assumed as standard.

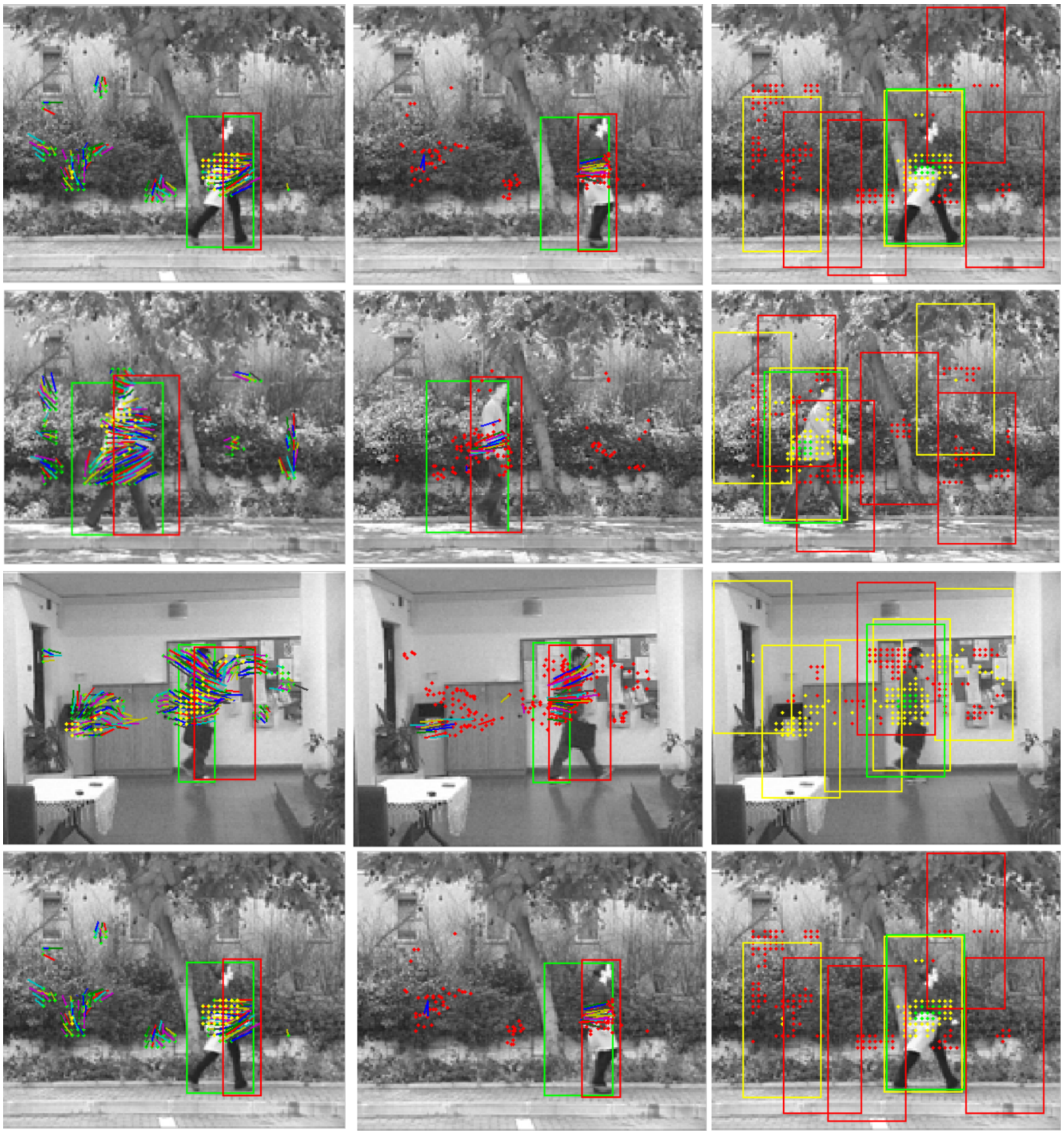


Figure 3.19: Several examples for sorting out wrong detections by forward and backward propagation with the proposed AUCCA method. Left column depicts the original detections within frame τ, with detections and estimated forward voting as colored points and lines. Green and red bounding boxes annotate the related ground truth for τ and $\tau + 3$, respectively. Middle column contains detection evaluation results on positions given by the voting lines from the left column in frame $\tau + 3$. Only valid detections again have a backward voting pointing to the estimated position in frame τ. Final validated detections in frame τ are visualized in the right column. Green bounding boxes depict new positive and red ones are handled as negative training samples for the next boot strapping round. Yellow samples are neglected for further training as they are assumed to be too uncertain to be added as new training data.

In contrast to those assumptions, the importance and advantages of observing local time windows in the context of their neighboring video frames has been shown within this chapter on two examples. First, the temporal prototype feature weighting (TPFW) and second the augmented canonical correlation analysis (AUCCA).

TPFW allows for learning the relations between different feature cues and their specific weighting for different temporal segments of an activity. This is realized with a prototype learning approach independently for each feature cues, followed by a probabilistic weighting scheme, were each prototype predicts the upcoming likelihood of all feature cues for following video frames. Experimental results show the advantage of TPFW compared to own previous publications and related state-of-the-art approaches. Even the comparison to feature cue learning mechanisms like MKL is favorable for TPFW.

AUCCA on the other hand, is a lightweight and efficient method to learn and predict classification and motion specific data in a combined way. The proposed AuCCA allows for unsupervised mining new training samples and post-processing of false positive detections by exploiting activity specific motions. Integrated into a particle filter framework the number of required particles can be drastically reduced, while still showing competitive results.

Although both methods have been discussed separately within this chapter, they share more than the underlying idea of exploiting temporal knowledge for better activity recognition and detection. A sequential or iterative combination of AUCCA for detection and regularization of activity sequences which would be input to TPFW would be obvious extension. Further, the intermediate representation found by AUCCA may be a vital information and pre-processing step before classification. The coefficients learned with AUCCA are much lower dimensional compared to the original one used within this paper, and optimize already for a representation of different classes within the coefficient space. To merge prototype based representation within hierarchies and AUCCA for feature representation and prediction of augmented data would have been logical extensions of the work presented in this chapter. Within literature one can find at least two examples of applying this idea for successful classification tasks, like the work by Donoser and Schmalstieg [39] and Rainforth and Wood within their Canonical Correlation Forests [128]. Within both publications CCA is applied as a pre-processing step to generate more natural feature representation using the CCA coefficients before training the classifier. In [128] this idea is extended to the Random Forest Framework by generating coefficient within each node of the trees.

Beside my papers which form the basis to this chapter [136, 159–161] , further outcome of this developments has found their way into the master thesis of Markus Murschitz [113], and Georg Waltner [165], who both were able to publish their findings on online activity classification for video games [114] and analysis of sports videos [166].

4 Learning and Detecting Activities from Ambiguously Labeled Data

The previous chapter analyzed the possibilities of substituting missing label information by temporal knowledge of activity motions. A lack of annotation information and background knowledge has been compensated by transferring activity detectors and learned motion patterns, for robust detection and retraining in new environments. Once enough data and labeling information is given, training efficient activity classifiers is solely dependent on the complexity and intra-class variability of the underlying activity videos. But all approaches discussed within the previous chapter assumed some kind of full supervision, according localization and labels of activities in videos. Training gets much more complex if such labeling is only given on a global basis, e.g., no bounding-box annotation of activities is given, or labeling information is even ambiguous, meaning more than one activity are happening simultaneously. In contrast, data annotation is a time consuming task and therefore, especially for all computer vision tasks within space and time domain, difficult to manage. This motivated the research presented within this chapter which focuses on training activity descriptors from ambiguous labeled data and tackling the task of solely global video descriptions and simultaneously executed activities.

Contents

4.1 Motivation

Many successful approaches are inspired by object detection and recognition methods, which are adapted into the spatio-temporal domain. Therefore, we have to deal with a large amount of data, which also increases the labeling effort. Moreover, as illustrated in Figure 4.1, the annotation of activities suffers from multiple problems as simple bounding-box descriptions may be too loose, non-rigid activities might change the required extent of the bounding box over time and thus often also incoherent background information is captured. If multiple activities may occur simultaneously, a global representation is insufficient and ambiguous, therefore leading to multi-instance multi-label problems when attempting to learn without supervision. But one would be interested in learning a compact, task specific and discriminative representation, which would be beneficial when dealing with large amount of data. But training of discriminative code-books is challenging when no distinct labels are available. Learning sparse code-books serving as compact and efficient mid-level representation has recently gained much attention for activity recognition. This is in particular beneficial if additionally label information is provided, and discriminative code-books can be generated. As already discussed within the experimental section Chapter 3, trained compact and low dimensional representation by basis vectors may generate comparable results in relation to high dimensional feature vectors. However, we are facing two main problems: First, providing labels for the data is not only very time consuming (especially, for large datasets) but also very complicated for the given task due to strongly articulated, prolonged and non-rigid activities. Second, existing discriminative sparse coding approaches are not robust in case of label noise or ambiguously labeled data (e.g., background motion or multiple parallel activities). In this chapter, both problems are tackled in parallel, by proposing a new sparse coding framework allowing for learning activities from ambiguous and global labels. This is realized by training pair-wise excluding code-books for class specific weighted representations. Such pair-wise comparison are exploited to iteratively decrease the ambiguity of labels. The proposed approach is capable of relabeling ambiguous samples and to create sparse discriminative representations jointly in an iterative adaptive manner. Within a bagging approach the term of pair-wise excluding sets is defined, where weighted sparse code-books are trained for each pair-wise label ambiguity separately within a bagging framework. Samples within a bag are encoded by those code-books, trained on independent out-of-bag data. Experimental results will show how to robustly relabel samples by ranking the coefficients of that encoding. As we build on ranking by pair-wise code-books trained with a soft-weighted non-negative matrix [96] factorization we denote our approach PR-SNMF. In addition we adopt discriminative sparse coding approaches, like JNMF [59] and LC K-SVD [6], to handle ambiguous labeled data. Within the experimental evaluation, we demonstrate the generality of our approach for diverse labeling problems with emphasis on activity recognition and detection. We evaluate our PR-SNMF approach on different ambiguous label problems and compare to related encoding (JNMF, LC-KSVD) and machine learning approaches like DA-Mil [55] and negative data mining [152].

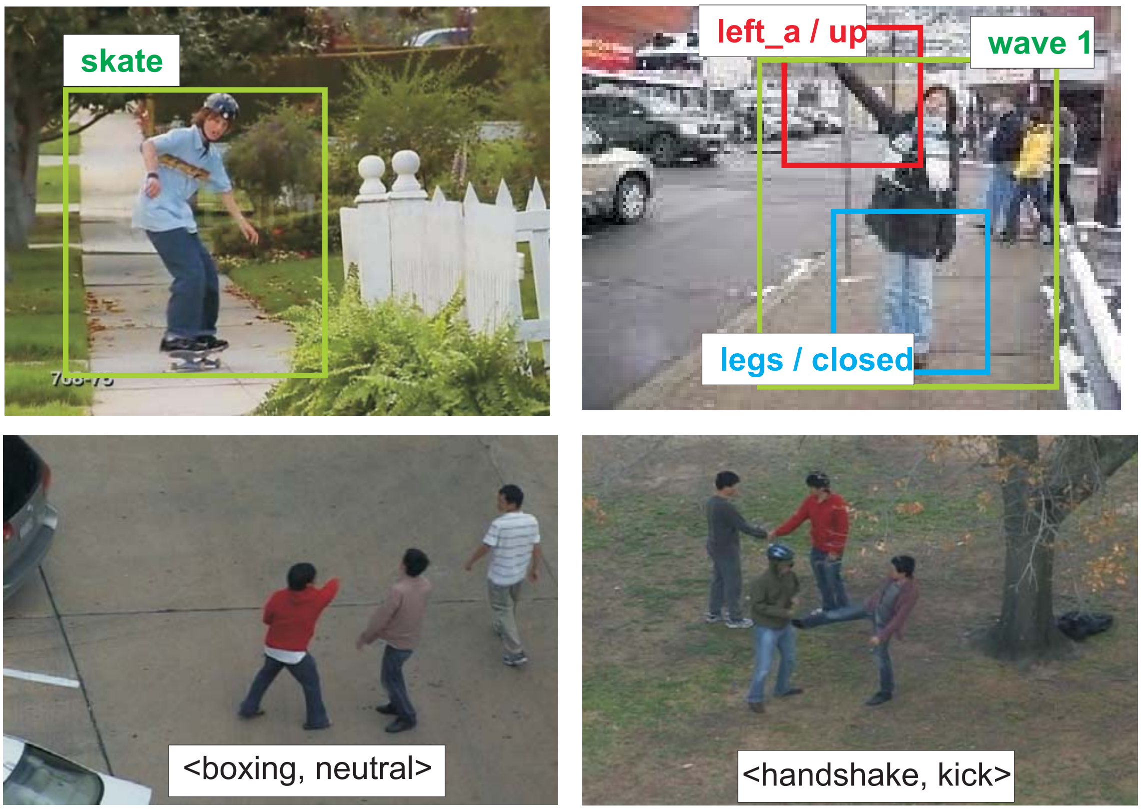


Figure 4.1: Different examples for activity annotation. Top row depicts an examples for coarse spatial bounding-box (left) and detailed body-parts annotation (right). Bottom row shows typical ambiguities if only global class labels are given for multiple activities within a video.

4.2 Related Approaches for Learning from Weakly Labeled Data

Solving the ambiguities of weakly labeled data for activity recognition was up to now strongly related to learn correct positions of activities versus the background noise, given a global temporal annotation on activities within a certain time slot [66, 92, 95, 146, 152]. This binary classification problem has often been interpreted as a multiple instance learning (MIL) problem, structuring the data into positive and negative bags, where positive bags must contain at least one positive instance, and there are not further restrictions for the negative bags. Hence, the global annotation information can be used to label those bags [66, 92, 146, 152]. Similarly, subtitle information was taken as weak label information in [17, 95].

Even though in this way the labeling effort can be drastically reduced, the correct selection of instances per bag is still a critical problem. To alleviate this problem, often general pedestrian detectors are applied [66, 152]. However, such approaches are failing if we have to deal with strong deformations or interactions, which are wrongly classified as non-person regions. Furthermore, general pedestrian detectors often generate many additional detections in crowded scenes, and therefore create additional ambiguities if not handled separately, e.g., by considering the amount of spatio-temporal key-points with the detection [152].

A more general approach is to use of a generic *objectness* detector [7] for generating instances for positive bags. But such approaches tend to favor highly textured or salient regions, which need not necessarily to be the activity regions as can be seen in [146]. Another variant is the separate training of a generic detector for specific datasets [52, 92], but this can be interpreted as additional indirect annotation information. [130, 185] train with a small amount of spatio-temporal labeled data and an additional larger set of ambiguous data. The problem of relabeling multiple concurrent activities per video has not been tackled by the above mentioned approaches.

Concerning feature representation, the above mentioned activity recognition methods build on highly engineered but generic spatio-temporal key-point detectors and descriptors. Without deprecating the importance of those developments, the strong variety of parametrization for individual parts of the feature pipeline for distinct datasets shows that there must be a better way of learning features for representation.

Ren and Ramanan [132] have shown that for object detection a learned sparse representation of local patches, aggregated into a histogram of sparse coefficients, can be much more expressive than HOG-based features. Also earlier works in activity recognition have demonstrated the beneficial behavior of sparse codes for supervised annotated data [25, 70, 136, 162, 193]. In particular [25, 162] show that the label likelihood of samples can be estimated from encoding coefficients, when distinct training labels are given. Learning discriminative code-books for classification of activities was limited to fully supervised data, like in [79, 124]. Nevertheless, all approaches showed the superior performance of a trained discriminative sparse code-book description compared to generic features representation. In a different domain Chen et al. [29] have learned discriminative sparse code-books for learning from ambiguously labeled data while creating a more expressive dictionary for classification. Concurrent possible active labels are problem known to learning faces from weakly labeled data like TV series [33]. Those approaches for learning from ambiguous labeled data are closely related to our approach

4.3 Learning from Ambiguous Labels

When learning a discriminative code-book or classifier, typically a sample $\mathbf{x}_i \in \mathbb{R}^d$ is associated with a distinct label y_i. However, as illustrated in Fig. 4.2, a sample can also be assigned multiple labels $l \in \{1, \ldots, L\}$, which can be represented by a vector $\mathbf{y}_i$. In the following, $\mathbf{V} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$ refers to a multi-label dataset, with data matrix $\mathbf{X} \in \mathbb{R}^{d \times N}$ and label matrix $\mathbf{Y} \in \mathbb{R}^{L \times N}$. In our case, each column $\mathbf{y}_i$ in $\mathbf{Y}$ is treated as a vector of label probabilities with $\mathbf{y}_i(l) \in [0, 1]$ and $\sum_l \mathbf{y}_i(l) = 1$. The goal is to estimate the correct labels for individual samples $\mathbf{x}_i$ while simultaneously creating a discriminative code-book for efficient representation of the data.

In the following, the generative and discriminative code-book learning via Non-negative Matrix Factorization (NMF) is reviewed for completeness. Following that, a new SoftNMF (SNMF) is proposed in Section 4.4.1, which allows incorporating label probability weights when building a sparse code-book. Then, in Section 4.4.2 and 4.4.6 the Pairwise Comparison Ranking (PR) is motivated with respect to SNMF representations and describe an iterative training scheme, finally yielding discriminative codes and providing the correct labels within T iterative updates of $\mathbf{Y}$.

4.3.1 Dictionary Learning with Non-negative Matrix Factorization

To create a mid-level representation for learning from ambiguous data, basic ideas from non-negative matrix factorization (NMF) are adopted for two reasons. First, many representations fulfill the non-negativity constraint and, second, the strictly additive combination allows for direct combination of individual bases. In general, let $\mathbf{X} \in \mathbb{R}^{d \times N}$ be a non-negative matrix, then the goal of NMF is to find a non-negative code-book matrix $\mathbf{D} \in \mathbb{R}^{d \times K}$ with K dictionary elements and an encoding matrix $\mathbf{H} \in \mathbb{R}^{K \times N}$ that approximate the original data by $\mathbf{X} \approx \mathbf{DH}$. Therefore, we consider the optimization problem,

$$\min \|\mathbf{X} - \mathbf{DH}\|_F^2 \quad \text{s.t.} \;\; \mathbf{D}, \mathbf{H} \geq 0 \;, \tag{4.1}$$

where $||.||_F$ denotes the Frobenius norm. Since there exists no closed-form solution, the matrices $\mathbf{D}$ and $\mathbf{H}$ have to be estimated in an iterative way. Using a gradient formulation, one gets an iterative solution for the optimization problem in Eq. (4.1) by the multiplicative update rules [96]:

$$\mathbf{H} \leftarrow \mathbf{H} \circ \frac{[\mathbf{D}^\top \mathbf{X}]}{[\mathbf{D}^\top \mathbf{DH}]} \qquad \mathbf{D} \leftarrow \mathbf{D} \circ \frac{[\mathbf{XH}^\top]}{[\mathbf{DHH}^\top]} \;, \tag{4.2}$$

where $\circ$ denotes the Hadamard product and $[.]/[.]$ an element-wise division. In the following, this this formulation will be referred to as Non-negative Matrix Factorization (NMF). To encode a new sample $\mathbf{x}_i$ during an evaluation, $\mathbf{D}$ is kept fixed and only the encoding step $\mathbf{H}$ in Eq. 4.2 is iterated.

4.4 Incorporating Label Information

The original NMF formulation cannot include information given by the label matrix $\mathbf{Y}$. However, due to the strictly additive combination of weights, it allows for direct combination of individual code-book matrices for a joint representation. Therefore, a straight forward way to incorporate label information is to *hard* select only samples $\mathbf{X}_l \subset \mathbf{X}$ with $\mathbf{y}_i(l) > 0$ or $\arg\max_{k\in L} \mathbf{y}_i(k) = l$. In this way, one can create L individual dictionaries $\mathbf{D}_l$ for each class and afterwards combine them again to one code-book $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_L]$. In the following, we will refer this approach to as *ClassNMF*. This process, however, leads to indistinct bases when strongly ambiguous labeled sets are given, where no distinct label information and class membership is given.

4.4.1 Soft Decision Non-negative Matrix Factorization

Although the above introduced ClassNMF approach incorporates discriminative information, it relies on a rather hard selection of training samples. Especially within an ambiguously labeled dataset, samples are either neglected or fully incorporated independent from their label probability.

Hence, this *hard* selection of training samples is replaced by a *soft* Non-negative Matrix Factorization (SNMF) formulation, taking into account the sample specific class probability $\mathbf{y}_i(l)$. This is realized by introducing the priority matrix $\mathbf{P}_l \in \mathbb{R}^{d\times N_l}$, where each column in $\mathbf{P}_l$ is equal to the label confidence $\mathbf{y}_i(l), \forall \mathbf{x}_i \in \mathbf{X}_l$. Hence, one can re-formulate the optimization problem in Eq. (4.1) to minimize

$$\|\mathbf{X}_l - \mathbf{D}_l\mathbf{H}_l\|^2_{\mathbf{P}_l} \;=\; \sum_{ij} [\mathbf{P}_l \circ (\mathbf{X}_l - \mathbf{D}_l\mathbf{H}_l) \circ (\mathbf{X}_l - \mathbf{D}_l\mathbf{H}_l)]_{ij} \;, \tag{4.3}$$

where $||\cdot||_{P_l}$ denotes the weighted Euclidean norm defined by the weights in $\mathbf{P}_l$. Thus, one gets the following multiplicative update rules:

$$\mathbf{H}_l \leftarrow \mathbf{H}_l \circ \frac{[\mathbf{D}_l^\top (\mathbf{P}_l \circ \mathbf{X}_l)]}{[\mathbf{D}_l^\top (\mathbf{P}_l \circ (\mathbf{D}_l\mathbf{H}_l))]} \quad \mathbf{D}_l \leftarrow \mathbf{D}_l \circ \frac{[(\mathbf{P}_l \circ \mathbf{X}_l)\, \mathbf{H}_l^\top]}{[(\mathbf{P}_l \circ (\mathbf{D}_l\mathbf{H}_l))\, \mathbf{H}_l^\top]} \;. \tag{4.4}$$

In this way, training samples are incorporated to each class-specific code-book according to their individual label confidence. Thus, a possible error induced by a *hard* selection of samples is minimized. Similar as for Eq. (4.2) it can be shown that the weighted Euclidean distance is non-increasing under these update rules. For the mathematical proof we would refer the reader to the PhD thesis of N.-D. Ho [62]. Finally, the thus obtained class-specific code-books are combined to a joint global code-book concatenation: $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_L]$. To encode a new sample $\mathbf{x}_i$ during evaluation, keeping $\mathbf{D}$ fixed and only the encoding step $\mathbf{H}$ in Eq. (4.2) is iterated.

4.4.2 Building Pairwise Ranking Sets for Solving Ambiguous Label Information

Even though the weighting SNMF approach introduced in Section 4.4.1 provides much better representations, the goal would be to generate code-books with high label purity per code-book vector, i.e., each vector should only be directly related to one specific label (see [124]). Although we strengthen the label-to-code-book relations with our SNMF formulation, the challenging problem remains to estimate the influence of a single code-book vector $\mathbf{d}$ for a distinct label l. Applying a weighted update scheme, samples from different labels are still intermingled and not directly related to specific coefficients in $\mathbf{H}$, caused by the ambiguity of the training data. To decrease the ambiguity, an iterative procedure is proposed in the following, where in each iteration the ambiguous samples are ranked according to their reconstruction by pairwise-label-excluding code-books, instead of label specific code books.

4.4.3 Out-of-Bag Training

To avoid a bias, the iterative label updates is performed in a bagging procedure. We split the training set $\mathbf{V}$ into B non overlapping bags, where $\mathbf{V}^b \subset \mathbf{V}$ denotes the evaluation bag and $\hat{\mathbf{V}}^b$ the corresponding out-of-bag training set $\hat{\mathbf{V}}^b \subset \{\mathbf{V} \backslash \mathbf{V}^b\}$. Within each $\mathbf{V}^b$ we analyze all pairwise label ambiguities, while samples from $\hat{\mathbf{V}}^b$ are the basis to create a discriminative code-book to relabel $\mathbf{Y}^b$.

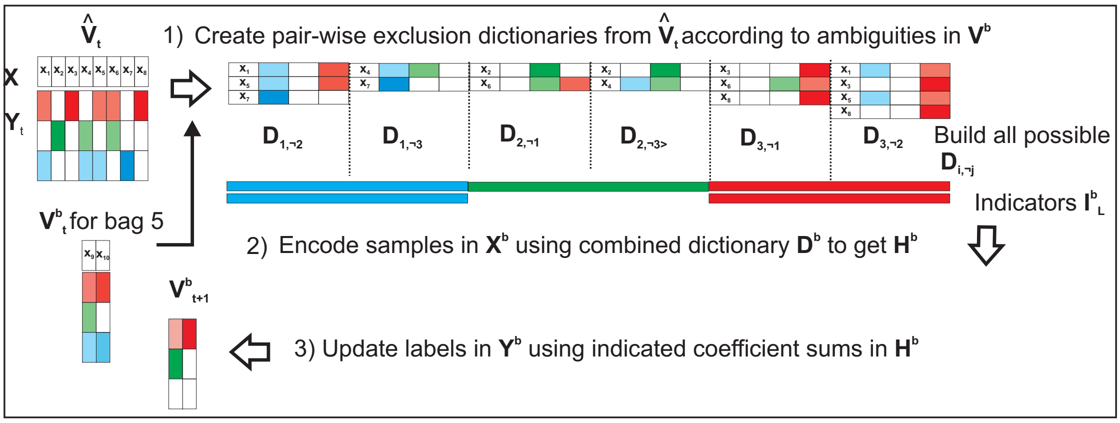

$\hat{\mathbf{V}}_t$
1) Create pair-wise exclusion dictionaries from $\hat{\mathbf{V}}_t$ according to ambiguities in $\mathbf{V}^b$
X
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
$\mathbf{Y}_t$
x_1 x_5 x_7
x_4 x_7
x_2 x_6
x_2 x_4
x_3 x_6 x_8
x_1 x_3 x_5 x_8
$\mathbf{D}_{1,\neg 2}$
$\mathbf{D}_{1,\neg 3}$
$\mathbf{D}_{2,\neg 1}$
$\mathbf{D}_{2,\neg 3>}$
$\mathbf{D}_{3,\neg 1}$
$\mathbf{D}_{3,\neg 2}$
Build all possible $\mathbf{D}_{i,\neg j}$
Indicators $\mathbf{I}^b_L$
$\mathbf{V}^b_t$ for bag 5
x_9 x_{10}
2) Encode samples in $\mathbf{X}^b$ using combined dictionary $\mathbf{D}^b$ to get $\mathbf{H}^b$
$\mathbf{V}^b_{t+1}$
3) Update labels in $\mathbf{Y}^b$ using indicated coefficient sums in $\mathbf{H}^b$

Figure 4.2: Visualization of one training iteration PR-SNMF for relabeling ambiguous data in a three class problem (blue,green,red), for one bag. 1) Create pairwise-exclude sets for all pairwise ambiguities, e.g., bag 5, from remaining data in $\hat{\mathbf{V}}$, and train dictionaries by SNMF. 2) Encode samples in $\mathbf{V}^b$ by a joined code-book and relabel $\mathbf{Y}^b$ using weighted coefficient sums, indicated by label priorities of individual code-books.

4.4.4 Pairwise Ranking with Excluding Code-Books

To cope with the ambiguously labeled samples, we adopt the idea of pairwise ranking for comparison (PR), e.g., [67]. As ranking ambiguously labeled samples cannot be learned using other samples sharing/including the same labels, the main idea of PR is to run pairwise excluding comparisons of labels. Within the context of discriminative code-book learning this can be realized as follows.

For each pair-wise label ambiguity $\langle i, j \rangle$ in $\mathbf{V}^b$, we build excluding code-books from samples $\hat{\mathbf{X}}_{\langle i, \neg j \rangle}$ and $\hat{\mathbf{X}}_{\langle j, \neg i \rangle}$ from of $\hat{\mathbf{V}}^b$, where $\langle i, \neg j \rangle$ denotes samples labeled with $y(i) > 0$ and $y(j) = 0$ (i.e., excluding samples labeled with class j), while all other labels are neglected. In this way, we can generate a maximum of $L(L-1)$ pairwise excluding code-books if all labels L occur in combination with each other within a bag.

For each combination $\langle i, \neg j \rangle$ in $\mathbf{V}^b$ we then compute $\mathbf{D}_{\langle i, \neg j \rangle}$ from samples $\hat{\mathbf{X}}_{\langle i, \neg j \rangle}$ using the weighted NMF updates defined in Eq. (4.4), with the corresponding priority matrix $\mathbf{P}_i$, describing the label confidences of samples in $\hat{\mathbf{X}}_{\langle i, \neg j \rangle}$ being from class i given by $\hat{\mathbf{Y}}$. Now we exploit the strictly additive combination of coefficients in NMF mentioned in Section 4.3.1, for encoding all elements $\mathbf{X}^b$ by a combined code-book. Let $\tilde{L}$ be the set of possible labels for $\mathbf{V}^b$, we get $\eta = s(s-1)$ pairwise exclusion sets with $s = |\tilde{L}|$. Therefore, we create a joint dictionary for encoding samples in $\mathbf{X}^b$:

$$\mathbf{D}^b = \left[\mathbf{D}_{\langle \tilde{L}(1), \neg \tilde{L}(2) \rangle}, ..., \mathbf{D}_{\langle \tilde{L}(s), \neg \tilde{L}(s-1) \rangle} \right] . \tag{4.5}$$

Encoding $\mathbf{X}^b$ with $\mathbf{D}^b$ using the update rules in Eq. (4.2), we obtain $\mathbf{H}^b$, with $\mathbf{h}_i^b$ an $\eta \cdot K$-dimensional column vector, where K is the number of dictionary elements per $\mathbf{D}_{\langle i, \neg j \rangle}$. The NMF update rule allows selecting bases from all subsets created beforehand, knowing which specific label was involved in creating the corresponding bases. As all NMF code-book vectors are normalized, the value of each coefficient in $\mathbf{h}_i^b$ is directly proportional to the influence of a code-element for the reconstruction of $\mathbf{x}_i$. Next, we define how to rank labels from the obtained $\mathbf{h}_i^b$ and update $\mathbf{y}_i$.

4.4.5 Label Ranking by Encoding Coefficients

As discussed above, the coefficients $\mathbf{h}_i^b$ are directly related to their importance for reconstructing $\mathbf{x}_i$. In other words, if many coefficients correspond to a certain code-book set $\mathbf{D}_{\langle i, \neg j \rangle}$, $\mathbf{x}_i$ has a smaller reconstruction error if built from samples labeled with i but certainly not with j (see Figure 4.2). As we know which coding vectors in $\mathbf{D}^b$ are constructed with priority to a certain label, we can derive a ranking for labels in $\tilde{L}_b$ according to selected coefficient sums by

$$r\left(y_i(l) | \mathbf{x}_i, \mathbf{H}^b, I_l\right) = \frac{\sum_{k \in I_l} \mathbf{h}_i^b(k)}{\sum_{k=1}^{\eta \cdot K} \mathbf{h}_i^b(k)} , \tag{4.6}$$

where I_l indicates all entries in h_i^b built with priority $\mathbf{P}_l$ for label l, during training the individual bases for $\mathbf{D}_b$ (see Figure 4.2 for a visualization). In particular, $r(y_i(n), \mathbf{x}_i) \succ r(y_i(m), \mathbf{x})$ denotes that label n is more likely for $\mathbf{x}_i$ than label m, as the normalized coefficient sum is higher, see [162].

4.4.6 Update Labels According to Encoding Coefficient

Assuming that all $r\,(y_i(l)|\mathbf{x}_i)$ sum up to 1, we can interpret $r(y_i(l)|\mathbf{x}_i)$ as an approximation of the likelihood $p(\mathbf{x}_i|l)$ of $\mathbf{x}_i$ being from a certain class, and perform a Bayesian update step:

$$\mathbf{y}_i^{t+1}(l) \approx \frac{p(\mathbf{x}_i|l)p_i(l)}{\sum\limits_{l\in L} p(\mathbf{x}_i|l)p_i(l)}\ , \tag{4.7}$$

where $p_i(l)$ denotes the prior given by $\mathbf{y}_i(l)$ at iteration t, e.g., the original labeling information when starting at $t = 0$. After performing the pair-wise ranking and re-labeling steps for all B bags, we obtain a new label matrix $\mathbf{Y}$. This alternating procedure of codebook generation and out-of-bag re-labeling is iteratively repeated during training until a maximum of T iterations is reached, or the average label difference of $\mathbf{Y}^t - \mathbf{Y}^{t-1}$ is below a predefined threshold. Solving these previous steps for each sample individually is infeasible for large scale data. Our approach is scalable in that respect, as we form bags and create necessary PR only according to pairwise labels within the bag (see Figure 4.2), and share $\mathbf{D}^b$ for encoding and relabeling individual $\mathbf{x}_i$. Thus, our approach scales with the number of pair-wise exclusion sets η per bag, see 4.4.4. However, since the degree of ambiguity (number of labels per samples) is in general relatively small, this can be easily compensated as the bagging allows effective parallelization. This leads to favorable run-times compared to other approaches, as will be shown in the experiments.

4.5 Comparison to Related Approaches

Besides the classical NMF variants developed within this chapter, let us compare to other discriminative sparse coding approaches, namely JNMF [43] and LC-KSVD [79]. As they are originally formulated to train from supervised data we adopt them to handle and work with ambiguously labeled training data as described in the following.

4.5.1 Joint Non-negative Matrix Factorization

To enforce the discrimination between different code-book entries and to leverage the supervised information (i.e. labels), several attempts of coupling feature and label space have been proposed ([59, 103]), summing up in a Joint Non-negative Matrix Factorization (JNMF). Learning a fully shared space among feature and auxiliary binary label information:

$$\min_{\mathbf{D},\mathbf{H},\mathbf{Q}} (1-\gamma) \left\|\mathbf{X} - \mathbf{DH}\right\|_2^2 \ + (\gamma) \left\|\overline{\mathbf{Y}} - \mathbf{WH}\right\|_2^2 \quad \text{s.t. } \ \mathbf{D}, \mathbf{H}, \mathbf{W} \geq 0 \ , \tag{4.8}$$

where $\overline{\mathbf{Y}} \in \mathbb{R}^{L \times N}$ denotes a *hard* selected binary label indicator matrix, e.g., by thresholding $\mathbf{Y}$, and $\mathbf{W} \in \mathbb{R}^{L \times K}$ the labeling code-book. The encoding $\mathbf{H}$ is jointly optimized for both input spaces, therefore JNMF is also denoted a multiview learning approach. As stated by [43], this adaptation could lead to descends into local minima. Thus, the joined factorization step is performed individually for all L classes, similar to SNMF in Section 3.2. The extension of the iterative updates in Section 4.3.1 are straight forward and can be found, e.g., in [43].

4.5.2 Label Consistent K-SVD

The original K-SVD algorithm [6] is an iterative approach to minimize the reconstruction error for $\mathbf{X}$ in Eq. (4.9) and learns a reconstructive dictionary for sparse representations of the input features. Γ is the sparsity constraint factor, defining maximum number of non-zero code coefficients per sample. The orthogonal matching pursuit algorithm (OMP) [163] is an efficient way for solving the coding of new test samples, given a trained dictionary. Jiang et al. [79] extended the optimization in Eq. (4.9) by a label consistency (LC-KSVD01) and an additional classification error (LC-KSVD02) term for enforcing discriminative solutions, weighted by α and β respectively:

$$\min_{\mathbf{D},\mathbf{H},\mathbf{A},\mathbf{W}} \left\|\mathbf{X} - \mathbf{DH}\right\|_2^2 \ + \alpha \left\|\mathbf{Q} - \mathbf{AH}\right\|_2^2 + \beta \left\|\overline{\mathbf{Y}} - \mathbf{WH}\right\|_2^2 \text{s.t.} \forall i, \left\|h_i\right\|_0 \leq \Gamma \tag{4.9}$$

In the original formulation in [79], the label consistency is defined in $\mathbf{Q} \in \mathbb{R}^{K \times N}$, a binary coding matrix describing where dictionary item and samples share the same label. It enforces the signals from one class to be represented by similar sparse codes, i.e. dictionary entries.

4.5.3 Extensions for Ambiguous Labels

During our evaluations we found that, as for the ClassNMF, the *hard* indication via a binary indicator matrix is not beneficial for ambiguously labeled data. We therefore implemented and adopt the JNMF by replacing the binary information in $\overline{\mathbf{Y}}$ by the corresponding label confidences given by $\mathbf{Y}$, denoted JNMF-soft within the experiments. For LC-KSVD we replace the *hard* binary selection by setting $\mathbf{Q}(k,i) = p(j|\mathbf{x}_i)$, if dictionary item k and sample $\mathbf{x}_i$ are supposed to share label j. As the dictionary is initialized from small independent dictionaries per class the coding in $\mathbf{Q}$ can be directly estimated from the label information in $\mathbf{Y}$. The labeling information in $\mathbf{Y}$ is directly optimized in Eq. (4.9). Again we adopt this part by using $\mathbf{Y}$, the real valued label confidences, directly instead of binary class information. We directly adopted the LC-KSVD code[1] provided by the authors of [79], denoted LC-KSVD-soft in the reminder of this paper. For a more detailed explanation how to build $\mathbf{Q}$ and to initialize $\mathbf{D}$, $\mathbf{A}$ and $\mathbf{W}$ we would like to refer the reader to [79]. For a fair comparison, both methods (JNMF, LC-KSVD) are trained in the same scheme of iteratively updating $\mathbf{Y}$ and also $\mathbf{Q}$ after each training iteration.

4.6 Experiments

The purpose of the experiments is twofold. First, to demonstrate the benefits of the proposed PR-NMF algorithm when learning from ambiguously annotated samples, which is demonstrated for two totally different tasks within Section 4.6.1. On the one hand, a novel training sets with controlled ambiguity building on the famous Weizmann dataset is generated. On the other hand, a balanced subset of the famos *Labeled Faces in the Wild* data set, namely FIW(10b) [33], to demonstrate robust re-labeling of ambiguous labeled samples. Those experiments shall evaluate all steps proposed in Section 4.4 and vindicate the improvements from NFM, to SNMF and PR-SNMF. In addition, a fair comparison to related state-of-the-art can be performed, and the proposed improvements within Section 4.5 can be evaluated.

Second, in Section 4.6.2 we present joint activity localization and recognition results for the the UT-Interaction dataset [139], being the first to learn such models without any geometric annotation information. In fact, compared to the state-of-the-art we get competitive results, even though only very loose labels (*i.e.* , on video level) have been used. Such experiments shall verify the applicability of PR-SNMF for localizing multiple parallel performed activities in videos, and the meaningfulness of the generated SNMF dictionaries as representation for activity classification.

4.6.1 Robust Re-Labeling of Ambiguous Training Data

A profound evaluation of algorithm parameters and comparison to state-of-the-art methods is done using two datasets with controllable ambiguity. Following the experimental setups in [33] and [29], the label confidence after training and testing for different levels of ambiguity is evaluated. In fact, the variations of the parameters q (the number of additional labels per sample) and ϵ (the degree of ambiguity – the maximum probability of an extra label co-occurring with the true label) are evaluated.

For a fair comparison of all evaluated approaches iteratively re-label the training data, and test errors are computed with identical classifier, namely linear regression classifiers for Weizmann and minimum per-class reconstruction errors for FIW(10b). The code-book size per class is set to $K = 10$ for all methods in the ambiguous Weizmann dataset, and to $K = 20$ for FIW(10b) respectively. The number of training-update iterations T is set to

[1] http://www.umiacs.umd.edu/ zhuolin/projectlcksvd.html

5, while the number of internal iterations for code-book generation, e.g., updates in NMF Eq. (4.2) and (4.4) and for K-SVD Eq. (4.9), is set to 50 for all methods.

4.6.1.1 Results on Ambiguous Weizmann dataset

The Weizmann dataset consists of 10 single person activities repeated by 9 different actors. A more detailed description is given within the previous chapters 2 and 3. For our purpose, let us create ambiguous labeled training data by combining up to 5 activities simultaneously in one video, *i.e.* , q = 4 (see Figure 4.3 or 4.4).

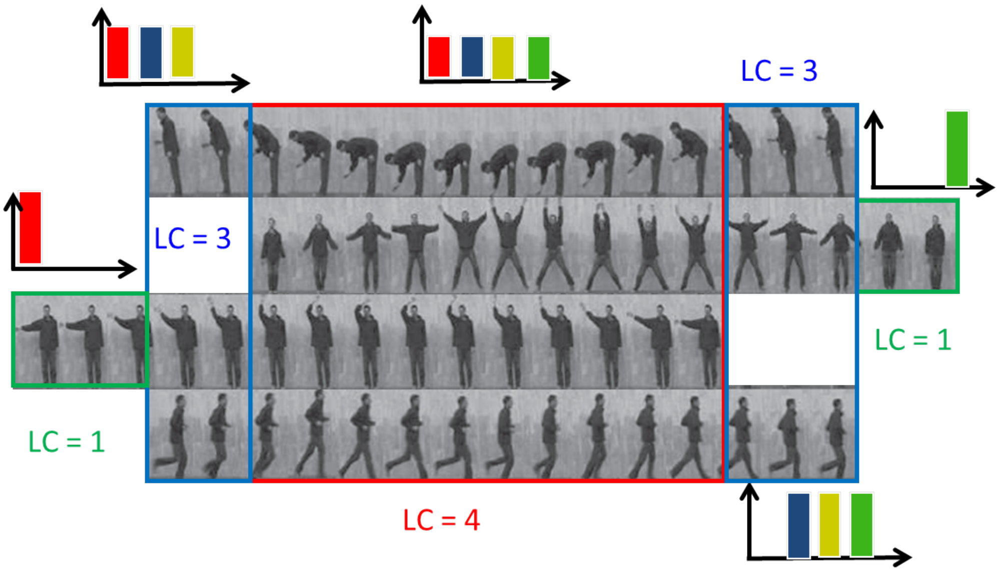


Figure 4.3: Example of a ambiguous training bag generated for the ambiguous Weizmann dataset experiments, with different label cardinality (LC).

The combination of activities is chosen at random, meaning ϵ is set to 0, and each *combined* video with concurrent labels is forming a bag within the PR-SNMF approach. The higher q, the smaller is the number of bags, as activities are randomly selected without replacement. Assuming that detections are given, the corresponding patches are represented using HoG/HoF features (similar to [70, 162] or [136]). To demonstrate the benefits of the proposed PR-SNMF approach, we show two comparisons. First, we compare the PR-SNMF and its evolutionary approaches (NMF, SNMF) to different encoding methods and adaptions discussed in 4.5. Optimal parameters for JNMF ($\gamma = 0.5$) and LC-KSVD ($\alpha = \beta = 0.1$ and $\Gamma = 30$) where found empirically and are in the range of recommendations from literature. Second, a comparison to related Multiple Instance Learning (MIL) approaches is given.

Exemplary labeling matrix for an training run with $q = 4$ is depicted in Figure 4.4. As activity videos are grouped together, subsequent samples share the same labeling information, but due to the random combination the matrix is rather noisy. Average label confidence show the low initial confidence, except for some longer videos with unique labels at the beginning or end of longer sequences, see also Figure 4.3 for explanation.

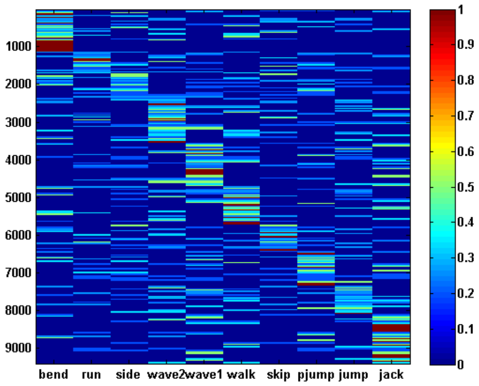


Figure 4.4: Exemplary initial labeling information $\mathbf{Y}$ of our ambiguous Weizmann dataset for $q = 4$. The average number of labels per training sample, also known as *Label Cardinality* (LC), is 3.85.

Running the pair-wise ranking NMF without weighting clearly optimizes the label distribution (high diagonal values in Figure 4.5), but still with several outliers and overall low confidence for the expected labels. The average label certainty of the diagonal elements is clearly lower compared with the PR-SNMF results in Figure 4.6.

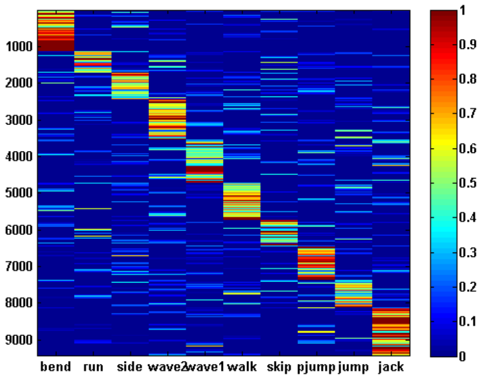


Figure 4.5: Resulting confidence matrix for training data in 4.4 after 2 iterations when applying pair-wise comparison with standard NMF code-books (PR-NMF) without weighting

The overall label statistics for all cross-validation runs is depicted in Figure 4.7. Figure 4.7a depicts the average label confidence, meaning the probability of the correct label after training, for different levels of ambiguity. These are more expressive than pure training errors in terms of correct classification rates given for completeness in Fig. 4.7b. As expected, the generative approach (NMF) and binary selecting discriminative approaches

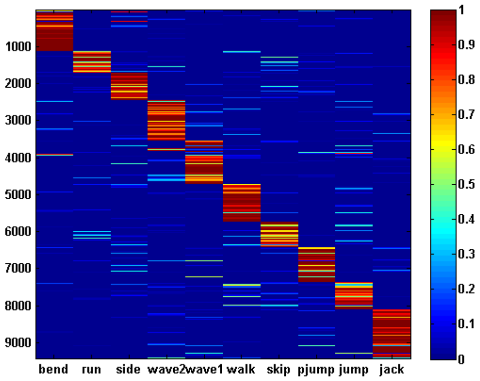


Figure 4.6: Resulting confidence matrix for training data in 4.4 after 2 iterations using the proposed PR-SNMF approach.

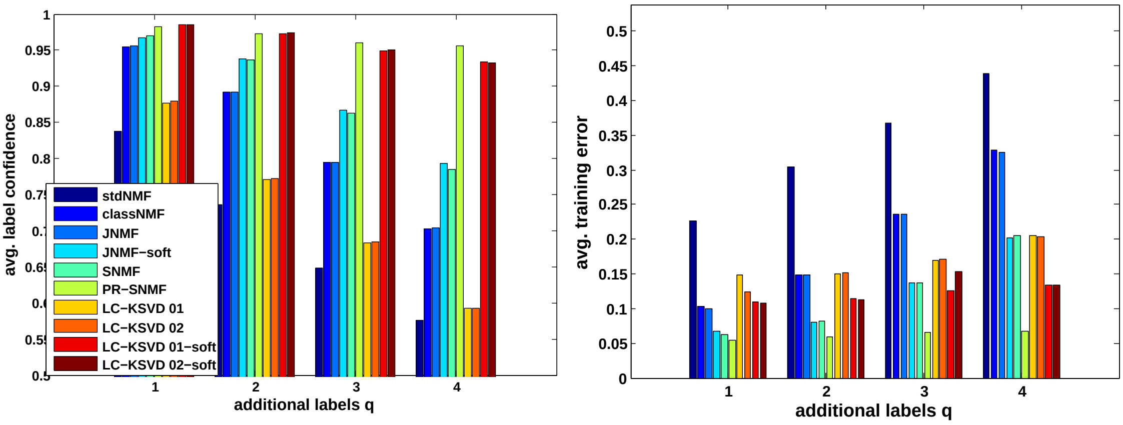
(classNMF, JNMF, LC-KSVD) have a strongly decreasing performance for higher degrees of ambiguity. Overall the label purity for KSVD versions is lower, while the training error it self seems not to be affected by that.

The benefit of the proposed *soft*-selection and the adaptations to state-of-the-art algorithms described in this chapter (SNMF, JNMF-soft, LC-KSVD-soft) is clearly visible, especially for KSVD and the training errors. The great improvement for LC-KSVD arise from the missing capability to adopt the label confidences as the optimization is chained to reconstruct the binary label information in $\mathbf{Q}$ and $\overline{\mathbf{Y}}$, see Section 4.5 for more details. The overall training error of LC-KSVD is comparable, only the confidence of the predicted correct labels is low. Clearly, PR-SNMF compensates an increased ambiguity better than all other methods, followed by the modified LC-KSVD-soft versions. Concerning the progress of relabeling during the training iterations all approaches show similar behavior as depicted in Figure 4.8.

When training a discriminative code-book from ambiguous data, the desired results after the relabeling training process are high dictionary purity (the maximum label probability per dictionary item) Figure 4.9a and compactness (low similarity of dictionary items estimated by $\mathbf{D}^T\mathbf{D}$) Figure 4.9b. More compact and distinct labeled code-books should lead to better classification results, as depicted in Figure 4.10. Interestingly, all NMF variants create more compact and distinct code-books, compared to K-SVD (see Fig. 4.9)

In addition, let us compare the proposed approaches to two Multiple Instance Learning (MIL) approaches, which can be adopted to our multi-label problem by forming several binary one-vs-all tasks. First, to a recently proposed *negative-data-mining* MIL (NM-MIL) framework [152], second, to a *deterministic annealing* Support Vector Machine (DA-SVM) [55], from a publicly available machine learning toolbox[2], which performs well on classical MIL benchmarks. Both approaches are trained directly on concatenated

[2] http://people.kyb.tuebingen.mpg.de/pgehler/mil/mil.html



(a) (b)

Figure 4.7: a) Average label confidences for correct classes after training with different levels of ambiguity and b) classification error on the training data of ambiguous Weizmann setup. The proposed PR-SNMF compensated higher ambiguity during training better than all compared approaches.

HoG/HoF features, while our baseline is a linear SVM, trained with a $200-dimensional$ code-book (20 per class), created with NMF/SNMF after PR-SNMF for solving the ambiguity, on top of the HoG/HoF features.

As first result we found that these approaches are strongly affected by the used bag size, and do not compare favorable if only trained on large bags as the proposed PR-SNMF. Therefore, an adaption has to be made by changing the bag-size such that they contain only data of 5 consecutive frames – typically, several hundred bags per training – significantly improving their results, which, however, are still far below the PR-SNMF approach.

Table 4.1 summarizes recognition results of PR-SNMF, compared to NM-MIL and DA-SVM, over different ambiguity levels q. Although having smaller dimensionality, the benefit of a task specific and discriminative representation is clear, compared to the underlying HoG/HoF features. PR-SNMF approach scales with a higher degree of ambiguity during training, as more pair-wise comparisons per bag are generated and the information from all labels may be used, compared to the binary one-vs-all setup in NM-MIL and DA-SVM. In addition, this shows the difficulty of our ambiguous Weizmann experiment, and that joint learning of discriminative codes and relabeling is very beneficial.

4.6.1.2 Results on Labeled Faces in the Wild dataset

Besides human activities, facial expressions are an related approach. Next, let us evaluate the re-labeling capabilities of our proposed PR-SNMP using an artificial dataset building on the well known and widely used Labeled Faces in the Wild dataset. This should also

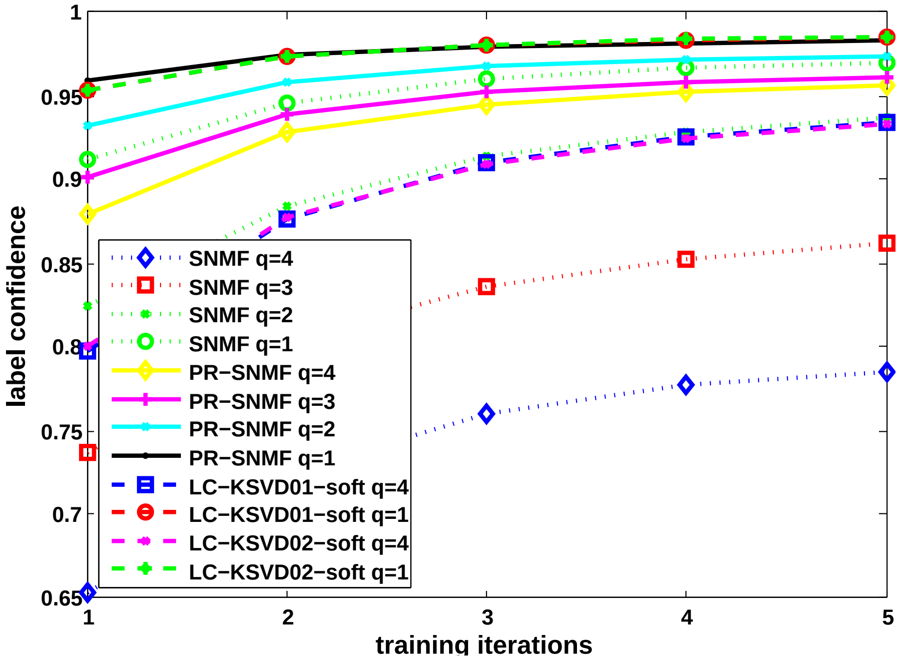


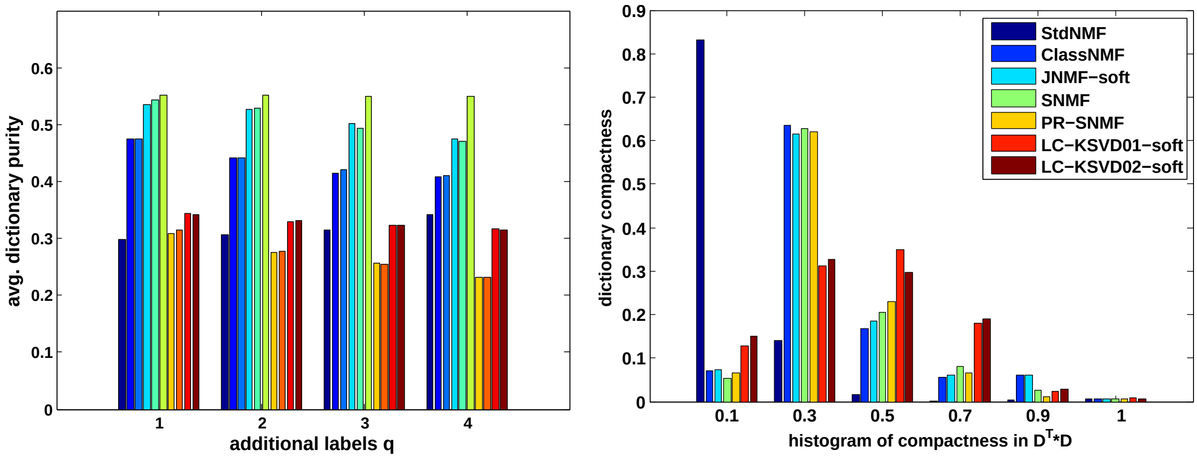
Figure 4.8: Behavior of the label confidences during the iterative training steps on ambiguous Weizmann setup. Proposed PR-SNMF and adopted JNMF and LC-KSVD *soft*-versions make significant progress in the first two iterations.

	Rec. rate / frame [%]		
q =	1	2	3
SNMF 01	$\mathbf{89.77 \pm 2.1}$	89.1 ± 2.2	87.5 ± 2.9
SNMF 05	88.8 ± 3.4	$\mathbf{89.2 \pm 2.5}$	$\mathbf{87.7 \pm 2.5}$
NMF 01	87.5 ± 3.6	86.1 ± 3.3	84.4 ± 4.8
NMF 05	86.08 ± 3.7	87.1 ± 3.8	85.7 ± 1.9
NM-MIL [152]	71.1 ± 6.7	68.1 ± 9.6	-
DA-SVM [55]	44.0 ± 2.24	37.5 ± 5.4	-

Table 4.1: Results of experiments on per frame classification for different levels of ambiguity within our tests on Weizmann dataset. The overall video classification results are between 95% and 99%, depending on the ambiguity of the training set.

demonstrate that the capability of PR-SNMF is quite general and not limited to activity recognition. In particular, the FIW(10b) dataset is used, consisting of the first 50 images of the 10 most frequent subsets of the FIW dataset.

Following the experimental setup in [33], to compare to their results (*mean*, *naive*, *model_one*) and in addtion to [29] (DLHD, DLSD) to the PR-SNMF approach, using the same features as input for all approaches, namely histogram equalized images as input. In particular, the dataset contains cropped faces which are additionally to the true label augmented with a varying number of q additional ambiguous labels. In contrast to the



(a) **(b)**

Figure 4.9: a) Average purity of dictionary items after with different levels of ambiguity. With more ambiguous labels q initialized, results of less adoptive values break down, while adoptive ones can compensate for this ambiguity. b) Histogram of dictionary compactness when training with ambiguous Weizmann with $q = 4$, visualized as a histogram. Lower values mean less correlation between dictionary vectors. As the classical NMF does not allow for sharing labels, the dictionary compactness is outstanding, but on cost of worse dictionary purity and classification results. Overall NMF dictionaries are more compact compared to SVD approaches.

ambiguous Weizmann setup beforehand, the ambiguity between labels is not random but correlated ($\epsilon \in [0.1, 1]$ with $q = 1$). The proportion p of ambiguous labeled samples is always 1 (all samples have q additional labels). Thankfully, the authors of [29] provided the result curves of the compared methods.

The results of evaluation against corresponding results, obtained by the authors of [29, 33] are illustrated in Fig. 4.13. It can be seen that our method compares favorable to state-of-the-art or even outperforms [29] if the number of ambiguous labels per sample is increasing. In fact, our approach scales with the complexity of the ambiguity, *i.e.* , more pairwise exclusions $\langle i, \neg j \rangle$ are created if more labels are active per sample. This leads, if required, to a higher-dimensional and a more discriminative representation, providing more distinct labels as a result. For smaller values of q the results are comparable to the state-of-the-art, while all methods are going to fail if q is getting too large (see Fig. 4.13a). The same behavior can be observed for correlated label ambiguities within the ϵ evaluation in Fig. 4.13b. Please note that no distinct parameters for generating the code-books used in [29] are mentioned in their paper.

4.6.2 Localizing Activities in Weakly Labeled Videos

Finally, we demonstrate our approach for localization activities within weakly, ambiguously labeled data. We start our training from global labels per video frame, defining which

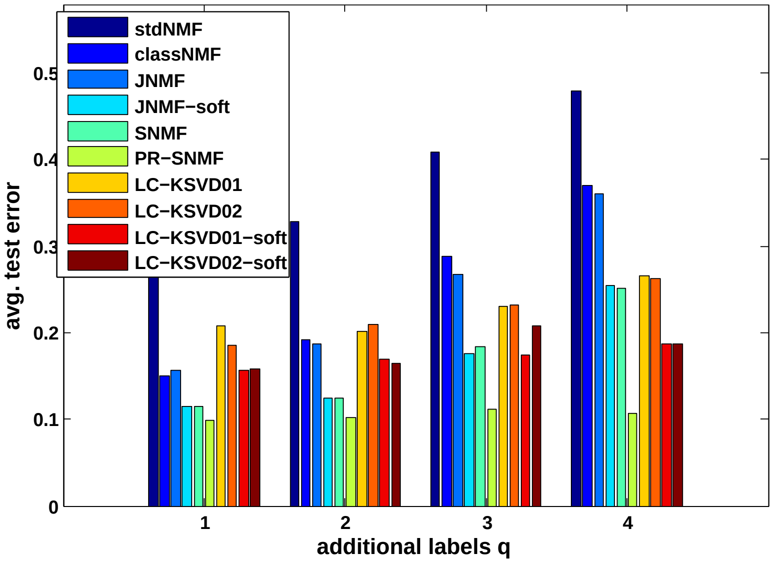


Figure 4.10: Average test errors on ambiguous Weizmann setup.

Figure 4.11: Labeled faces in the wild.

activities are currently active. The only additional information given is the average size of a person for defining the scale σ. In contrast to previous works [92, 146, 152], we do not rely on any pre-trained detector. Instead, we apply a simple motion compensation to handle camera jitter and sample features densely from moving areas, with a patch size of $\frac{\sigma}{4}$. Every local patch i is described by HoG/HoF features [95], mapping texture and motion information of two consecutive frames, which form $\mathbf{X}_i$.

4.6.3 Learning Activity Descriptors from Videos

The above described PR-SNMF relabeling approach can be applied in general to all possible applications on learning from ambiguously labeled data (see Section 4.6.1). Now we describe the general processing cue for detecting local spatio-temporal activity annotations from global temporal labels. Finally, we show how obtained detections are described by a feature vector based on weighted SNMF bases and the histogram of sparse weighted codes (HSWC).

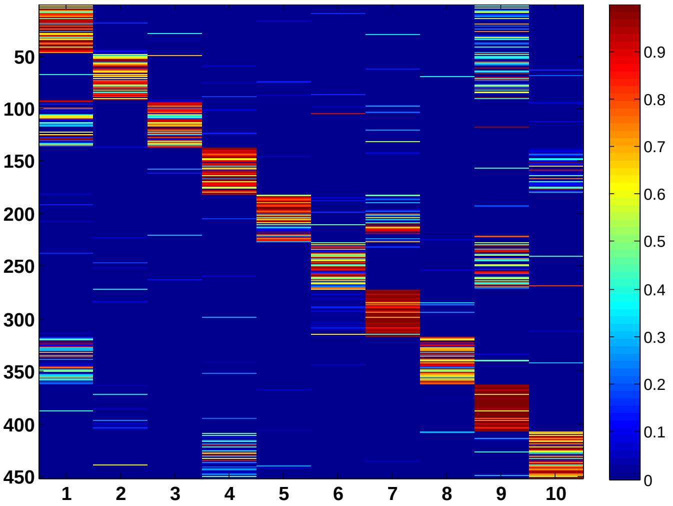


Figure 4.12: Results on FIW(10b) dataset by relabeling training data with strong correlated ambiguity ($q = 1, \epsilon = 0.95$).

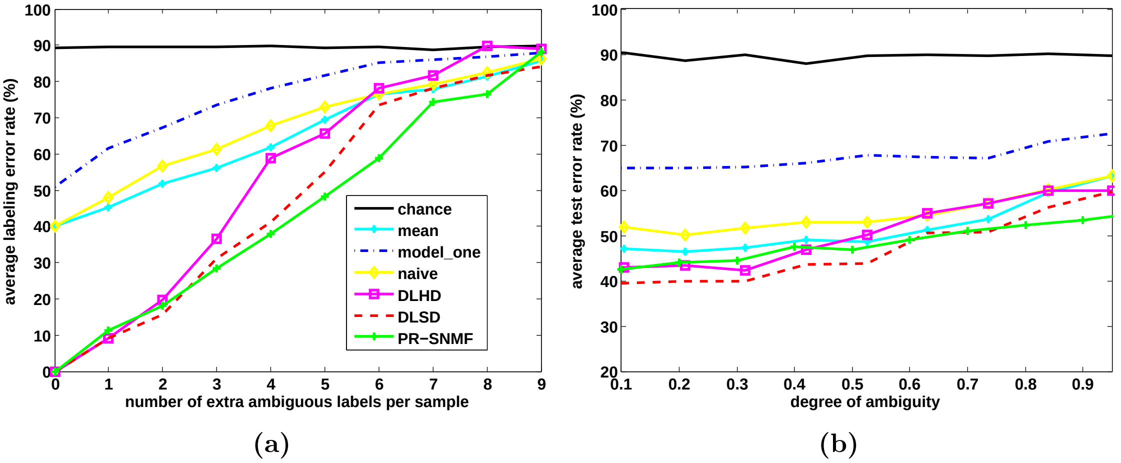


Figure 4.13: Experiments on FIW(10b) dataset. a) Average labeling error rates versus the number of extra labels per sample. b) Average test error rates versus the degree of ambiguity ϵ while $q = 1$. Although rather simple, PR-SNMF outperforms the more specialized approaches.

4.6.3.1 Initializing local Label Confidences

Each local patch gets assigned the corresponding global label information $\mathbf{y}_i$. In addition, and different to previous work, we define an additional neutral label. This is not only important for collecting samples generated by background motion, but also for parts of activities that are shared between activities and therefore are not discriminative. Figure 4.14 (top row) depicts the initial probabilities for salient samples within the selected global annotations. After 3 iterations of our PRS algorithm presented in Section 4.3 we obtain new labels for each sample (again color coded shown in Figure 4.14 (middle row)). Discriminative patches clearly define activity related parts, while background or ambiguous motion is suppressed. After local smoothing we obtain vote maps for each active label, shown for active activities in Figure 4.14 (bottom), used for non-maxima-suppression to obtain the final detections passed to the next training stage (shown in green).

4.6.3.2 Histograms of Weighted Sparse Codes

Previous work on activity recognition has adopted NMF for replacing the underlying feature representation of HoG and HoF by NMF coefficient vectors [70, 136, 162]. In general, the whole patch containing an object of interest was described by such a coefficient feature vector. Similarly, in [132] small local images patches are described by their sparse coefficients and an object is described with a histogram of those sparse codes. For both setups an improvements compared to working with the underlying raw gradient and motion features can be obtained, as the sparse coding is specifically learned for the task.

After performing T iterative updates as described in Section 4.4.2, we obtain our final set of training detections (Figure 4.14 (middle)). The first difference to the above mentioned approaches is that we create again a concatenated representation of weighted bases from labeled samples in training set $\mathbf{X}$. All samples in $\mathbf{X}_l$ with $\mathbf{y}_i(l) = \arg\max_l y(l)$ and within a training detection for $y(l)$ are taken to create $\mathbf{D}_l$ using Eq. (4.4). Their corresponding label confidence for l are the priority weighting in $\mathbf{P}_l$. Running these steps for all L possible classes we obtain a $d \times (L \cdot b)$ dimensional basis matrix $\mathbf{D}^c$ through concatenation, where b is the number bases per $\mathbf{D}_l$. Projecting a sample $\mathbf{x}_i$ onto $\mathbf{D}^c$, we get an $L \cdot b$ dimensional representation. Again high values in $\mathbf{h}^c(j)$ denote the importance for the basis j in $\mathbf{D}^c$ for the reconstruction of $\mathbf{x}_i$. Following [132] we build the final descriptor by aggregating the coefficients of patches within a training detection into a histogram of sparse codes. We perform a simple version of a pyramidal histogram approach, where a detection with given temporal extend τ is additionally divided into four slightly overlapping sub-volumes, to represent spatial and temporal relations within activities.

4.6.3.3 Activity Localization on the UT-Interaction Dataset

Originally introduced by Ryoo and Aggarwal [139], the UT-interaction data set[3] contains multiple interacting persons, individual pedestrians and multiple activities executed simultaneously, and is divided into two sets #01 and #02, with 10 videos each. The dataset describes six realistic human interactions including shaking hands, pointing, hugging, pushing, kicking and punching. The strong intra-class similarity and the non-repetitive nature of the activities make the recognition task very complex. In contrast to other activity datasets, the UT-Interaction videos contain several activities in parallel and vast majority of neutral phases without any significant activity, but nevertheless humans in motion. The

[3] http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html

simultaneous interactions are clearly a multi-label problem, if one wants to avoid labeling exact spatio-temporal positions for training.

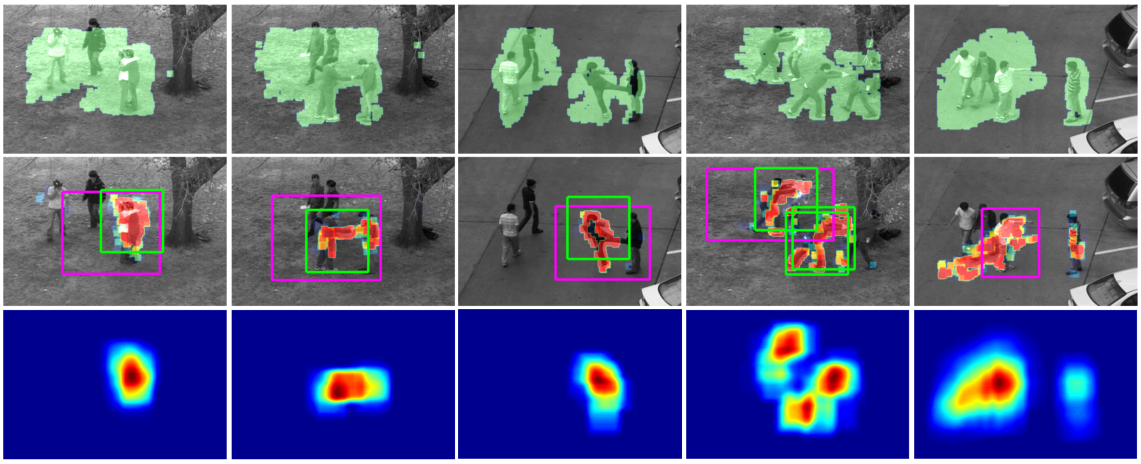


Figure 4.14: Visualization of the learning process on UT-interaction dataset. Each column represents an independent annotation. Top: Initial label probability is uniform over all patches (aggregated over time). Middle: Confidences of patches corresponding to the global activity label after PR-SNMF updates, with final estimated local spatio-temporal annotation (green) generated from vote maps (bottom row) and ground-truth annotations (magenta) for comparison. Last column: Missed annotation due to spatial overlapping background motion.

Figure 4.14 (top row) depicts the initial probabilities for salient samples within the selected global annotations. After 3 iterations of our PR-SNMF algorithm we obtain new labels for each sample (again color coded shown in Figure 4.14 (middle row)). Discriminative patches clearly define activity related parts, while background or ambiguous motion is suppressed. After local smoothing we obtain vote maps for each active label, shown for active activities in Figure 4.14 (bottom), used for non-maxima suppression to obtain the final detections passed to the next training stage (shown in green). It has to be mentioned that we do not limit our approach to return only one detection per active label. Therefore, several instances of the same activity per frame may be detected. From this final detections we learn our discriminative features by forming 20 bases per activity which leads to a final 140 dimensional feature (6 activities plus one neutral label) per sampled patch. The final descriptor for learning a linear SVM in a one-vs-all setup is a 700 dimensional pyramidal histogram of the PR-SNMF coefficient within the detected positions (σ pixel) and temporal extent of $\tau = 40$ frames.

Evaluation is done on cropped sequences for comparing recognition results to state-of-art. As reported by others, we evaluate using a 10-fold cross-validation per set, always keeping one video for testing. Comparisons to related approaches are shown in Table 4.2. Although we train our proposed approach without any spatial prior, we compare favorable or equal to other histogram based approaches like [139, 186] or even approaches with stronger spatial and temporal relation descriptors like [21, 54]. Approaches with part-based annotations like [130] or Hough voting [185, 186] yield superior results, however, with a significantly higher annotation effort. We want to highlight at this point that our approach could also serve as an initial step before learning stronger descriptors with higher flexibility for spatial-temporal relations.

method	anno	Set 01	Set 02
proposed	weak	80.0%	75.2%
[186]	bb	66.67%	-
[139]	full	70.6%	-
[138]	bb	85.0%	-
[21]	full	78.9%	-
[54]	bb	65% - 79%	-
[186] +RF	bb	83.33%	-
[185]	weak	75.0%	-
	bb	93.3%	91.7%
[130]	full	93.3%	-

Table 4.2: Comparison of UT-interaction recognition results with related work. The *anno* column depicts the level of supervision, where *weak* stays for global temporal labels, *bb* denotes bounding box annotations and *full* describes additional annotations/segmentations of sub-parts or atomic actions for training.

4.7 Discussion of Results

Learning detectors or classifiers without clearly labeled data is a challenging task. Within this chapter a novel method for learning from ambiguous labeled activity was presented. This PR-SNMF approach was evaluated and compared on various challenging tasks, and in addition several related state-of-the-art methods have been adopted to get further understandings. By building code-books from pair-wise excluding sets within a bagging framework, we are able to solve the ambiguity of samples, while in addition finding discriminative samples per activity.

The pair-wise comparison within PR-SNMF scales better with the degree of ambiguity, therefore showing better results compared to related approaches when the level of ambiguity increases. These samples can be exploited to find spatio-temporal annotations in the context of learning from ambiguous global temporal labels. The proposed learning approach is not limited to the task of activity recognition and has demonstrated its performance on diverse applications. As shown, PR-SNMF could also serve as an initial annotation step for other activity recognition methods by providing spatio-temporal annotations or discriminative features, which bears good prospects for further enhancements.

One major outcome of the studies was the influence of background samples or neutral activities, not related to a specific activity and shared over the majority of activities. Simple motion estimation and compensation for global motion showed to be not discriminative enough to be applied as a pre-selection process for generating training samples. Those findings let to the ideas of applying so called saliency estimation to activity videos, for weighting prospective areas and samples. Unsupervised activity detection by saliency estimation is developed within chapter 5.

Besides the work described within this chapter, my studies on non-negative matrix factorization for activity recognition found their way into several publications [136, 159]. In addition, together with my colleague Mahdi Jampour, we extended the approaches to facial gesture estimation under various poses and successfully published the following conference papers and journals [73, 74].

5

Unsupervised Activity Detection by Encoding based Saliency

In previous chapters, the presented methods have been based on different degrees of supervision during training, either by given ground truth annotations or knowledge about activity presence or absence in videos. In this chapter, the circumstance that activities form salient areas within the video volume is exploited to find activity relevant areas. Therefore, a novel video saliency detection method is presented, to support human activity recognition and weakly supervised training of activity detection algorithms. Recent research has emphasized the need for analyzing salient information in videos to minimize dataset bias or to supervise weakly labeled training of activity detectors. In contrast to previous methods the proposed method does not rely on training information given by either eye-gaze or annotation data, but propose a fully unsupervised algorithm to find salient regions within videos. In general, the Gestalt principle of figure-ground segregation is enforced for both appearance and motion cues. The introduced encoding approach allows for efficient computation of saliency by approximating joint feature distributions. Evaluations of the approach on several datasets, including challenging scenarios with cluttered background and camera motion, as well as salient object detection in images are performed within the experimental part. Overall, the encoding-based saliency estimation demonstrates favorable performance compared to state-of-the-art methods in estimating both ground truth eye-gaze, activity annotations and improved classification and detection scores.

Contents

5.1 Motivation

Estimating saliency maps or predicting human gaze in images or videos recently attracted much research interest. In particular, by selecting interesting information based on saliency maps irrelevant image or video regions can be filtered. Thus, saliency estimation is a valuable pre-processing step for a large domain of applications, including activity recognition, object detection and recognition, image compression, and video summarization. Especially activity recognition, and in particular activity localization can be subject to variations of positions, viewpoint change, size and therefore hard to train by few examples or rules. This leads to a higher annotation effort compared to classical object detection methods. Moreover, the majority of videos contain background information which may lead biased classifiers when based on global video descriptors.

Salient regions contain per definition important information which in general is contrasted with its arbitrary surrounding. For example searching the web for the tag ”horse riding” returns images and videos which all share the same specific appearance (someone on a horse) and specific motion (riding), within whatever context or background. Therefore, the region containing the horse is the eponymous region, and in general the horse should be at least part of the most salient region.

As a consequence of evolution, the human visual system has evolved towards an eclectic system, capable to recognize and analyze complex scenes in a fraction of a second. Therefore, much effort in computer vision research has been put on predicting human eye-gaze. Capturing fixation points and saccadic movements via eye-tracking [110, 127] allows for creating training data and analyzing spatial and temporal attention shifts. It is well known that humans are attracted by motion [80] or other human subjects, respectively their faces [82] if the resolution is good enough. Furthermore, human saliency maps are sparse and change if content is analyzed per image or embedded within a video [164]. Besides the drawback that a sufficient number of individuals have to observe the same image or video to obtain expressive saliency maps, above mentioned human preferences may even be misleading for general salient object detection tasks.

Thus, the defined goal of this chapter is to find eponymous and therefore salient video or image regions for guiding activity classification and detection, rather than predicting human eye-gaze. In contrast to estimating human gaze, these salient regions are not required to overlap with human fixation points but must identify the eponymous regions including the activity of interest. Within the proposed saliency estimation method, we enforce the Gestalt principle of figure-ground segregation, i.e. visually surrounded regions are more likely to be perceived as distinct objects. In contrast to previous approaches which globally enforce objects to be segregated from the image border,e.g., [191], no such assumption is enforced. Instead it finds visually segregated regions by a local search over several scales.

The contributions of this chapter are as follows. An encoding method is proposed for approximating the joint distribution of feature channels (color or motion) based on

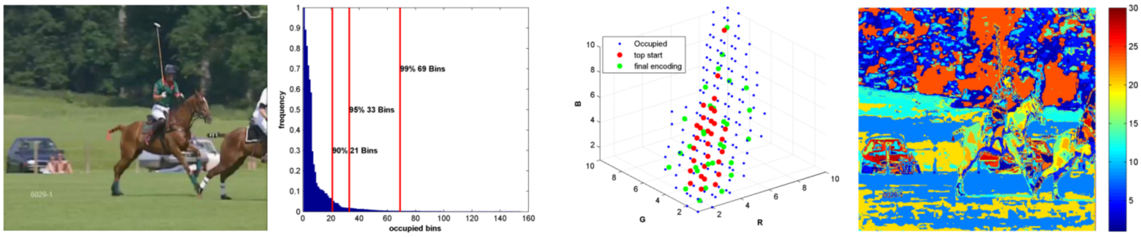


Figure 5.1: Encoding image content (from left to right): Input image. Occupancy distribution of bins within color cube. Only a small number of 33 bins contains 95% of the pixels. Occupied bins **o**, initial and final encoding vectors **E** within the color cube. Encoded image by assigning closest of 30 encoding vectors per pixel.

analyzing the image or video content, respectively. This efficient representation allows to scan images on several scales, estimating foreground distributions locally instead of relying on global statistics only. Finally, a measurement of salience quality is proposed that allows for dynamically weighting and combining the results of different maps, e.g., appearance and motion. Evaluating the proposed encoding based saliency estimation (EBS) on challenging activity videos and salient object detection tasks, benchmarking against a variety of state-of-the-art video and image saliency methods shows the superior performance of the proposed methods.

5.2 Related Work

Bottom-up vision based saliency has started with fixation prediction [72], and training models to estimate the eye fixation behavior of humans, either based on local patch or pixel information is still of interest today [164]. In contrast to using fixation maps as ground truth, [104] proposed a large dataset with bounding-box annotations of salient objects. By segmenting relevant objects within 1k images of this dataset [3], and later 10k images [31] elaborated the salient object detection task, see [18] for a review. Grouping image saliency approaches one sees methods working on local contrast [60, 104] or global statistics [3, 30, 31, 97]. Recently, segmentation based approaches [172, 180, 191] have emerged which often impose an object-center prior, i.e. the object must be segregated from image borders, mainly motivated by data-sets such as [3].

In contrast to salient object detection, video saliency or finding salient objects in videos is a rather unexplored field. Global motion saliency methods are based on analyzing spectral statistics of frequencies [58], the Fourier spectrum within a video [34] or color and motion statistics [190]. Local contrast between feature distributions is measured by [127], where independence between feature channels is assumed for simplifying the computations. In addition, motion information is described rather weak by magnitudes which causes problems with optical flow artifacts and global camera motion. Most recent, [64] developed an video saliency approach on evaluating the number of bits taken by an compression algorithm, like, e.g., H.264/AVC coding standard, followed by a MRF step for smoothing the block structure results according to image regions. The authors of [137] learned the transition between saliency maps of consecutive frames by detecting candidate regions created from analyzing motion magnitude, image saliency by [60], and high level cues like face detectors. Their combination of bottom up and top-down information has been applied mainly for eye-gaze prediction on specific datasets like, e.g., DIEM [110]. [194] over-segment the input video into color-coherent regions, and use several low level features for computing the feature contrast between regions. They show interesting results by sub-sampling salient parts from high-frame-rate videos and simple activity sequences. As a drawback they impose several priors in their feature computation, such as foreground estimation or center prior, which do not hold in more challenging videos with moving cameras, cluttered backgrounds, and low image quality.

Two different trends to incorporate have came up in the literature within the context of activity recognition and detection. Video classification is still the most common variant in activity classification. In fact, the volume of a video directly related to an activity is rather small. Previous research [92] has shown that this could bias the classification towards learning rather the activity related background than the activity itself, by global statistical models. Such approaches profit from unbalanced or too homogenous datasets. Finding activity related regions before creating video signature vectors, creates more robust and dataset independent representations and classifications. Recently, [157] motivated video saliency for foreground estimation to support cross dataset activity recognition and decrease the influence of background information. They adopted the image saliency method by [60] and aggregated color and motion gradients, followed by 3D MRF smoothing. Their research showed both an increase in classification accuracy and robustness of the representation when evaluated between different datasets containing the same activities. In [10], the authors computed three different saliency maps for key-point detection based *cornerness* value, and color information as well as motion magnitude with the method of [127]. During training a ranking of those salieny maps is learned, allowing for selecting the most representative saliency cue for each activity individually. Most recent [48] published a saliency weighted creation of feature vectors, similar to [10], but instead of optical flow applied directional motion energy for measuring video saliency followed by a super-pixel regularization. The second direction follows the idea of classifying videos solely by detected activity regions or volumes, respectively. This should avoid any background bias, also results by [146] showed that for some unbalanced datasets background regions could be learned as the most discriminative ones, and allows for classifying several activities happens within an video simultaneously. Within [108], the authors captured eye-gaze tracking data for the UCF Sports activity recognition dataset and showed that spatio-temporal key-point detections differ from human fixations. Later, [145] utilized this human gaze data for weakly supervised training of an activity detector and saliency predictor. Although applying eye-gaze tracking information should not be termed un- or weak-supervised, as it rather takes a lot human interaction to create such data, their research creates valuable insights.

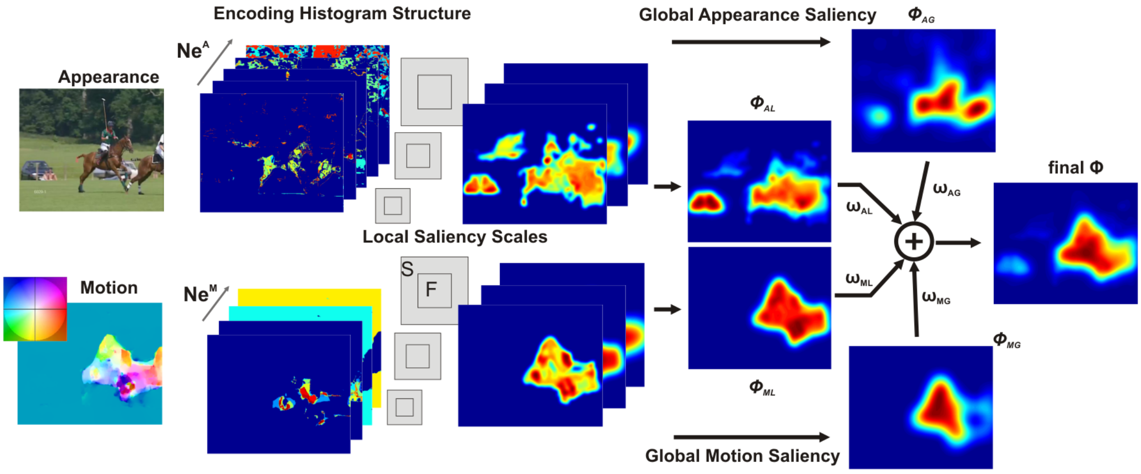


Figure 5.2: Overview of the proposed approach (from left to right): Input data for appearance and motion. Individual data dependent encoding for each feature cue. Estimation of local saliency on several scales by foreground and surrounding patches. L_∞ normalized saliency maps Φ_i and weighted combination according to reliability of individual saliency maps.

Summarizing the bottom-up video saliency methods one may see adaptions from visual saliency methods, that incorporate motion information by rather simple means like magnitude or gradient values. In contrast, in the proposed approach the joint distribution of appearance or motion features is modeled, which yields favorable performance. Moreover, our approach requires neither training data nor human eye-gaze ground truth as opposed to pre-trained methods, such as [137, 145].

5.3 Encoding Based Saliency

In the following, the individual steps of the proposed encoding based saliency (EBS) described and adaption for image and video saliency methods are discussed.

5.3.1 A Bayesian Saliency Formulation

Following the Gestalt principle for figure-ground segregation, we are searching for surrounded regions as they are more likely to be perceived as salient areas [121]. In other words, we analyze the contrast between the distribution of an image region (e.g., rectangle) with its surrounding border. Similar to [106, 127], we first define a Bayesian salience measurement. To distinguish salient foreground pixels $\mathbf{x} \in \mathcal{F}$ from surrounding background pixels, lets employ a histogram based Bayes classifier on the input image I. Therefore, let $H_\Omega(b)$ denote the b-th bin of the non-normalized histogram H computed over the region $\Omega \in I$. Furthermore, let $b_\mathbf{x}$ denote the bin b assigned to the color components of $I(\mathbf{x})$. Given a rectangular object region F and its surrounding region S (see Figure 5.2), we apply Bayes rule to obtain the foreground likelihood at location $\mathbf{x}$ as

$$P(\mathbf{x}\in\mathcal{F}|F,S,b_{\mathbf{x}}) \approx \frac{P(b_{\mathbf{x}}|\mathbf{x}\in F)\,P(\mathbf{x}\in F)}{\sum\limits_{\Omega\in\{F,S\}} P(b_{\mathbf{x}}|\mathbf{x}\in\Omega)\,P(\mathbf{x}\in\Omega)}. \tag{5.1}$$

In particular, we estimate the likelihood terms directly from color histograms, *i.e.* $P(b_{\mathbf{x}}|\mathbf{x}\in F) \approx H_F(b_{\mathbf{x}})/|F|$ and $P(b_{\mathbf{x}}|\mathbf{x}\in S) \approx H_S(b_{\mathbf{x}})/|S|$, where $|\cdot|$ denotes the cardinality. Furthermore, the prior probability can be approximated as $P(\mathbf{x}\in F) \approx |F|/(|F|+|S|)$. Then, Eq. (5.1) simplifies to

$$P(\mathbf{x}\in\mathcal{F}|F,S,b_{\mathbf{x}}) = \begin{cases} \frac{H_F(b_{\mathbf{x}})}{H_F(b_{\mathbf{x}})+H_S(b_{\mathbf{x}})} & \text{if } I(\mathbf{x})\in I(F\cup S) \\ 0.5 & \text{otherwise,} \end{cases} \tag{5.2}$$

where unseen pixel values are assigned the maximum entropy prior of 0.5. This discriminative model already allows to distinguish foreground and background pixels locally. However, modeling the joint distribution of color values, represented by 10 bins per channel, within a histogram based representation as described above, would lead to 10^3 dimensional features for solely describing color information. Assuming independence between channels as in [127] would simplify the problem to 3×10 dimensions and allow for efficient structures (e.g., integral histograms), but information is lost. Therefore, an efficient approximation by lower dimensional joint distributions using encoding vectors is proposed.

5.3.2 Estimating Joint Distributions via Encoding

Analyzing the content of image or video frames yields in general an exponential distribution of occupied bins as shown in Figure 5.1. The majority of image content is represented by a small number of occupied bins within a $10 \times 10 \times 10$ color cube representing the joint distribution, namely 33 bins account for 95% of the data in this example, while overall only 150 of 1000 possible bins are occupied (blue dots). Taking only the bins covering for 95% (red dots) has two major weaknesses. First, their spatial distribution is not efficiently covering all occupied volume within the color cube, leading to approximation artifacts. And second, the threshold for 95% may increase the number of taken bins to more than 80 as stated in [30], limiting the applicability for efficient sliding window computations.

Instead, let us represent the image content by a fixed number of encoding vectors. Let $\mathbf{O} \in \mathbb{R}^{o\times d}$ represent all occupied bins and $\mathbf{E} \in \mathbb{R}^{Ne\times d}$ the set of Ne encoding vectors where $Ne \leq |\mathbf{O}|$. We initialize $\mathbf{E}$ with the Ne most occupied bins (*i.e.* red dots in Figure 5.1) and perform *kmeans* clustering to optimize for the spatial distribution of encoding vectors as

$$\arg\min_{\mathbf{E}} \sum_{i=1}^{Ne} \sum_{o\in E(i)} \|\mathbf{o}-\mathbf{e}_i\|^2\,, \tag{5.3}$$

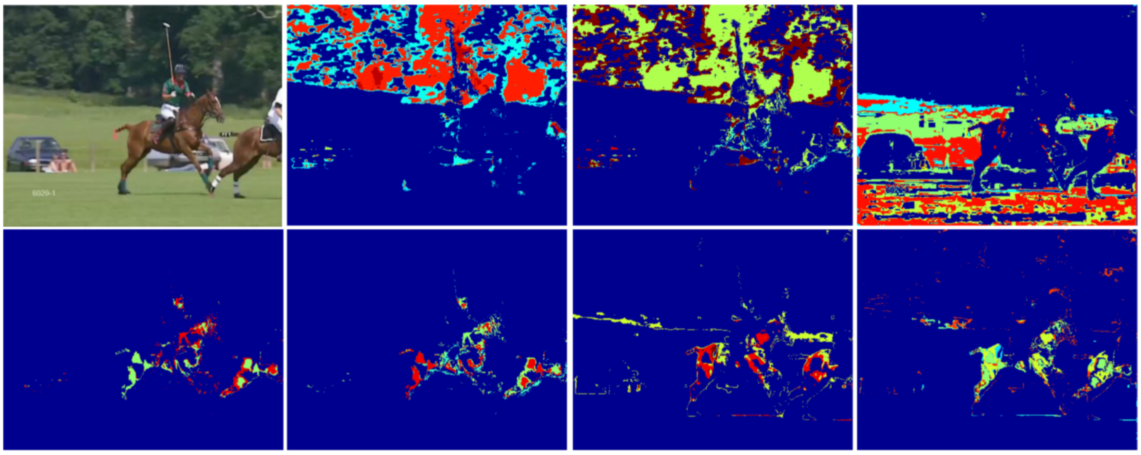


Figure 5.3: EBS encoding results for a color image belonging to activity horse riding of the UCF Sports dataset. Input image is depicted top left, while top rows shows per pixel weights for encoding vectors representing background regions (warm colors indicate high weights). Obviously a small amount of encoding vectors is sufficient to represent the complete background. Bottom row contains encoding weights for activity related encoding vectors.

where $E(i)$ denotes the set of bins $\mathbf{o}$ clustered to the encoding vector $\mathbf{e}_i$. The number of encoding vectors is set to the number of occupied bins covering 95% image pixels if this number is smaller than a maximum Ne. The final encoding vectors $\mathbf{E}$, visualized with green dots, and the resulting encoded image with $Ne = 30$ are shown in Figure 5.1. Homogenous regions are encoded by a small number of codes, while detailed structures are preserved. Please note that the final encoding vectors are not required to correspond to bins in the color cube. For further relaxing the hard binning decisions of color histograms, a weighted encoding over the nearest encoding vectors of each element in $\mathbf{O}$ is performed. When creating the integral histogram structure H (for simplicity we use the same notation as for histograms in Section 5.3.1), the entry for the k-th bin at pixel position $\mathbf{x}$ is computed by

$$H\left(\mathbf{x}, k\right) = \begin{cases} 1 - \frac{\|o(\mathbf{x}) - \mathbf{e}_k\|^2}{\sum\limits_{j \in \mathcal{N}(o(\mathbf{x}), \mathbf{E})} \|o(\mathbf{x}) - \mathbf{e}_j\|^2} & \text{if } k \in \mathcal{N}(o(\mathbf{x}), \mathbf{E}) \\ 0 & \text{otherwise,} \end{cases} \tag{5.4}$$

where $H \in \mathbb{R}^{h \times w \times Ne}$ and $o(\mathbf{x})$ defines the occupied bin $I(\mathbf{x})$ belongs to. The set of j encoding vectors nearest to $o(\mathbf{x})$ is given by $\mathcal{N}(o(\mathbf{x}), \mathbf{E})$.

Compared to other saliency approaches based on segmenting or clustering images, our overall process is very efficient as number and dimensionality of vectors in $\mathbf{O}$ is relatively small (in general around 200 occupied bins have to be considered) compared to pixels per image (above 200k), and converges in a fraction of a second. In addition, all operations for mapping $I(\mathbf{x})$ to $H(\mathbf{x})$ can be efficiently performed using lookup-tables. The result of such soft-encoded histogram structures is visualized in Figure 5.2. Next, we discuss how to enforce the Gestalt principle of figure-ground segregation on local and global scales.

5.3.3 Saliency Map Computation

Once the integral histogram structure of encoding vectors is created as described above, one can efficiently compute the local foreground saliency likelihood Φ for each pixel by applying Equ. 5.2 in a sliding window over the image. Therefore, the inner region F of size $[\sigma_i \times \sigma_i]$ is surrounded by the $[2\sigma_i \times 2\sigma_i]$ region S. Then, the following processing steps are performed on each scale σ_i.

First, the algorithm iterates over the image with a step size of $\frac{\sigma_i}{4}$ to ensure that the foreground likelihood for each pixel is estimated against different local neighboring constellations. Within each calculation, the foreground likelihood values of all pixels inside F are set. The final likelihood value for $\Phi_i(\mathbf{x})$ is obtained as the maximum value over all neighborhood constellations. Second, following our original motivation by Gestalt theory, the foreground map for scale i should contain highlighted areas for salient regions of size σ_i or smaller. In contrast, a region significantly larger than σ_i would have likelihood values $\Phi_i(\mathbf{x}) \leq 0.5$ for $\mathbf{x} \in F$. Therefore, this figure-ground segregation can be easily controlled after computing the foreground likelihood map by applying a box filter of size $[\sigma_i \times \sigma_i]$, and setting $\Phi_i(\mathbf{x})$ to zero if the average foreground likelihood $\bar{\Phi}_i(\mathbf{x}) \leq 0.5$. Finally, local foreground maps $\Phi_i(\mathbf{x})$ are filtered by a Gaussian with kernel width $\frac{\sigma_i}{4}$. The local foreground maps of individual scales are linearly combined to one local foreground saliency map Φ_L, which is L_∞ normalized.

Besides these locally computed foreground maps, global estimation of salient parts can offer valuable information. In particular, empirically observations showed that videos or images with global camera motion or homogenous background regions benefit from such global information. For computing the global foreground likelihood map Φ_G, we set S to the image border (typically 10–20% of the image dimensions) and F is the non-overlapping center part of the image. The resulting foreground saliency map Φ_G is Gaussian filtered and L_∞ normalized.

5.3.4 Processing of Motion Information

Studying related approaches for video saliency, the incorporation of optical flow information is in general treated with less care than appearance information. Measurements like pure flow magnitude [127, 137], motion gradients [157] or simple attributes like velocity, acceleration or average motion [194] are treated independently respectively without motion orientation information. However, considering the pseudo-color optical flow representation in Figure 5.2, one can directly observe that magnitude or simple attributes are prone to fail as large global camera motion is present, and motion gradients will create a noisy response. On the other hand, we observe a very discriminative visual representation of the scene context, which motivated us to have a closer look on the creation of such pseudo-color representations for optical flow. Following [140], the motion components for horizontal and vertical directions given in $U(\mathbf{x})$ and $V(\mathbf{x})$ are mapped to a color wheel representing the transitions and relations between the *psychological primaries* red, yellow,

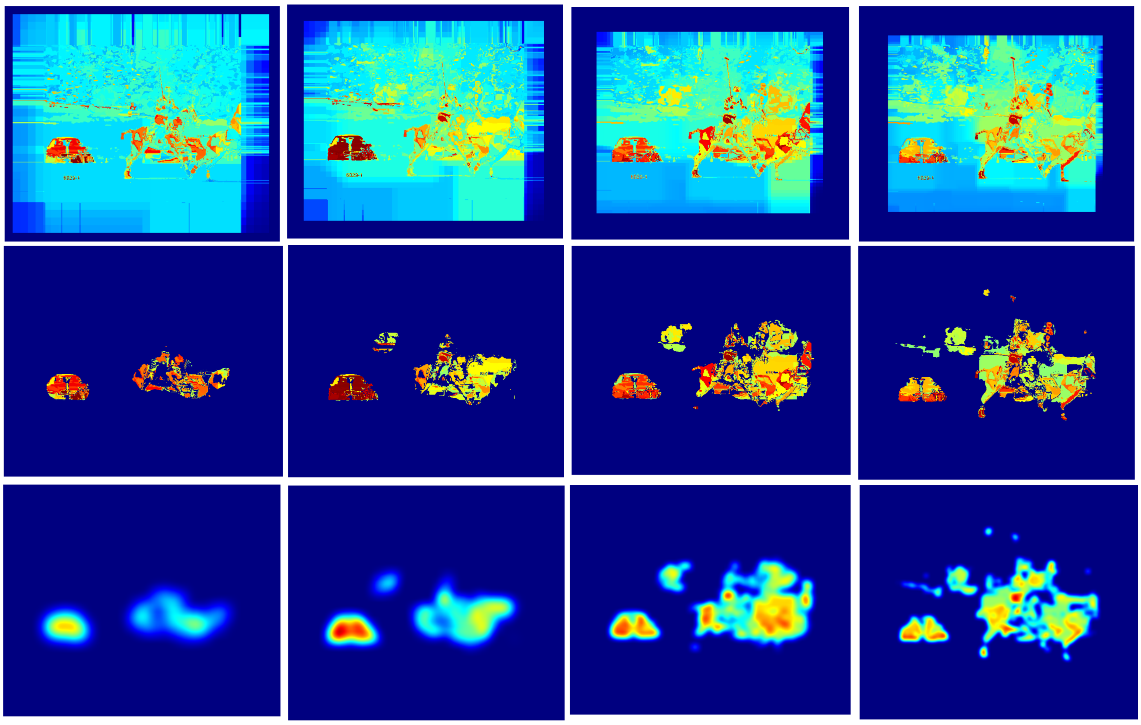


Figure 5.4: Creation of saliency maps within different scales for exemplar image from Figure 5.1. Top row: saliency map created with specific size of foreground and surrounding window. Middle: Saliency map after enforcing local Gestalt constrains by computing fill ratio as described in Section 5.3.3. Bottom: Final Gaussian smoothed saliency maps.

green, and blue. The color wheel, also known as *Munsell color system*, arranges colors such that opposite colors (at opposite ends of the spectrum, e.g., red and blue) are most distant to each other on the wheel. Similar to that, we want to represent opposite motion directions most distant to each other.

Therefore, the approach represented in Sections 5.3.2 and 5.3.3 is directly applied on the pseudo-color motion representation. For that let us compute the magnitude $M(\mathbf{x})$ and orientation $\Theta(\mathbf{x})$ of $\hat{U}(\mathbf{x})$ and $\hat{V}(\mathbf{x})$, which are the optical flow components normalized by the maximum magnitude of the corresponding frame. The orientation $\Theta(\mathbf{x})$ defines the hue value in the color wheel, while saturation is controlled by $M(\mathbf{x})$. Applying precomputed color wheel look-up tables, one can directly generate a three dimensional pseudo-color image taken as input for our motion saliency pipeline. Similar to the appearance likelihood maps Φ_{AL} and Φ_{AG} this yields the motion-based local Φ_{ML} and global Φ_{MG} likelihood maps. Although relatively simple, experimental evaluations show the beneficial behavior of this motion representation compared to related approaches discussed at the beginning of this section.

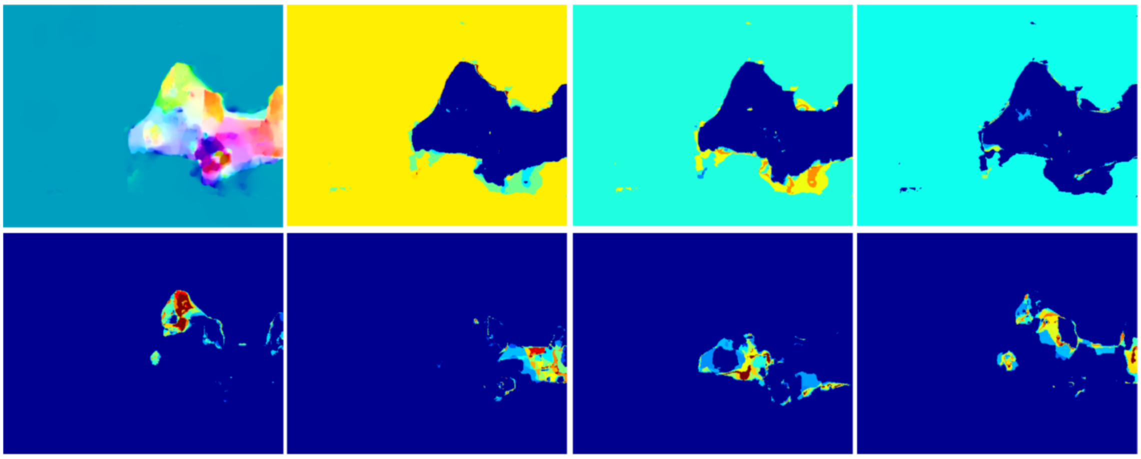


Figure 5.5: EBS encoding results for optical flow corresponding to the image within Figure 5.3. Flow field is depicted top left, while top rows shows per pixel weights for encoding vectors representing background motion (warm colors indicate higher weights). Obviously the majority is assigned to one encoding vector, which is sufficient to represent the camera motion. Bottom row contains encoding weights for object motion. This example nicely illustrates how EBS scales on more complex regions within a image.

5.3.5 Adaptive Saliency Combination

Given the above described steps, up to four foreground maps are generated for local and global estimation of appearance (i.e. Φ_{AL} and Φ_{AG}) and motion (i.e. Φ_{ML} and Φ_{MG}) saliency. Combining these in a proper way can compensate for failures of individual maps. Previous works either directly merged cues [157] or performed coarse global measurements like pseudo-invariance [190] without incorporating the spatial distribution of maps. In contrast, the uncertainty within our individual saliency maps is approximated by computing weighted covariance matrices of each map. A weighted covariance for saliency map Φ_j is given as

$$\Sigma_j = \begin{bmatrix} \frac{\sum\limits_{x,y\in I} \Phi_j(x,y)(\bar{x}-\bar{\mu}_x)}{\sum\limits_{x,y\in I} \Phi_j(x,y)} & \frac{\sum\limits_{x,y\in I} \Phi_j(x,y)(\bar{x}-\bar{\mu}_x)(\bar{y}-\bar{\mu}_y)}{\sum\limits_{x,y\in I} \Phi_j(x,y)} \\ \frac{\sum\limits_{x,y\in I} \Phi_j(x,y)(\bar{x}-\bar{\mu}_x)(\bar{y}-\bar{\mu}_y)}{\sum\limits_{x,y\in I} \Phi_j(x,y)} & \frac{\sum\limits_{x,y\in I} \Phi_j(x,y)(\bar{y}-\bar{\mu}_y)}{\sum\limits_{x,y\in I} \Phi_j(x,y)} \end{bmatrix}, \tag{5.5}$$

where $\bar{x}$, $\bar{y}$ denote normalized image coordinates to avoid bias for rectangular images and $\bar{\mu}_x, \bar{\mu}_y$ are the corresponding mean coordinates. Taking Σ_u as the baseline covariance of an unweighted uniform distribution, the reliability or weighting score for map j is computed as

$$\omega_j = 1 - \frac{det\left(\Sigma_j\right)}{det\left(\Sigma_u\right)} \quad \text{where} \quad \sum\nolimits_j \omega_j = 1. \tag{5.6}$$

Then, the final saliency map can be directly obtained by $\Phi = \sum_j \omega_j \Phi_j$. In the following, we denote our encoding based saliency approach EBS for linear combination of local saliency maps. In contrast, EBSL uses the proposed weighted combination of solely local and EBSG the weighted combination of all available (local & global) likelihood maps.

5.3.6 Saliency based Segmentation

While the general saliency approach presented above creates a per-pixel saliency sore, it may be required to segment objects from the background. In addition, precise salience predictions score higher on evaluations with segmented ground truth objects. Previous approaches have incorporated separated segmentation approaches like Mean-shift [51] or super-pixels [48], for propagating high saliency values within local connected segments. An advantage of the proposed encoding based saliency method lies in the indirect over-segmentation of the image, without defining any parameters on average segment size and such adopting nicely to the image content.

Taking a closer look on Figure 5.3, we can see larger regions for background regions, while foreground regions consist of several smaller, and therefore detail preserving, segments in different encoding layers. Each encoding layer describes the per pixel weights for a specific encoding vector. After computing the EBSG saliency values per pixel, as described in previous sections, we compute connected components in each encoding layer, to define segments. Moreover, each segment has a weight according to the significance the corresponding encoding vector contributes to the representation of this, while each pixel may be part of j different regions in j encoding layers, see Section 5.3.2 for more details. For each connected component in each layer we analyze the saliency values of the underlying pixels, and propagate high saliency values within components while suppressing components with low saliency values. Finally, the region based saliency values of each layer are again combined into one final EBSGR saliency map. Exemplar results and a comparison between EBSG and EBSGR can be seen in Figures 5.21 and 5.22.

5.4 Experiments

In the following, various experiments are performed for both video saliency and object saliency tasks. First, the favorable performance of our approach for challenging video saliency tasks using the Weizmann [57], UCF Sports [133] and Olympic Sports [116] activity datasets is demonstrated. Second, a comparison of EBS to related saliency approaches and evaluation of the influence of parameter settings is performed, analyzing the effect of prior image information as discussed in [99], on the widely used ASD [3] and MSRA10k [31] salient object segmentation datasets.

As ground truth annotations are given in different formats (*i.e.* coarse bounding boxes, detailed binary segmentation or eye-fixation maps), we apply the following metrics correspondingly. If ground truth segmentation is available, we compute precision/recall values as well as the area under curve (AUC) by varying thresholds to binarize saliency maps, and measure the overlap with the ground truth segmentation. For experiments where solely bounding box annotations are available we add spanning bounding boxes to the binarized saliency maps before computing the scores denoted AUC-box (please see Section 5.4.1 for more details). For given eye-gaze ground truth data we measure the exactness of the saliency maps by computing the normalized cross correlation (NCC). For all benchmark comparisons we use code or precomputed results published by the corresponding authors, except for [157] which we reimplemented according to the paper (without 3D MRF smoothing which could be optionally applied to all methods).

5.4.1 Evaluation Metrics

Within the experimental section several metrics are applied to compare different saliency detection methods. As ground truth annotations are given in different formats i.e. coarse bounding boxes, detailed binary segmentation or eye-fixation maps, we apply the following metrics correspondingly. If ground truth segmentation is available, we compute precision/recall values as well as the area under curve (AUC) by varying thresholds to obtain binarized saliency maps and measure the overlap with the ground truth segmentation. For experiments where solely bounding box annotations are available we add spanning bounding boxes to the binarized saliency map before computing the scores (denoted AUC-box, see Figure 5.6). For given eye-gaze ground truth data we measure the exactness of the saliency maps by computing the normalized cross correlation (NCC).

AUC For evaluation on segmented object ground truth, we compute true-positives (TP), false-positives (FP), true-negatives (TN) and false-negatives (FN) for each threshold image (see Figure 5.6, second row). The TP is the number of pixels with saliency values greater than a threshold and overlapping with the ground truth annotation. In contrast, FN are all pixels within the groundtruth region < threshold. By varying this threshold one can compute recall-precision curves by

$$precision = \frac{TP}{TP + FP} \tag{5.7}$$

and

$$recall = \frac{TP}{TP + FN}. \tag{5.8}$$

The area under this recall precision curve is denoted as AUC. Examples in Figure 5.6 show, that although the saliency maps of different algorithms align nicely with the object, the bounding box annotation causes many false-negatives, see [18] for details and discussion. Therefore, we defined AUC-Box as an additional metric as follows.

AUC-Box Filling the binary saliency maps with spanning bounding boxes before computing TP, FP, TN and FN compensates for coarse annotation (see Figure 5.6, third row).

We denote this measure AUC-box within the following sections. Given both scores for the UCF sports dataset, the reader may extract additional information about performance and robustness of methods.

NCC If non-binary ground truth information is given, as in form of eye-gaze tracking data, we apply normalized cross correlation as a measurement of performance. Eye-gaze data is generally given as a set of sparse local points of fixations or saccadic motions. As defined by the collectors of the data [108], we apply Gaussian blur for each gaze measurement to compensate measurement errors and create smooth eye fixation ground truth maps.

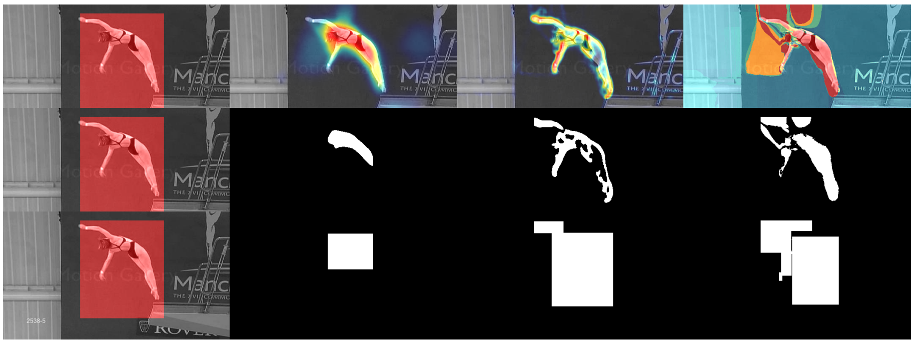
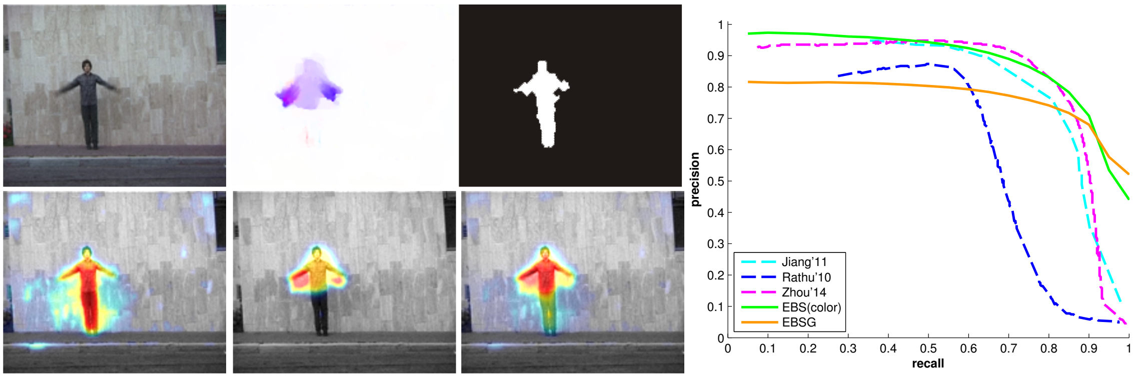


Figure 5.6: Comparison of AUC and AUC-box computation. Left column shows the ground truth bounding box annotation. Top row depicts results from individual saliency detection methods (f.l.t.r) EBSG, [157] and [127]. Second row shows true-positives (TP) pixels for different thresholds. Especially non-segmentation results like in column three [157] have many false-negatives (FN), due to the coarse bounding box annotation, and hence low recall and AUC values. Applying spanning bounding-boxes around sub-segments, as shown in row three, compensates for the coarse segmentation, allowing for a fairer comparison.

5.4.2 Saliency for Activity Localization in Videos

Recent research emphasized the benefits of video saliency for two main tasks within activity recognition, namely weakly supervised training of activity detectors [145, 146], and saliency weighted descriptors for superior video classification results [48, 157]. Both fields of application depend on robustly highlighted foreground regions, marking the activity related areas. Therefore, the proposed EBS method is evaluated on different activity recognition dataset with varying characteristics, with the goal of estimating the human ground truth activity annotations. In addition, comparison against a variety of state-of-the-art saliency methods, ground-truth eye-gaze data [108] and generic object detectors [7] allows a critical evaluation about advantages and drawbacks of such approaches.

Weizmann. We follow the recent evaluation of image and video saliency methods by [194] on the Weizmann activity dataset [57] and compare our proposed EBS method to the top-performing methods [77, 127, 194] of that study. The dataset contains videos of ten activities performed by 9 actors captured with object-centered static cameras in front of a homogenous background, as further described in Chapter 2. This simplifies the video saliency estimation to a foreground estimation problem. In fact, results in [194] have shown the superior performance of solely color-based methods, while video saliency approaches (which include motion information) performed worse. Figure 5.7 shows this bias of the evaluation strategy. Although our weighted saliency approach (EBSG) yields visually plausible results for the hand waving activity, the segmentation ground truth prefers fully segmented objects. In particular, considering purely local activity (e.g., hand waving), the motion-based saliency focuses on such active regions, which results in reduced performance scores on the binary ground truth masks, which cover the whole person. However, our solely color-based approach EBS(color) shows competitive results in comparison to the top-performing methods. As a result of this experiments, the following evaluations focus on more realistic videos and separately evaluate the influence of camera and background motions in comparison to static scenes.



(a) (b)

Figure 5.7: Results for video saliency on the Weizmann dataset. (a) Top row: appearance, motion, and ground truth segmentation information. Bottom row: saliency results for taking color, motion or combining both cues by our proposed weighting scheme within our EBS method. Average recall precision curves are shown in (b). Our EBS method taking solely color information performs favorable, as full body segmentation ground truth does not favor activity related motion information.

UCF Sports dataset. Recent evaluation of video saliency methods by [194] on the Weizmann activity dataset [57] has shown the surprisingly well performance of solely color-based methods. For completeness, we compared against them in the previous experiment but based on our findings, as depicted within Figure 5.7, we further evaluate on a more selective activity dataset, namely the UCF Sports dataset [133]. UCF Sports is a collection of low-quality television broadcasts, containing 150 videos of various sports. This dataset depicts challenging scenarios including camera motion, cluttered backgrounds, and non-rigid object deformations. Furthermore, it provides ground truth bounding box annotations for all activities. In addition, [108] captured eye-gaze data from 16 subjects, which allows to compare saliency results with these human attention maps given as probability density functions (see Figure 5.18). This makes the dataset well suited for benchmarking our EBS approach with other video saliency methods. For comparison, we apply all top performing methods from [194] and additionally evaluate [157]. Furthermore, we use the objectness detector of [7], as previously applied for weakly supervised training of activity detectors on UCF Sports within [146]. The objectness detector [7] is trained on the PASCAL object detection benchmark dataset. We follow their parametrization and take the top 100 boxes returned by the objectness detector to create a max-normalized saliency map per frame. For completeness, we quote NCC scores from [145] for supervised eye-gaze estimation trained and evaluated via cross-validation on UCF Sports. Please note that all other saliency methods, state-of-the-art and the proposed EBS variants, are fully unsupervised and require no training.

For a distinct evaluation we split the videos into two sets, namely *static* and *dynamic*, where the first contains activities with less severe background clutter or camera motion like golfing, kicking, lifting, swinging, and walking. The second set consists of activities with strong camera motion, clutter, and deformable objects, such as diving, horse-riding, skating, swing on bar, and running. As can be seen from the resulting recall/precision curves in Figure 5.8, all methods perform better on the static videos than on the dynamic ones visualized in Figure 5.9. The most significant performance decrease between static and dynamic videos can be observed for [194] which is the top-performing methods on the simpler Weizmann dataset experiments. On the contrary, our EBS versions show almost no degradation when switching from simpler static to more complex dynamic scenes. Furthermore, one may notice a larger gap between using solely local saliency information with EBSL and incorporating additional global information within EBSG on the dynamic videos. This can be explained by our optical flow representation which acts as a kind of global motion compensation when computing the global motion saliency. In particular, our optial-flow processing performs favorably compared to [127, 157, 194] which rely on simple motion magnitude values. Overall, all compared methods benefit from the box prior when evaluating recall and precision, as it compensates for coarse annotations and supports *sparse* saliency maps as generated by [157, 194].

Olympic Sport dataset. For additional experiments on further videos with more diverse and severe backgrounds and camera motions, results on the Olympic Sports dataset [116] were performed. Activities within videos of the test set were annotated manually with bounding box information, as shown in Figure 5.20. The top performing algorithms from experiments above were evaluated against the proposed EBSG method. The parameter settings stayed unchanged from the UCF Sports experiments. Results are summarized within Figure 5.11. Although, the Olympic dataset has been quoted as more challenging compared to UCF Sports, surprisingly the work of [7], working without motion

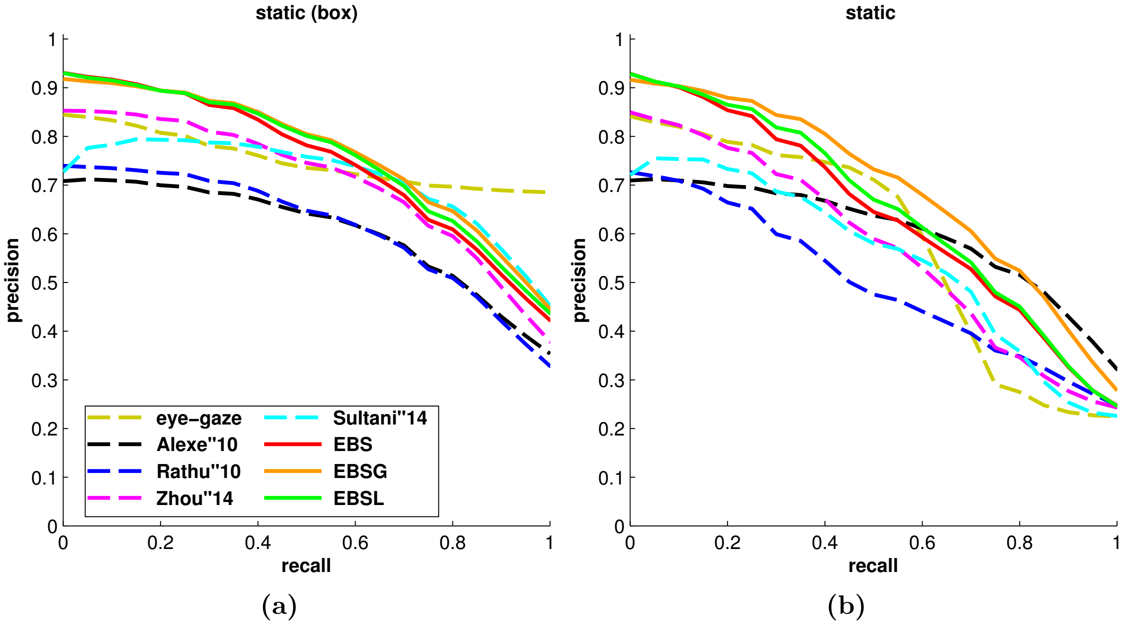


Figure 5.8: Average recall-precision plots of various saliency methods on static videos within the UCF-sports dataset. Results over static videos with (a) or without (b) box prior. See text for further discussion.

information, performs on par or better compared with more recent video saliency methods like, e.g., [127] or [157]. The proposed EBSG again performs superior and within a similar performance range as for the UCF Sports evaluation given in Figure 5.10.

A closer look within a detailed per activity evaluation, depicted in Table 5.1 reveals more insights into the dataset. The *objectness* detector by [7] score well on activities recorded within frontal viewpoints of upright standing persons performing an activity. Originally trained, e.g., on the PASCAL Visual Object Challenge dataset [42], human and man-made shaped structure score high within the algorithm. Furthermore, it returns bounding box hypotheses, which can be seen as an inherently box-prior described in Section 5.4.1. As a drawback, the *objectness* detector is often distracted by rectangular background structures highly contrasted with their local surrounding. The video saliency methods have severe problems with jittered optical flow [127], or highly textured backgrounds [157], as depicted in Figure 5.20.

In summary one may draw the following conclusions: The proposed EBSG performs better or on a par, compared to a variety of different image or video saliency methods for the task of activity localization in videos. Especially the robustness to noisy and jittered optical flow fields as well as cluttered backgrounds leads to more stable and robust results throughout all evaluated datasets. In addition, evaluations showed that pre-trained detectors do not fulfill the requirements and may produce incorrect hypotheses for entire activity classes. In the following, a comparison to eye-gaze ground truth gives further

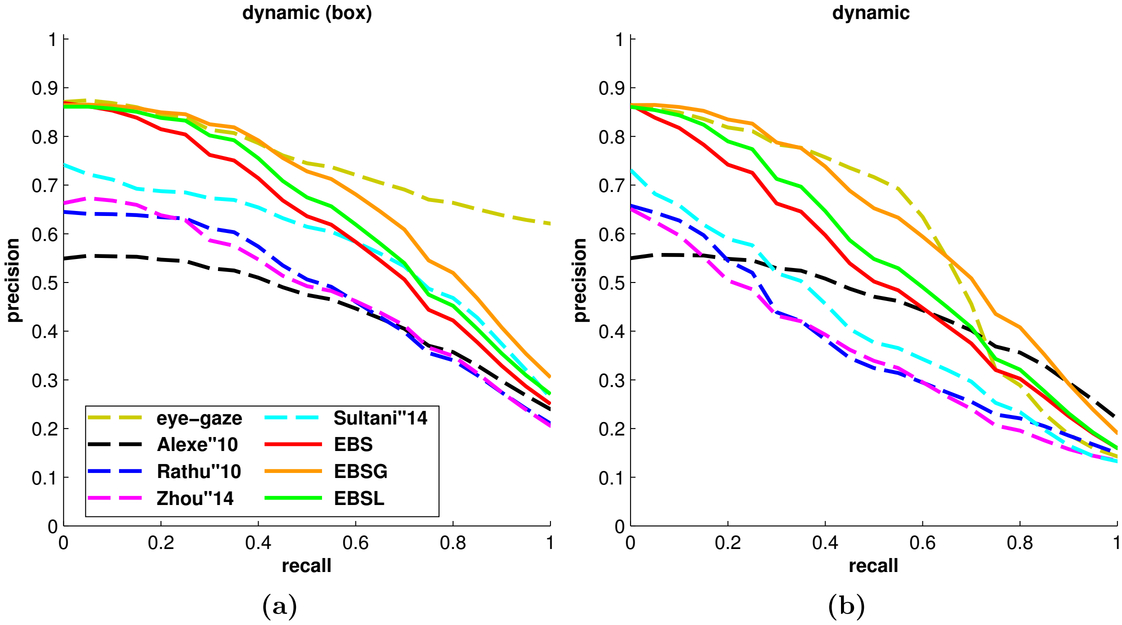


Figure 5.9: Average recall-precision plots of various saliency methods on dynamic videos within the UCF-sports dataset. Results over dynamic videos (a), (b). The dynamic subset contains much more challenging videos including moving cameras, cluttered background and non-rigid object deformations during actions. See text for further discussion.

insights how useful automatic video saliency is for weak supervised activity tasks.

5.4.3 Eye-gaze Prediction

One of the most notable points within the experiments within Section 5.4.2 is to see that human eye-gaze does not perform superior when evaluated against bounding box annotations, especially considering the simpler static videos. After having a closer look on the visual results it can be seen that human fixations are focused on faces if the image resolution is sufficiently high and the image context is less demanding as in Figure 5.18 and 5.19. On the other hand, for low resolution videos or rapidly changing actions the fixations are distributed over the whole person (see Figure 5.19). This also reflects the findings of [82], but questioning the general applicability of human eye-gaze as supervision for training activity detectors such as [145].

Table 5.2 summarizes the mean AUC and NCC results over all UCF Sports videos also shown in Figure 5.10. As can be seen, our EBSG methods performs favorably compared to other video saliency methods and on par with previously proposed supervised methods trained and tested on UCF Sports. The importance of individual cues within the EBS pipeline is depicted in Table 5.3. DJS depicts the results for directly modeling the joint distribution of color and motion channels for saliency estimation, as described

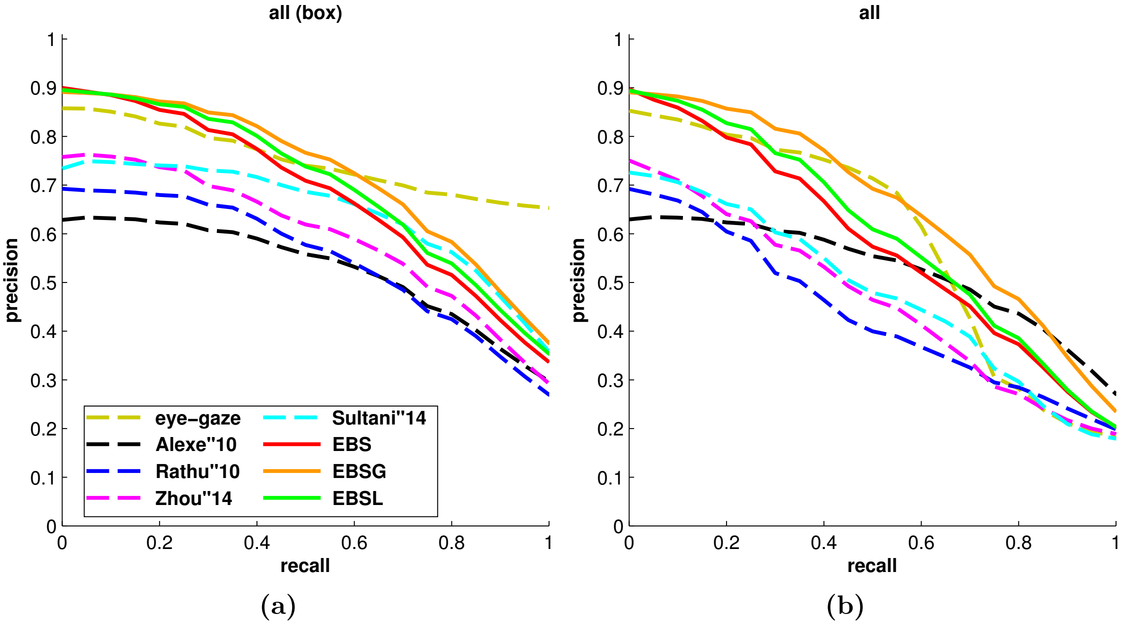
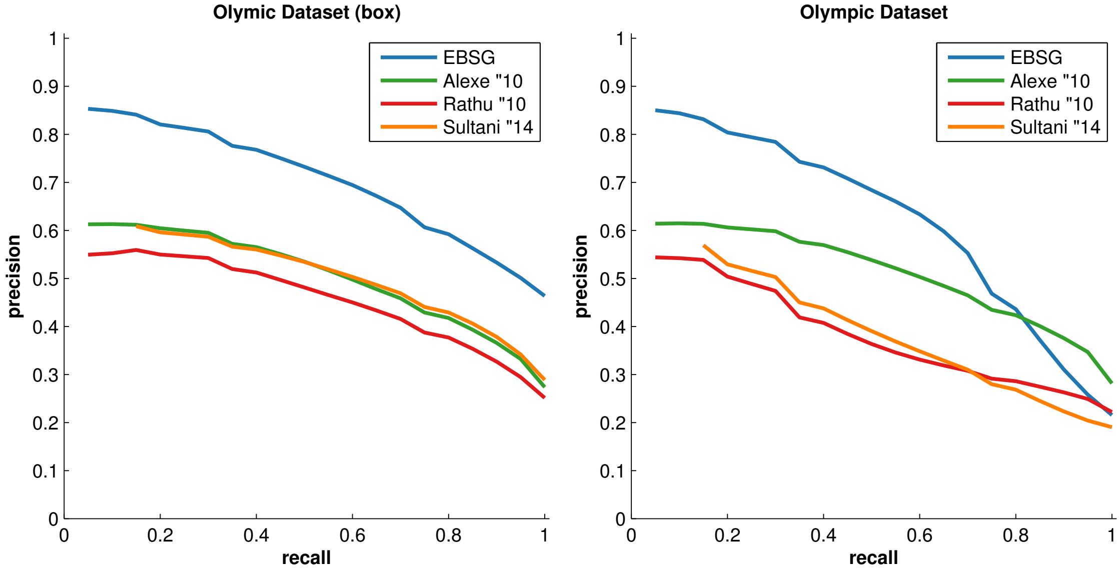


Figure 5.10: Average recall-precision plots of various saliency methods over all videos within the UCF-sports dataset. Results over all videos with (a) or without box-prior (b). The proposed EBSG method even creates higher precision values compared to results of the human eye-gaze ground truth.

in Section 5.3.1. As this incorporates a 1000 dimensional histogram when working with 10 bins per color channel, we cannot perform optimizations like integral histograms as described in Section 5.3.2, therefore leading to inferior run-times, while our MATLAB implementation of EBS is comparable to other bench-marked methods and still has potential for optimization. The difference between DJS and EBS is the loss of encoding up to several hundred color values per image with 30 or less encoding vectors. But this loss can be captured by our adaptive weighting of individual local saliency cues within EBSL and additional global saliency measures in EBSG.

5.4.4 Improved Activity Classification by Saliency Weighted Features

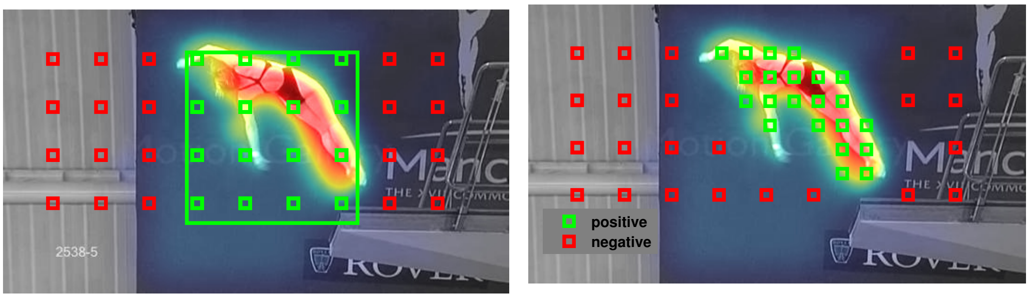
Besides the positive performance on general activity localization evaluated so far, the open question is if and how saliency maps can be applied to increase activity classification and support training of activity detectors. Especially detector training needs a huge effort of labeling time and, as been shown by [10] and [157], video classification can benefit from separating foreground and background information or weighting features by the underlying saliency values. To evaluate the performance gain in both cases, we perform experiments on the UCF Sports dataset following a train-test split similar to [92]. The data is split into non- overlapping train and test sets and a 3-fold cross validation is performed. This



(a) (b)

Figure 5.11: Average recall-precision plots of various saliency methods over all videos within the Olympic dataset. Results over all videos with (a) or without (b) bounding box prior for evaluation. The pre-trained *objectness* detector by Alexe et al. [7] performs surprisingly well (see text for further discussion).

has been shown to be more expressive than a classical leave-one-out test [92].



(a) (b)

Figure 5.12: Left: Classical grid sampling for feature extraction in the video classification experiment. According to bounding box annotations, green marks represents locations of positive training samples for the detector training experiment and red marks negatives, respectively. Right: Sampling additional features on high salient positions for the classification tasks, and positive and negative detector training samples according to underlying saliency information.

Classification Results Following the global bag-of-words (bow) approach, features are sampled in a fix grid as depicted in Figure 5.12a. For evaluating the benefit of saliency

	AUC-box				AUC			
	EBSG	[7]	[127]	[157]	**EBSG**	[7]	[127]	[157]
basketball layup	**0.72**	0.36	0.38	0.58	**0.66**	0.36	0.28	0.41
bowling	**0.65**	0.48	0.51	0.45	**0.58**	0.48	0.46	0.35
clean and jerk	0.81	**0.83**	0.67	0.60	0.73	**0.81**	0.61	0.49
discus throw	**0.68**	0.41	0.41	0.50	**0.57**	0.42	0.34	0.37
diving platform 10m	**0.49**	0.21	0.20	0.26	**0.44**	0.26	0.17	0.19
diving springboard 3m	**0.41**	0.05	0.16	0.28	**0.36**	0.06	0.12	0.20
hammer throw	**0.79**	0.68	0.56	0.65	**0.66**	0.63	0.43	0.44
high jump	**0.66**	0.40	0.38	0.43	**0.60**	0.42	0.29	0.34
javelin throw	**0.82**	0.47	0.45	0.57	**0.73**	0.48	0.37	0.42
long jump	**0.78**	0.43	0.30	0.47	**0.69**	0.43	0.23	0.36
pole vault	**0.65**	0.30	0.31	0.38	**0.57**	0.31	0.26	0.29
shot put	**0.66**	0.60	0.48	0.55	0.57	**0.62**	0.38	0.41
snatch	**0.78**	0.72	0.61	0.63	0.68	**0.70**	0.55	0.52
tennis serve	**0.70**	0.57	0.45	0.39	0.56	**0.57**	0.31	0.28
triple jump	**0.74**	0.38	0.36	0.47	**0.66**	0.38	0.27	0.36
vault	**0.71**	0.38	0.38	0.49	**0.63**	0.40	0.28	0.37

Table 5.1: Detailed AUC-box scores per activity classes for different saliency methods on Olympic dataset. The *objectness* detector results by [7] performs well on static videos with frontal viewpoints. In addition, bounding box hypotheses, which can be seen as an inherently box-prior described in Section 5.4.1, allows higher recall scores for same precision values within the AUC evaluation.

	Eye-gaze	[77]	[7]	[157]	[194]	[127]	EBSG	[145]	[108]
AUC	0.61	0.44	0.52	0.48	0.47	0.43	**0.66**	–	–
AUC-box	**0.77**	0.51	0.52	0.65	0.61	0.54	**0.73**	–	–
NCC	**1.00**	0.36	0.33	0.33	0.37	0.32	**0.47**	0.36*	0.46*

Table 5.2: Average AUC, AUC-box and NCC scores over all UCF Sports videos and comparison to state-of-the-art saliency estimation methods.* NCC scores for supervised methods trained on UCF sports published by [145]. EBSG methods performs favorable on both activity localization AUC scores and eye-gaze estimation NCC scores.

information for activity detection, an evaluation against human ground-truth annotation is performed. In the first setup we train according to bounding box information as shown in Figure 5.12a. Within the saliency training, the decision for positive and negative training samples is solely defined by the underlying saliency values. Again, positive training patches are densely sampled from highly salient regions, while negative patches are taken from patches with low saliency values. From the visualizations in Figure 5.12b one may recognize various benefits of saliency based sampling. According to the inaccurate bounding box annotation background patches may be trained as foreground, and in adaptation it allows for data dependent denser sampling. We extract HoG and HoF features and create a codebook of size 200 using fisher vectors [122]. The encoding is learned from 100k randomly

	DJS	EBS	EBSL	EBSG
AUC	0.64	0.58	0.60	**0.66**
AUC-box	**0.73**	0.68	0.70	**0.73**
NCC	0.43	**0.47**	0.45	**0.47**

Table 5.3: Average AUC, AUC-box and NCC scores of different EBS version on UCF Sports dataset.

selected samples. Based on those generated features, we train one-vs-all linear support vector machines using LIBSVM [27].

Overall classification performance gained an increase between $3-5\%$, depending on the activity. The improvement was not as high expected compared to the numbers proposed in related literature, but may be mainly influenced by different feature descriptor and tuning parameters. Such optimization have not been performed in more detail within this work, nevertheless cleaner activity localizations during training by the proposed EBS methods should influence related approaches in a positive way, especially in the case of activity detection.

5.4.5 Salient Object Detection in Images

Although comparison on activity data-sets within this chapter is comprehensive, a variety of saliency approaches exist for the domain of image segmentation. As this task is most similar to localizing activities in videos is salient object detection in still images, additional experiments are performed evaluating especially the influence of center-priors on the algorithms and the robustness of the generated saliency maps to non-centered objects.

Both tasks, image and video saliency detections, have the goal of finding eponymous regions in the data. Although the focus of this work is on saliency estimation for activity videos, EBS can easily be applied to standard image saliency tasks by switching off the motion components. In addition, image saliency datasets contain per-pixel segmentation ground truth annotations for a more detailed evaluation of the proposed methods. Moreover, it shows the robustness of the proposed EBS methods and allows for additional comparisons and evaluations.

Many models and data-sets have been proposed in the image domain (see, e.g., [18, 23] for a review). In particular, we use the ASD dataset [3], which comprises 1000 images from the MSRA saliency dataset with ground truth segmentation masks an has been widely used within the community. We benchmark against recent state-of-the-art approaches, such as FT [3], HFT [97], BMS [191], Hsal [180], GSGD & GSSP [172], and RC & HC [30]. Recently, the authors of [30] have published an improved version within $10k$ ground truth segmented MSRA saliency images named MSRA10k [31]. Results of state-of-the-art image saliency methods published in [31] are available online[1].

[1] http://mmcheng.net/msra10k/

Within all experiments we applied 7 local scales between $\sigma_i = [\frac{1}{10},\dots,\frac{1}{2}]\min(width, height)$ of each individual test image. Post-processing at each scale level is performed as described in Section 5.3.3. We fixed the number of bins per color channel to 10 and the maximum number of encoding vectors Ne to 40 within all experiments, as the average number of encoding vectors chosen by EBS lies below 30 for both RGB and CIE Lab (see Section 5.3.2). Finally, we evaluate the influence of taking RGB or CIE Lab color spaces. Further, we evaluate the benefit of joint modeling feature channel probabilities within our EBS compared to saliency estimation with independent color channel probabilities as previously done by, e.g., [127, 190]. Figure 5.13a shows that the increase of the number of maximally available encoding bins Ne from 30 to 60 does not increase results because, as mentioned above, the number of encoding vectors is set the number of occupied bins responsible for 95% if this number is smaller than maximum Ne.

Results does not show considerable differences between EBSG using RGB or Lab color channels. But we see a strong improvement from applying our methods on distributions following the independents assumption between channels, similar to [127] denoted as (*independent Rgb,Lab*), to our approximated joint distributions in EBSG.

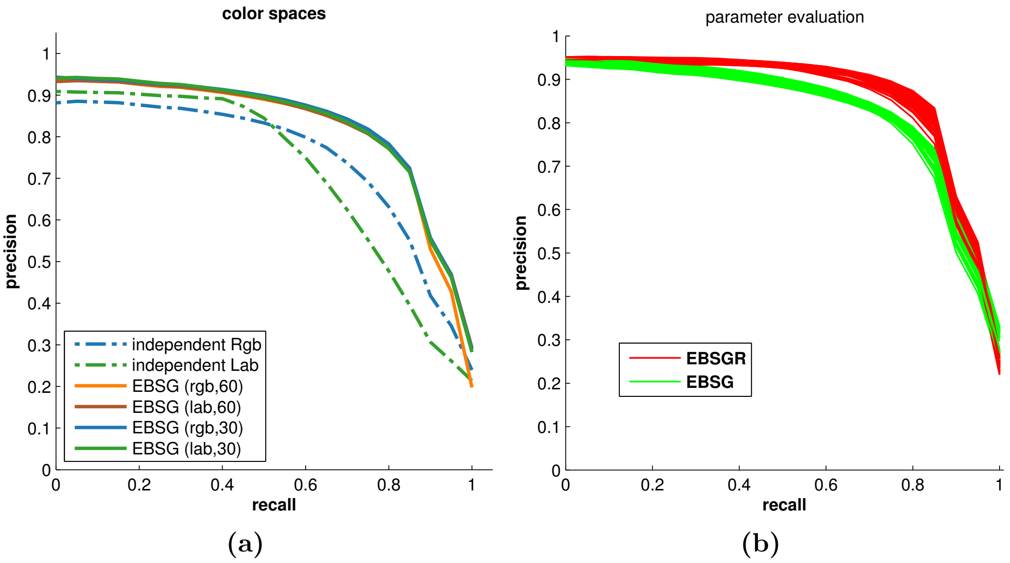
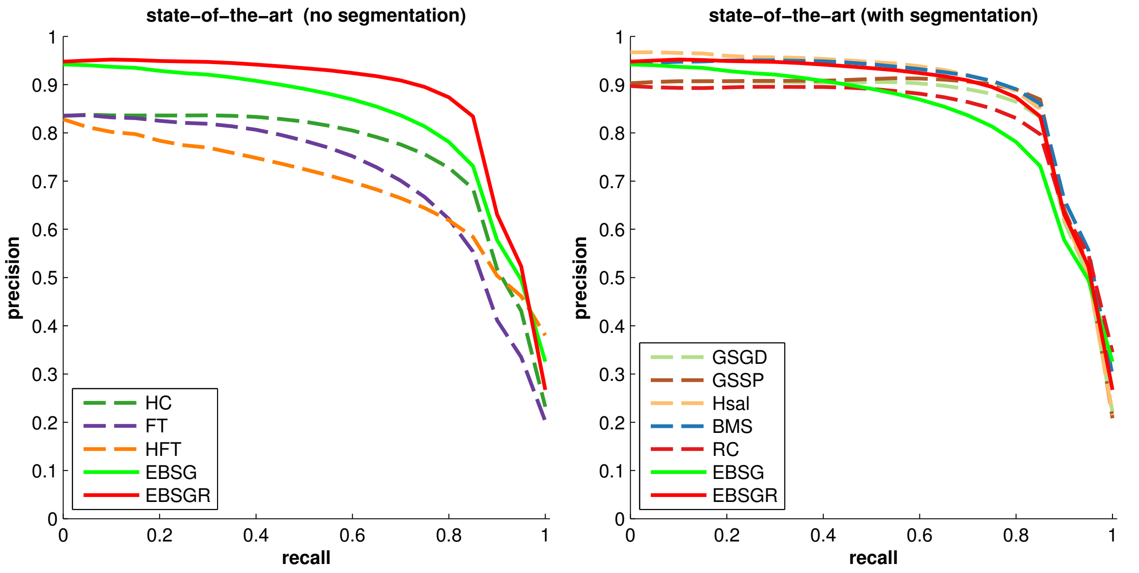


Figure 5.13: Evaluation of ESB parameter settings.(a) Importance of modeling the joint distribution of color channels within EBSG, instead of the independence assumption as previously done by, e.g., [127]. (b) Evaluating different parameter settings shows the robustness of our proposed approaches.

A comparison with the state-of-the-art in salient object segmentation is shown in Figure 5.14. EBSG performs better or equal than approaches without explicit segmentation steps, *i.e.* [3, 30, 97] in Figure 5.14a. The top-performers on the other hand, shown in Figure 5.14b enforce segmentation consistent results [180] or pose additional assumptions,



(a) (b)

Figure 5.14: Comparison of EBSG to state-of-the-art in salient object detection on the ASD dataset.(a) Evaluation against methods without segmentation or image center prior. (b) Methods applying pre- or post-segmentation or assuming salient regions to be centered within the image.

e.g., the object must not be connected to the image border [172, 191]. Both constraints are particularly beneficial for the ASD dataset, but questioned by the recent analysis in [99]. Therefore, we evaluate the impact of the latter object-center prior by cropping the ASD dataset images such that salient objects are touching the borders. As discussed previously in [99], salient object data-sets are biased towards centered objects without connection to the image border. All methods, performing favorable compared to our proposed EBS methods, exploit this circumstance. To evaluate robustness of methods if this assumption is violated, and to compare our EBSG against top performing BMS [191], we created two data-sets by cropping images of the ASD dataset such that salient objects are located near the borders. Two cropping levels are tested: First, salient objects touch the closest image border and second, intersect the closest border by 5 pixels as shown in Figure 5.17. Depicted in Figure 5.16, the robustness of BMS decreases drastically while EBSG stays almost constant within the first test and decreases slightly for severe *out of center* objects. Additional visual comparisons can be found in Figures 5.21 and 5.22.

5.5 Discussion of Results

A novel encoding based saliency detection method (EBS) inspired by Gestalt theory has been proposed within this chapter. Analyzing the image or video context respectively, we create encoding vectors to approximate the joint distribution of feature channels. This

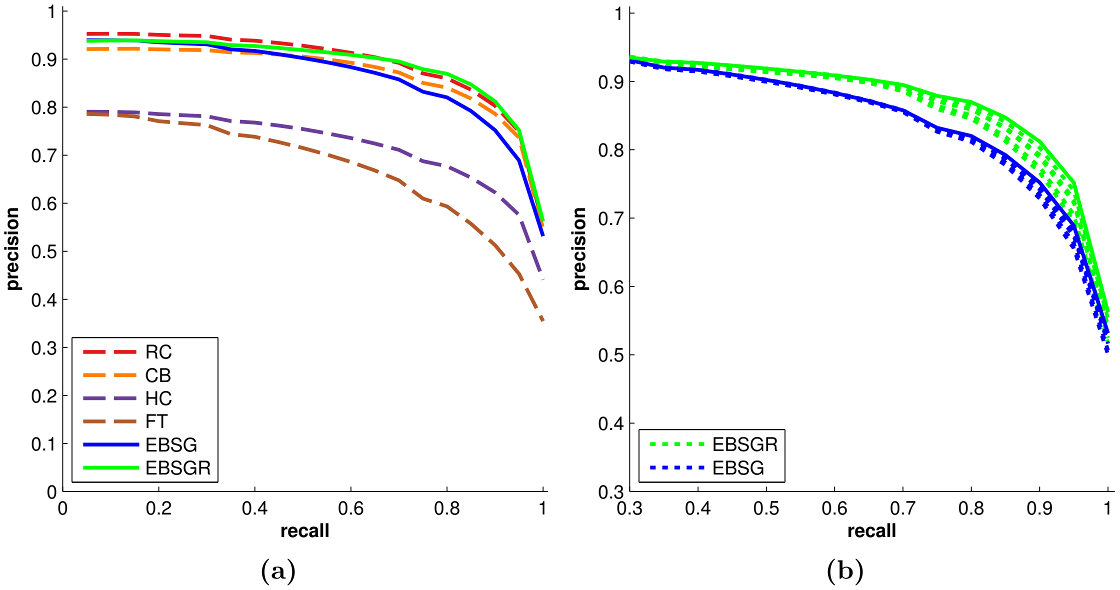


Figure 5.15: (a) Comparison of EBSG to state-of-the-art in salient object detection on the MSRA10k dataset. (b) Evaluating different parameter settings shows the robustness of our proposed approach (see text for more details).

low-dimensional representation allows for efficiently estimating local saliency scores. The advantage of approximating the joint distribution against previous approaches assuming independence between feature channels has been shown within the experiments. Implicitly enforcing figure-ground segregation on individual scales allows us to preserve salient regions of various sizes. In addition, the robust reliability measurement allows for dynamically merging individual saliency maps, leading to excellent results on challenging video sequences with cluttered background and camera motion, as well as salient object detection in images. Besides those positive findings, results of the evaluations questioning the applicability of pre-trained general object detectors for initialization of learning tasks like, e.g., activity detection. Furthermore, the connection between eye-gaze driven supervision of learning tasks is challenged, due to the huge effort of generating meaningful saliency maps and the human focus on specific landmarks like faces, which are in general not activity specific. In general, evaluations on image saliency datasets showed the robustness of the proposed EBS approach when compared to state-of-the-art methods, especially on images with off-centered objects. This underlines that the presented encoding based saliency work flow is less dependent on prior assumptions than related approaches, and therefore reveals a wide range of application fields.

Overall, the experiments in this chapter encourage further utilization of the proposed EBS saliency method for achieving goals like weakly supervised activity detection training or improved representations for activity classifications. We believe that further statistical

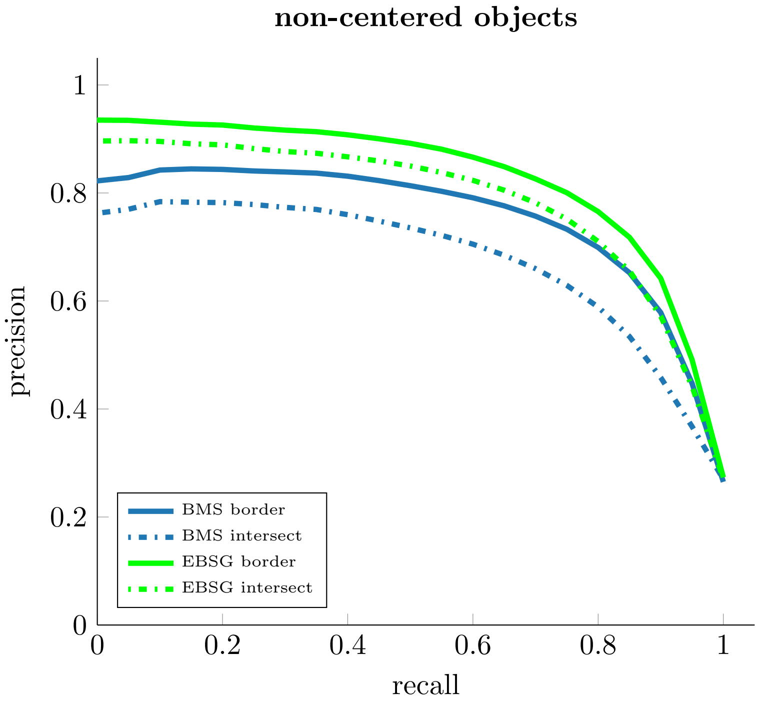


Figure 5.16: Results for off-centered object tests on the ASD dataset [3]. Proposed EBSG is compared against BMS [191] on different levels of off-centered objects, either slightly touching or intersecting the border. Our EBSG performs favorable against decreasing BMS results and, compared to results in Figure 5.14, shows nearly no decrease of performance for border touching objects.

measurements for saliency reliability and incorporation of additional top-down saliency maps could further augment our approach. Novel research in semantic image segmentation successfully applied saliency for weakly labeled training, like Oh et al. [81] or Shimoda and Yanai [147]. New challenges like actor-action recognition like proposed by Xu et al. [179] could benefit from saliency driven proposals too.

Further open questions for future research may be to incorporate high-level information as in top-down saliency methods (e.g., face, task specific objects or image and video context). Moreover, long-term temporal analysis of video context for eye-gaze and saliency prediction has been totally ignored so far within the literature, and will be needed for real prediction of human eye-gaze behavior, as it always incorporates the complete spatial and temporal context of a video.

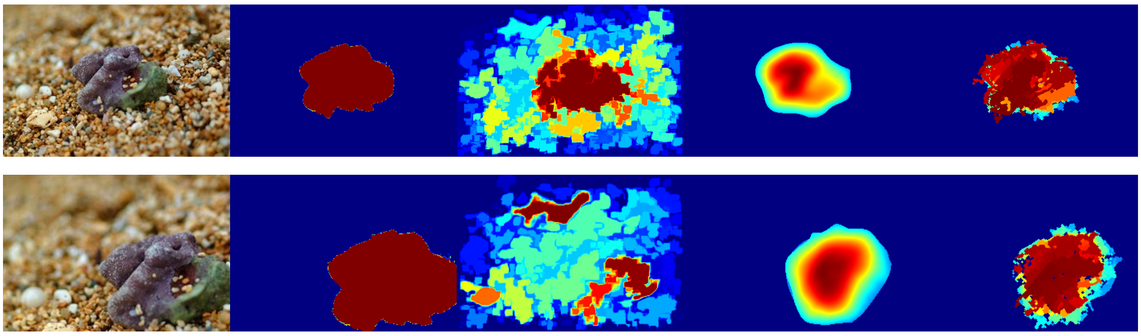


Figure 5.17: Example of image center prior for saliency estimation by generating off-centered objects. Top row shows (f.l.t.r) input image, ground truth, BMS, EBSG and EBSGR results. Bottom row shows the cropped image with the object touching the image border, but still defining the visual salient part of the image. Performance of BMS strongly decreases while EBSG still performs well. As we process region information after computation of the saliency, our EBSGR segmentation results still marks details of the object.

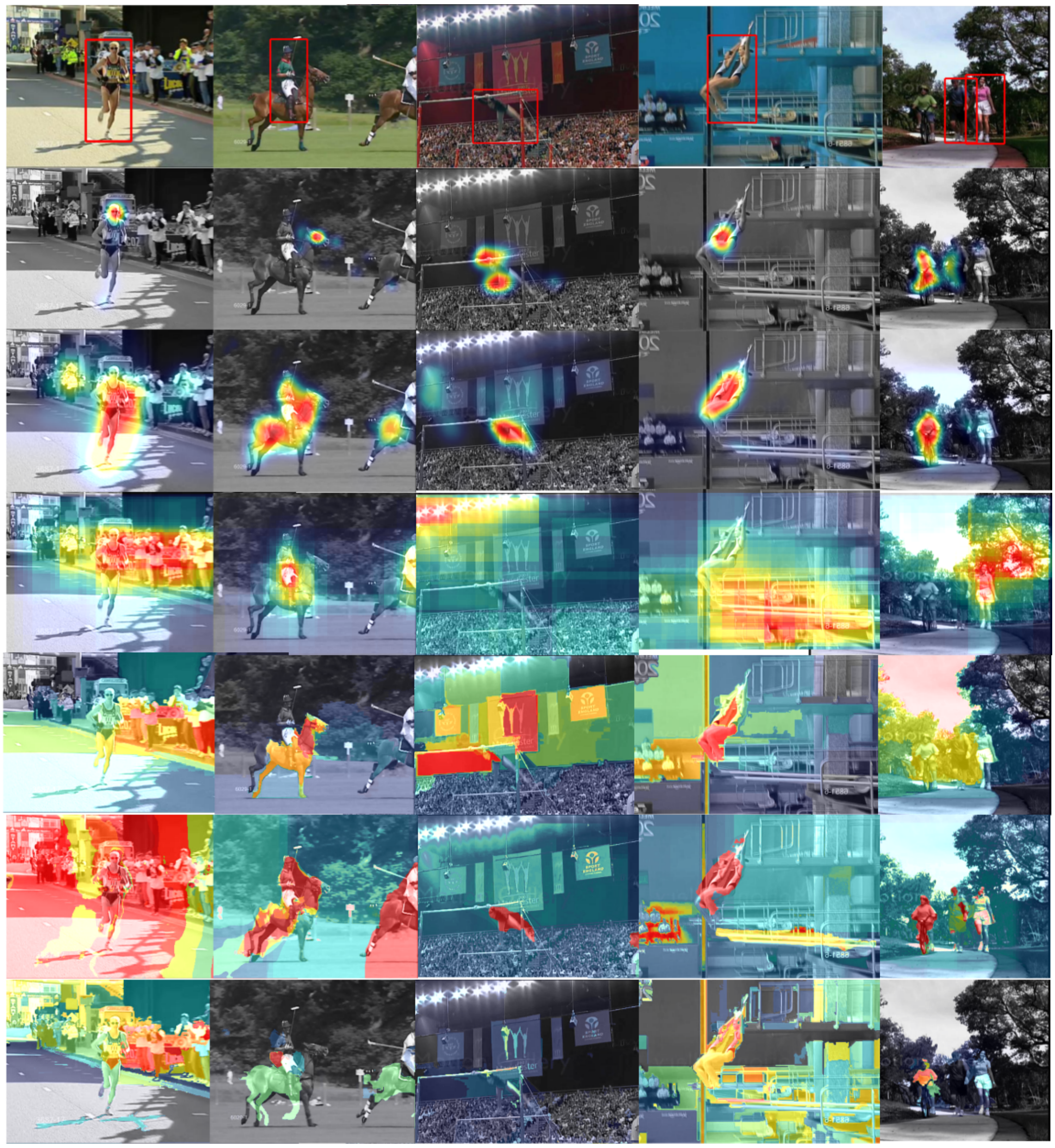


Figure 5.18: Examples of video saliency results on UCF sports. Top row: Input images with ground truth annotations. Second row: Eye-gaze tracking results collected by [108]. From row three to bottom: Our proposed method (EBS global), *objectness* detector [7], color saliency[77], video saliency methods [127] and [194]. See text for detailed discussion.

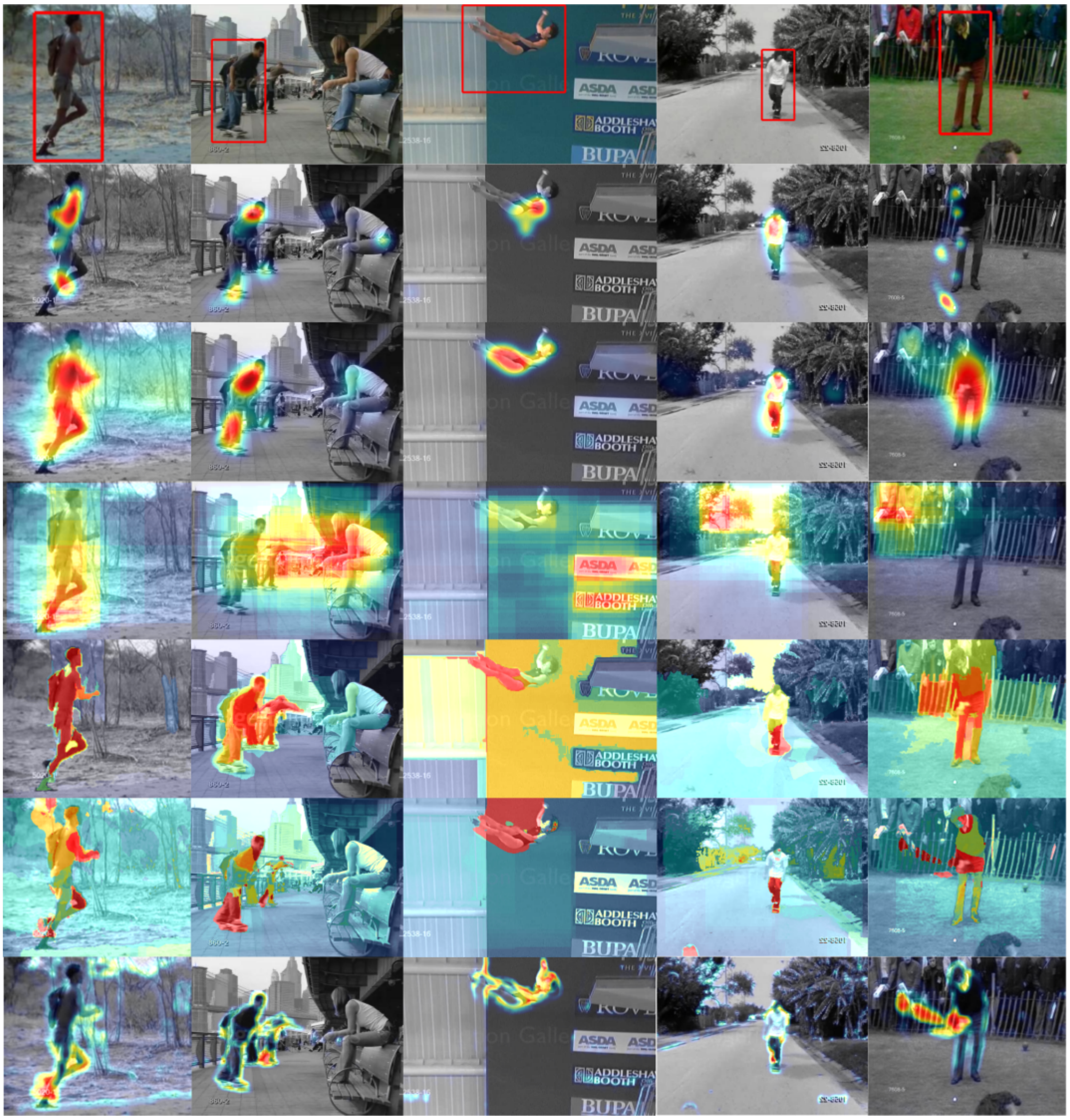


Figure 5.19: Additional examples of video saliency results on UCF sports. Top row: Input images with ground truth annotations. Second row: Eye-gaze tracking results collected by [108]. From row three to bottom: Our proposed method (EBS global), *objectness* detector [7], color saliency [77], video saliency methods [127] and [157]. See text for detailed discussion.

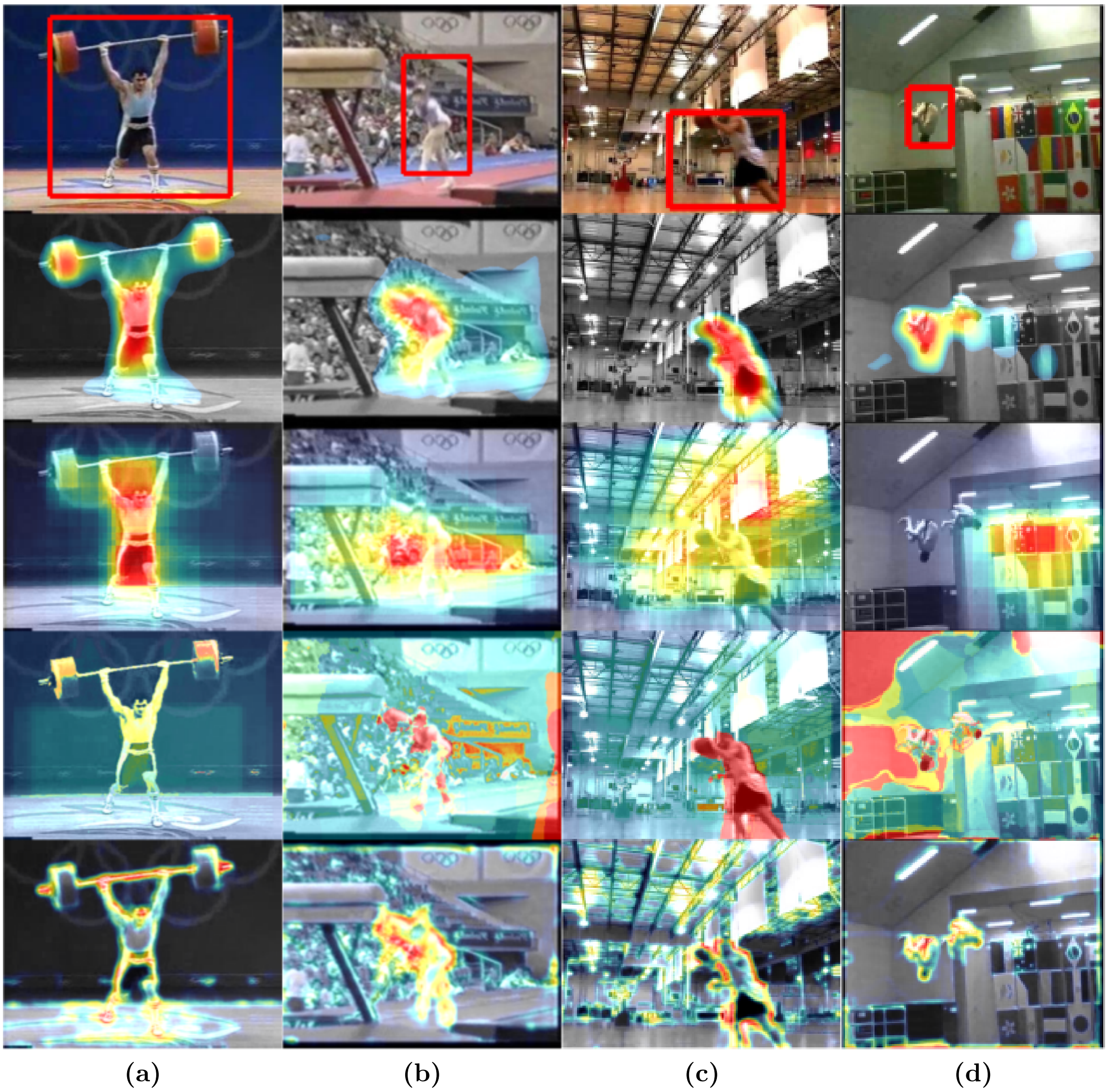


Figure 5.20: Examples of video saliency results on Olympic Sports dataset. From top to bottom: Image with ground truth bounding box, proposed EBSG results, *objectness* [7], and video saliency results [127, 157]. Videos with cluttered background or strong camera motion causes errors and false detections by state-of-the-art methods.

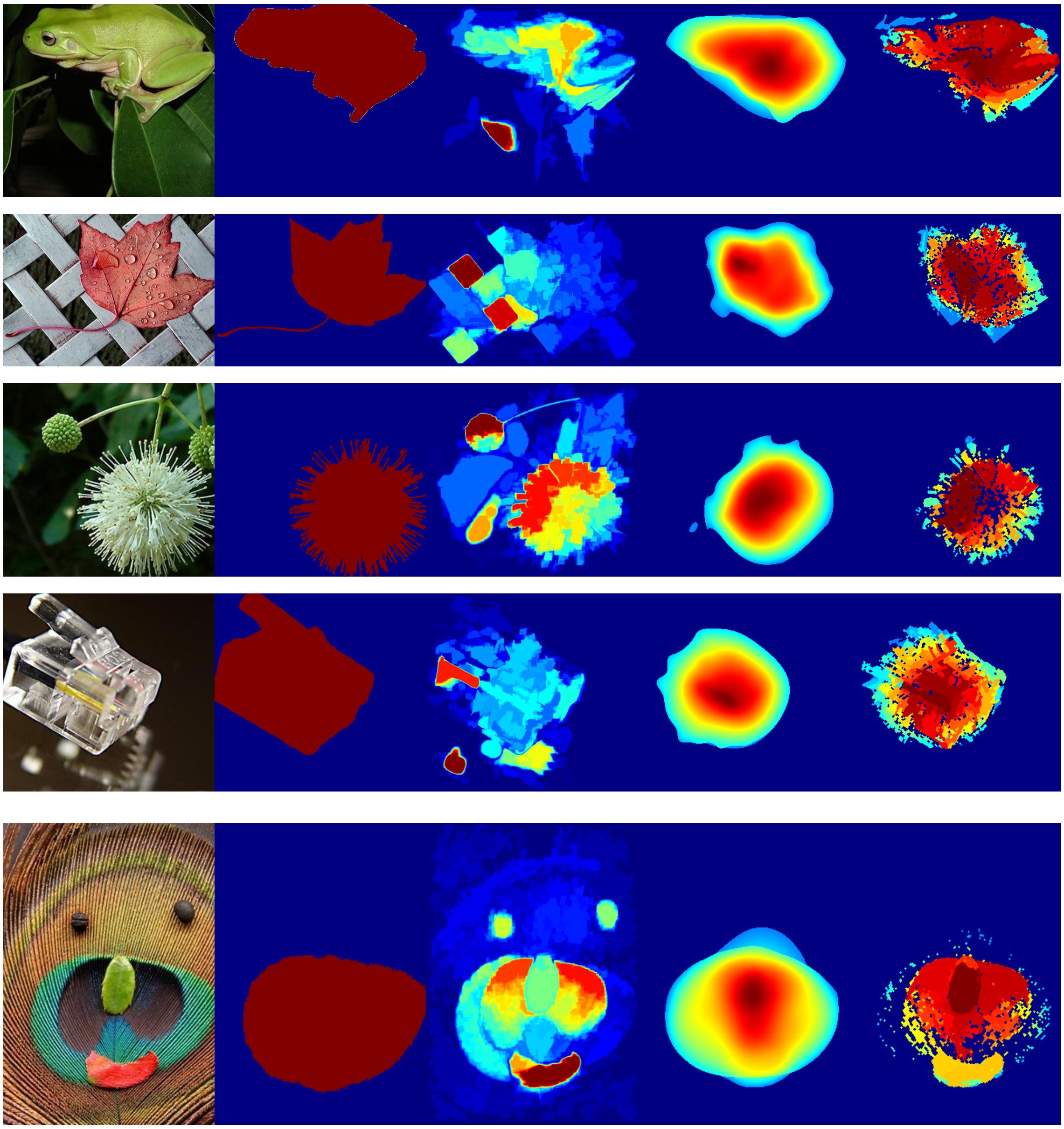


Figure 5.21: Visual comparison of EBSG and EBSGR with BMS [191] on off-centered examples objects touching the image border (f.l.t.r): Input image, ground truth, BMS and EBSG. Final column shows EBSGR using encoding information for over-segmentation and propagating high saliency values within this segments, which leads to sharper object contours.

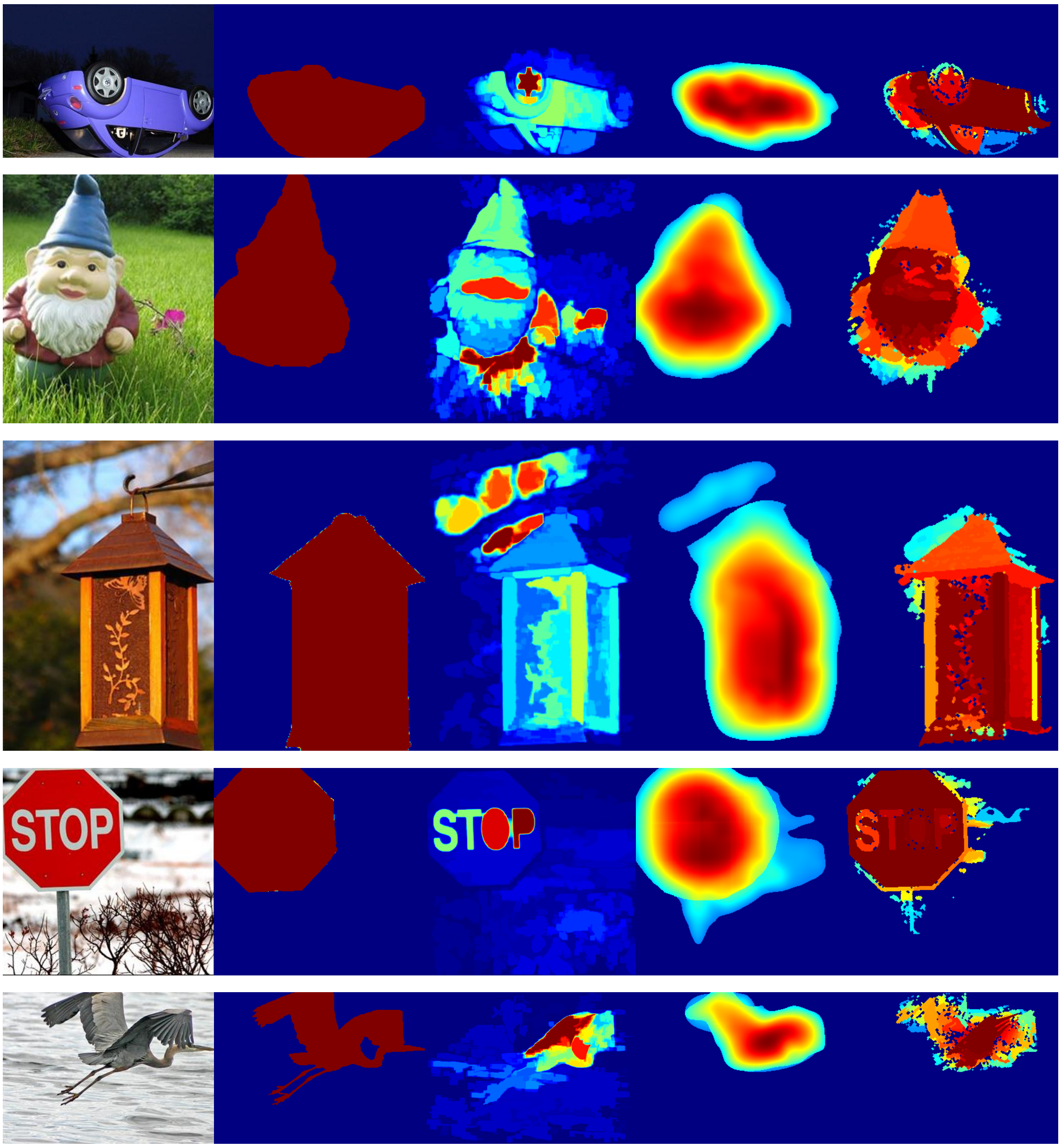


Figure 5.22: Further visual comparison of EBSG and EBSGR with BMS [191] on off-centered objects touching the image border (f.l.t.r): Input image, ground truth segmentation, BMS, EBSG and EBSGR.

6 Summary and Outlook

In the following, we summarize the ideas presented in this thesis and how the proposed approaches and results can be seen related to the potential future evolution of activity recognition. Several aspects and problem statements concerning ambiguities in the context of activity recognition have been discussed. Solutions for handling spatial, temporal and label ambiguities have been proposed and compared to related state-of-the-art. The related findings, besides the present thesis, have found their way into many research projects, publications and master theses of students. In relation to image classification, the combination of space and time information in videos does not only add additional dimensionality to the data, but exponentially increases the complexity and variability concerning relations and extend of different actors and events over time. Within the thesis, this important information of spatial and temporal relations within an activity could be exploited effectively for generating better classification and detection results.

Learning spatial and temporal relation and linking individual action detections over time is still an important task. The action tubes approach by Gkioxari and Malik [56] linked specifically trained detector results without prior motion knowledge. Looking into related research fields like object tracking, the authors of [189] augmented a CNN feature vector with motion actions like (left, right, up, down, scale,...), and retrained the network offline to guide a tracker during online tracking tasks with the predicted motion actions. To some extend both approaches have overlaps to ideas presented in chapter 3, and depicts the need for further research in that direction.

Furthermore, the importance of emphasizing the actor(s) within activity videos was enforced, by a solution for automatically generated saliency maps for marking the action related areas within videos and images in chapter 5. During the last years there was a rise of convolution neural networks (CNN) for feature representation, replacing hand crafted features and processing pipelines of the past. As discussed in [26] by Carreira and Zisserman, the change from applying image based CNNs to networks trained on huge video datasets nowadays available, like *YouTube1m* or *Kinetics* is the necessary next step. In addition, incorporation of human detections or focus attention into the action detection and classification process would be an interesting research direction. Similar ideas have been already successfully applied by using saliency maps in the area of semantic segmentation by Oh et al. [81] or Shimoda and Yanai [147] as well as the VideoLSTM approach [101].

As already discussed in the beginning of this thesis, the amount of videos for activity recognition benchmarks is drastically increasing. On the one hand this is necessary to avoid bias like within older datasets, on the other hand processing this amount of information becomes unfeasible. Such development would imply that researches may have to work on precomputed features and never analyze or even see the underlying video and its context. Thus, in addition we will need to develop more compact but still visualisable representations for activity videos, e.g., similar to the dynamic images proposed by [11].

Still open points are handling unconstrained video for temporal action localization and the incorporation of temporal and contextual information. Although the research community made a big enhancement applying CNNs for representation, connecting spatial and temporal context will be necessary to get more insight into the complex structures of understanding activity. Recently, interesting concept for representing temporal ordering within activity videos have been proposed by Sigurdsson et al. [148] as a fully-connected temporal CRF model on top of a two-stream networks. Other approaches describe the temporal evolution by ranking methods like Fernando et al. [49] or with an approximate rank pooling CNN layer [11] by Bilen and colleagues named *dynamic images*.

We will see a further fragmentation of video related activity recognition and human centered action recognition based on depth and wearable sensors. Data modalities and feature description as well as applications diver too much between those two areas, although they both contains activity recognition as their principal goal. Hence, activity recognition will stay one of the uprising fields in the currently extreme popular and dynamic field of computer vision.

A List of Publications

My work at the Institute for Computer Graphics and Vision led to the following peer-reviewed publications. For the sake of completeness of this thesis strongly related publications are listed with the respective abstracts in chronological order, otherwise only cited by title and authors.

A.1 2017

Spatiotemporal Saliency Estimation by Spectral Foreground Detection

Caglar Aytekin, Horst Possegger, Thomas Mauthner, Serkan Kiranyaz, Horst Bischof and Moncef Gabbouj
*IEEE Transactions on Multimedia , vol.PP, no.*99*, pp.1-1, 2017*
(preprint)

Abstract: We present a novel approach for spatio-temporal saliency detection by optimizing a unified criterion of color contrast, motion contrast, appearance and background cues. To this end, we first abstract the video by temporal super-pixels. Second, we propose a novel graph structure exploiting the saliency cues to assign the edge weights. The salient segments are then extracted by applying a spectral foreground detection method, Quantum Cuts, on this graph. We evaluate our approach on several public data sets for video saliency and activity localization to demonstrate the favorable performance of the proposed Video Quantum Cuts (VQCUT) compared to the state-of-the-art.

Pose-specific non-linear mappings in feature space towards multiview facial expression recognition

Mahdi Jampour, Vincent Lepetit, Thomas Mauthner and Horst Bischof
Image and Vision Computing, Volume 58*, Pages* $38 - 46$*, 2017*

Abstract: We introduce a novel approach to recognizing facial expressions over a large range of head poses. Like previous approaches, we map the features extracted from the input image to the corresponding features of the face with the same facial expression but seen in a frontal view. This allows us to collect all training data into a common referential and therefore benefit from more data to learn to recognize the expressions. However, by contrast with such previous work, our mapping depends on the pose of the input image: We first estimate the pose of the head in the input image, and then apply the mapping specifically learned for this pose. The features after mapping are therefore much more reliable for recognition purposes. In addition, we introduce a non-linear form for the mapping of the features, and we show that it is robust to occasional mistakes made by the pose estimation stage. We evaluate our approach with extensive experiments on two protocols of the BU3DFE and Multi-PIE data sets, and show that it outperforms the state-of-the-art on both data-sets.

A.2 2015

Encoding based saliency detection for videos and images

Thomas Mauthner, Horst Possegger, Georg Waltner and Horst Bischof
In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015*

Abstract: We present a novel video saliency detection method to support human activity recognition and weakly supervised training of activity detection algorithms. Recent research has emphasized the need for analyzing salient information in videos to minimize data-set bias or to supervise weakly labeled training of activity detectors. In contrast to previous methods we do not rely on training information given by either eye-gaze or annotation data, but propose a fully unsupervised algorithm to find salient regions within videos. In general, we enforce the Gestalt principle of figure-ground segregation for both appearance and motion cues. We introduce an encoding approach that allows for efficient computation of saliency by approximating joint feature distributions. We evaluate our approach on several data-sets, including challenging scenarios with cluttered background and camera motion, as well as salient object detection in images. Overall, we demonstrate favorable performance compared to state-of-the-art methods in estimating both ground-truth eye-gaze and activity annotations.

In defense of color-based model-free tracking

Horst Possegger, Thomas Mauthner and Horst Bischof
In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015*

Pairwise linear regression: An efficient and fast multi-view facial expression recognition

Mahdi Jampour, Thomas Mauthner and Horst Bischof
In: *Proceedings of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015*

Abstract: Multi-view facial expression recognition (MFER) is an active research topic in facial analysis. In fact, not only the accuracy but also time complexity is desirable for real applications. In this paper, we introduce a new fast and robust approach for recognizing facial expressions in arbitrary views. Our approach relies on learning linear regressions between pairs of non-frontal and frontal sets to virtually compensate occluded facial parts. We learn linear regression for projecting from non-frontal to frontal views. Such approximated frontal training features are applied for training view specific facial expression classifiers. We propose a number of different variants of our approach, including sparse encoding and ridge-regression for feature representation. While classical pose specific methods strongly depend on the quality of the pose estimation step, our approaches maintain their superior behavior even under severe pose noise. We evaluate on both BU3DFE and Multi-PIE data-sets and outperform the state-of-the-art in classification accuracy, even with a simple pose specific baseline method, while being extremely robust to feature noise and erroneous viewpoint estimation with our pairwise regression approaches.

Multi-view Facial Expressions Recognition using Local Linear Regression of Sparse Codes

Mahdi Jampour, Thomas Mauthner and Horst Bischof
In: *Proceedings of Computer Vision Winter Workshop (CVWW), 2015*

Abstract: We introduce a linear regression-based projection for multi-view facial expressions recognition (MFER) based on sparse features. While facial expression recognition (FER) approaches have become popular in frontal or near to frontal views, few papers demonstrate their results on arbitrary views of facial expressions. Our model relies on a new method for multi-view facial expression recognition, where we encode appearance-based facial features using sparse codes and learn projections from non- frontal to frontal views using linear regression projection. We then reconstruct facial features from the projected sparse codes using a common global dictionary. Finally, the reconstructed features are used for facial expression recognition. Our regression of sparse codes approach outperforms the state-of-the- art results on both protocols of BU3DFE data set.

A.3 2014

Indoor Activity Detection and Recognition for Automated Sport Games Analysis

Georg Waltner, Thomas Mauthner and Horst Bischof
In: *Proceedings of Austrian Association for Pattern Recognition (OAGM/AAPR) Workshop, 2014*

Abstract: Activity recognition in sport is an attractive field for computer vision research. Game, player and team analysis are of great interest and research topics within this field emerge with the goal of automated analysis. The very specific underlying rules of sports can be used as prior knowledge for the recognition task and present a constrained environment for evaluation. This paper describes recognition of single player activities in sport with special emphasis on volleyball. Starting from a per-frame player-centered activity recognition, we incorporate geometry and contextual information via an activity context descriptor that collects information about all player's activities over a certain timespan relative to the investigated player. The benefit of this context information on single player activity recognition is evaluated on our new real-life data set presenting a total amount of almost 36k annotated frames containing 7 activity classes within 6 videos of professional volleyball games. Our incorporation of the contextual information improves the average player-centered classification performance of 77.56% by up to 18.35% on specific classes, proving that spatio-temporal context is an important clue for activity recognition.

Improved Sport Activity Recognition using Spatio-temporal Context

Georg Waltner, Thomas Mauthner and Horst Bischof
In: *Proceedings of DVS-Conference on Computer Science in Sport (DVS/GSSS), 2014*

Abstract: Activity recognition in sport is an attractive field for computer vision research. Game, player and team analysis are of great interest and research topics within this field emerge with the goal of automated analysis. As the execution of same activities

differs between players and activities cannot be modeled by local description alone, additional information is needed. Inspired by the concept of group context ([Choi11], [Lan12], [Zhu13]), we employ contextual information to support activity recognition. Compared to other sport activity recognition systems, e.g., proposed by [Bialkowski13], we focus on single player activities rather than on general team activities.

Occlusion Geodesics for Online Multi-object Tracking

Horst Possegger, Thomas Mauthner, Peter M. Roth and Horst Bischof
In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014*

A novel method for the analysis of sequential actions in team handball

Paul Rudelsdorfer, Norbert Schrapf, Horst Possegger, Thomas Mauthner, Horst Bischof and Markus Tilp
International Journal of Computer Science in Sport Volume 13, Issue 1, 2014

Abstract: Performance in team sports crucially depends on the knowledge about the own and the opponents strengths and weaknesses. Since the analysis of single actions only provides restricted information on the game process, the analysis of sequential actions is from great importance to understand team tactics. In this paper, we introduce a novel method to analyze tactical behavior in team sports based on action sequences of positional data which are subsequently analyzed with artificial neural networks. We present custom-made software which allows annotating single actions with accurate manual position information. The process of building action sequences with the notational information of single actions in team handball is described step by step and the accuracy of the position determination is evaluated. The evaluation revealed a mean error of $0.16m(\pm 0.17m)$ for field positions on a handball field. Inter- and intra-rater reliability for identical camera setups are excellent (ICC=0.92 and 0.95 resp.). However, tests revealed that position accuracy is depending on camera setup (ICC=0.36). The results of the study demonstrate the applicability of the described method to gain action sequence data with accurate position information. The combination with neural networks gives an alternative approach to T-patterns for the analysis of sport games.

A.4 2013

Robust Real-Time Tracking of Multiple Objects by Volumetric Mass Densities

Horst Possegger, Sabine Sternig, Thomas Mauthner, Peter M. Roth and Horst Bischof
In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013*

A.5 2012

Unsupervised Calibration of Camera Networks an Virtual PTZ Cameras

Horst Possegger, Matthias Rüther, Sabine Sternig, Thomas Mauthner, Manfred Klopschitz, Peter Roth and Horst Bischof
In: *Proceedings of Computer Vision Winter Workshop (CVWW), 2012*

Learn to Move: Activity Specific Motion Models for Tracking by Detection

Thomas Mauthner, Peter Roth and Horst Bischof
In: *Proceedings of European Conference on Computer Vision Workshops and Demonstrations (ECCVW), 2012*

Abstract: In this paper, we focus on human activity detection, which solves detection, tracking, and recognition jointly. Existing approaches typically use off-the-shelf approaches for detection and tracking, ignoring naturally given prior knowledge. Hence, in this work we present a novel strategy for learning activity specific motion models by feature-to-temporal-displacement relationships. We propose a method based on an augmented version of canonical correlation analysis (AuCCA) for linking high-dimensional features to activity-specific spatial displacements over time. We compare this continuous and discriminative approach to other well established methods in the field of activity recognition and detection. In particular, we first improve activity detections by incorporating temporal forward and backward mappings for regularization of detections. Second, we extend a particle filter framework by using activity-specific motion proposals, allowing for drastically reducing the search space. To show these improvements, we run detailed evaluations on several benchmark data sets, clearly showing the advantages of our activity-specific motion models.

Real-time Activity Repetition Detection

Markus Murschitz, Thomas Mauthner, Peter Roth and Horst Bischof
In: *Proceedings of Computer Vision Winter Workshop (CVWW), 2012*

Abstract: In this work, we address the problem of real-time action recognition from video streams, in particular exploiting the information given by repetitive actions. Adapting the SPRING algorithm, a sequential variant of dynamic time warping (DTW), not only single frames can be classified but also repetitions of actions can be recognized. These particular benefits are demonstrated for two different publicly available data sets, showing promising results. Moreover, since we apply efficient GPU implementations for visual features and the classifier, we can ensure real-time capability, even using low cost consumer hardware. Thus, the system can also be applied in practical setups,e.g., for human computer interaction.

A.6 2011

Multi-camera multi-object tracking by robust hough-based homography projections

Sabine Sternig, Thomas Mauthner, Arnold Irschara, Peter M. Roth and Horst Bischof
In: *Proceedings of IEEE International Workshop on Visual Surveillance, 2011*

A.7 2010

Temporal feature weighting for prototype-based action recognition

Thomas Mauthner, Peter M. Roth and Horst Bischof
In: *Proceedings of the Asian Conference on Computer Vision (ACCV), 2010*

Abstract: In action recognition recently prototype-based classification methods became popular. However, such methods, even showing competitive classification results, are often limited due to too simple and thus insufficient representations and require a long-term analysis. To compensate these problems we propose to use more sophisticated features and an efficient prototype-based representation allowing for a single-frame evaluation. In particular, we apply four feature cues in parallel (two for appearance and two for motion) and apply a hierarchical k-means tree, where the obtained leaf nodes represent the prototypes. In addition, to increase the classification power, we introduce a temporal weighting scheme for the different information cues. Thus, in contrast to existing methods, which

typically use global weighting strategies (i.e., the same weights are applied for all data) the weights are estimated separately for a specific point in time. We demonstrate our approach on standard benchmark data sets showing excellent classification results. In particular, we give a detailed study on the applied features, the hierarchical tree representation, and the influence of temporal weighting as well as a competitive comparison to existing state-of-the-art methods.

FlowGames

Jakob Santner, Manuel Werlberger, Thomas Mauthner, Paier Wolfgang and Horst Bischof
In: *Proceedings of Int. Workshop on Computer Vision for Computer Games (CVCG) in conjunction with CVPR, 2010*
Computer vision-based interfaces to games hold the promise of rich natural interaction and thus a more realistic gaming experience. Therefore, the video games industry started to develop and market computer vision-based games recently with great success. Due to limited computational resources, they employ mostly simple algorithms such as background subtraction, instead of sophisticated motion estimation or gesture recognition methods. This not only results in a lack of robustness, but also in very limited interaction possibilities and thus reduced gaming experience. In this paper, we show a couple of concepts to control video games based on optical flow. We use a state-of-the-art optical flow algorithm able to be computed densely in real-time on GPUs, which are in fact built-in in nearly every gaming hardware available. Based on the estimated motion, we develop several computer games with increasing complexity: Starting with using the flow field as force acting on moveable objects, we span the spectrum to more sophisticated concepts such as controlling widgets and action recognition.

Efficient Object Detection Using Orthogonal NMF Descriptor Hierarchies

Thomas Mauthner, Stefan Kluckner, Peter Roth and Horst Bischof
In: *Proceedings of German Association for Pattern Recognition (DAGM/GCPR), 2010*

Abstract: Recently descriptors based on Histograms of Oriented Gradients (HOG) and Local Binary Patterns (LBP) have shown excellent results in object detection considering the precision as well as the recall. However, since these descriptors are based on high dimensional representations such approaches suffer from enormous memory and runtime requirements. The goal of this paper is to overcome these problems by introducing hierarchies of orthogonal Non-negative Matrix Factorizations (NMF). In fact, in this way a lower dimensional feature representation can be obtained without loosing the discriminative power of the original features. Moreover, the hierarchical structure allows to represent parts of patches on different scales allowing for a more robust classification. We show the effectiveness of our approach for two publicly available data sets and compare it to existing state-of-the-art methods. In addition, we demonstrate it in context of aerial imagery, where high dimensional images have to be processed requiring efficient methods.

A.8 2009

Tracking as Segmentation of Spatial-Temporal Volumes by Anisotropic Weighted TV

Markus Unger, Thomas Mauthner, Thomas Pock and Horst Bischof
In: *Proceedings of Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR), 2009*

Semantic Image Classification Using Consistent Regions and Individual Context

Stefan Kluckner, Thomas Mauthner, Peter M. Roth and Horst Bischof
In: *Proceedings of British Machine Vision Conference (BMWC), 2009*

Semantic Classification in Aerial Imagery by Integrating Appearance and Height Information

Stefan Kluckner, Thomas Mauthner, Peter M. Roth and Horst Bischof
In: *Proceedings of Asian Conference on Computer Vision (ACCV), 2009*

A Covariance Approximation on Euclidean Space for Visual Tracking

Stefan Kluckner, Thomas Mauthner, Peter M. Roth and Horst Bischof
In: *Proceedings of Austrian Association for Pattern Recognition (AAPR/OAGM) Workshop, 2009*

Efficient Human Action Recognition by Cascaded Linear Classification

Peter M. Roth, Thomas Mauthner, Inayatullah Khan, and Horst Bischof.
In: *Proceedings of IEEE Workshop on Video-Oriented Object and Event Classification in conjunction with ICCV, 2009*

Abstract: We present a human action recognition system suitable for very short sequences. In particular, we estimate Histograms of Oriented Gradients (HOGs) for the current frame as well as the corresponding dense flow field estimated from two frames. The thus obtained descriptors are then efficiently represented by the coefficients of a Non-negative Matrix Factorization (NMF). To further speed up the overall process, we apply an efficient cascaded Linear Discriminant Analysis (CLDA) classifier. In the experimental results we show the benefits of the proposed approach on standard benchmark data sets as well as on more challenging and realistic videos. In addition, since other state-of-the-art methods apply weighting between different cues, we provide a detailed analysis of the importance of weighting for action recognition and show that weighting is not necessarily required for the given task.

Instant Action Recognition

Thomas Mauthner, Peter M. Roth and Horst Bischof
In: *Proceedings of Scandinavian Conference on Image Analysis (SCIA), 2009*

Abstract: In this paper, we present an efficient system for action recognition from very short sequences. For action recognition typically appearance and/or motion information of an action is analyzed using a large number of frames. This is a limitation if very fast actions (e.g., in sport analysis) have to be analyzed. To overcome this limitation, we propose a method that uses a single-frame representation for actions based on appearance and motion information. In particular, we estimate Histograms of Oriented Gradients (HOGs) for the current frame as well as for the corresponding dense flow field. The thus obtained descriptors are efficiently represented by the coefficients of a Non-negative Matrix Factorization (NMF). Actions are classified using an one-vs-all Support Vector Machine. Since the flow can be estimated from two frames, in the evaluation stage only two consecutive frames are required for the action analysis. Both, the optical flow as well as the HOGs, can be computed very efficiently. In the experiments, we compare the proposed approach to state-of-the-art methods and show that it yields competitive results. In addition, we demonstrate action recognition for real-world beach volleyball sequences.

Action Recognition from a Small Number of Frames

Thomas Mauthner, Peter M. Roth and Horst Bischof
In: *Proceedings of Computer Vision Winter Workshop (CVWW), 2009*

Abstract: In this paper, we present an efficient system for action recognition from very short sequences. For action recognition typically appearance and/or motion information of an action is analyzed using a large number of frames, which is often not sufficient, if very

fast actions (e.g., in sport analysis) have to be analyzed. To overcome this limitation, we propose a method that uses a single-frame representation for actions based on appearance and on motion information. In particular, we estimate Histograms of Oriented Gradients (HOGs) for the current sample as well as for a flow field. The thus obtained descriptors are then efficiently represented by the coefficients of a Non-negative Matrix Factorization (NMF). Actions are classified using an one-vs-all Support Vector Machine. Since the flow can be estimated

Evaluation of visual position estimation in beach volleyball

Christina Koch, Thomas Mauthner , Markus Tilp and Norbert Schrapf
International Journal of Performance Analysis in Sport, Volume 9, Issue 3, 2009

Abstract: Position determination of game analysts is often performed by subjective visual estimation. The aim of this study was to evaluate human position estimations for setting actions in beach volleyball. Subjects were asked to assign the athlete's position to one of five cells representing the court. Position estimations from seven beach volleyball experts and seven non-experts were compared with results of video metrology also presented in this paper. A mean error-quote of 41% indicates that the perception of humans assessing athlete's position is very inaccurate. While beach volleyball experience did not influence the ability for position estimation, the accuracy was dependent on the distance from the camera. The obtained results should be taken into account by coaches and scientist, who deduct their strategy on such analysis. If it is not possible to get exact position information with technical systems, we recommend close-up views and adequate cell size to improve validity of results.

A.9 2008

A Probabilistic Approach for Tracking Fibers

Michael Donoser, Thomas Mauthner and Horst Bischof
In: *Proceedings of International Conference on Pattern Recognition (ICPR), 2008*

Robust Tracking of Spatial Related Components

Thomas Mauthner, Michael Donoser and Horst Bischof
In: *Proceedings of International Conference on Pattern Recognition (ICPR), 2008*

A.10 2007

A Robust Multiple Object Tracking for Sport Applications

Thomas Mauthner and Horst Bischof
In: *Proceedings of Austrian Association for Pattern Recognition (AAPR/OAGM) Workshop, 2007*

Visual Tracking of Athletes in Beach Volleyball Using a Single Camera

Thomas Mauthner, Christina Koch, Markus Tilp and Horst Bischof
International Journal of Computer Science in Sport, Volume 6, Edition 2, 2007

Bibliography

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *Proc. USENIX Symposium on Operating Systems Design and Implementation*, 2016. (page 13)

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv*, 1609.08675, 2016. (page 13)

[3] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. (page 83, 91, 101, 102, 105)

[4] J. K. Aggarwal and Michel S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):16:1–16:43, 2011. (page 5)

[5] Omid Aghazadeh, Josephine Sullivan, and Stefan Carlsson. Novelty detection from an ego-centric perspective. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011. (page 14)

[6] M. Aharon, M. Elad, and A. Bruckstein. K -SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, Nov 2006. (page 58, 66)

[7] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010. (page 60, 93, 95, 96, 99, 100, 107, 108, 109)

[8] Mohamed R. Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song-Chun Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *Proc. European Conf. on Computer Vision*. Springer, 2012. (page 14)

[9] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002. (page 39, 40)

[10] Nicolas Ballas, Yi Yang, Zhen-Zhong Lan, Bertrand Delezoide, Francoise Preteux, and Alexander Hauptmann. Space-time robust representation for action recognition. In *Proc. IEEE Intern. Conf. on Computer Vision*, December 2013. (page 84, 98)

[11] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic image networks for action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016. (page 114)

[12] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2005. (page 6, 27, 41, 49)

[13] V. Bloom, D. Makris, and V. Argyriou. G3D: A gaming action dataset and real time action recognition evaluation framework. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2012. (page 16)

[14] Victoria Bloom, Vasileios Argyriou, and Dimitrios Makris. Hierarchical transfer learning for online recognition of compound actions. *Computer Vision and Image Understanding*, 144(C):62–72, March 2016. (page 16)

[15] S. J. Blunsden and Robert Fisher. The behave video dataset: ground truthed video for multi-person. In *Proc. British Machine Vision Conf.*, 2010. (page 8, 14)

[16] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257 – 267, 2001. (page 5, 6)

[17] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2013. (page 59)

[18] Ali Borji, DickyN. Sihite, and Laurent Itti. Salient object detection: A benchmark. In *Proc. European Conf. on Computer Vision*, 2012. (page 83, 92, 101)

[19] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Image classification using random forests and ferns. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2007. (page 38)

[20] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. (page 38)

[21] William Brendel and Sinisa Todorovic. Learning spatiotemporal graphs of human activities. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2011. (page 77, 78)

[22] Xavier P. Burgos-Artizzu, Piotr Dollar, Dayu Lin, David J. Anderson, and Pietro Perona. Social behaviour recognition in continuous video. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012. (page 26)

[23] Zoya Bylinskii, Tilke Judd, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT saliency benchmark. http://saliency.mit.edu/, 2015. (page 101)

[24] Liangliang Cao, Zicheng Liu, and T.S. Huang. Cross-dataset action detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010. (page xix, 7, 8)

[25] Yu Cao, Daniel Barrett, Andrei Barbu, Siddharth Narayanaswamy, Aaron Michaux Haonan Yu and, Yuewei Lin, Sven Dickinson, and Jeffrey Mark Siskinda nd Song Wang. Recognize human activities from partially observed videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013. (page 60)

[26] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, July 2017. (page 113)

[27] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. (page 101)

[28] Jose M. Chaquet, Enrique J. Carmona, and Antonio Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633 – 659, 2013. (page 5)

[29] Yi-Chen Chen, Vishal Patel, Jaishanker Pillai, Rama Chellappa, and P. Jonathan Phillips. Dictionary learning from ambiguously labeled data. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013. (page 60, 67, 72, 73)

[30] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011. (page 83, 86, 101, 102)

[31] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015. (page 83, 91, 101)

[32] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *Proc. IEEE International Workshop on Visual Surveillance*, 2009. (page 14)

[33] T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *Journal of Machine Learning Research(JMLR)*, 12:1225–1261, 2011. (page 60, 67, 72, 73)

[34] Xinyi Cui, Qingshan Liu, and Dimitris Metaxas. Temporal spectral residual: Fast motion saliency detection. In *Proc. of ACM International Conference on Multimedia*, 2009. (page 83)

[35] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2005. (page 21, 29)

[36] Navneet Dalal, Bill Triggs, and Cordelia Schmid. Human detection using oriented histograms of flow and appearance. In *Proc. European Conf. on Computer Vision*, 2006. (page 21, 23)

[37] Giovanni Denina, Bir Bhanu, HoangThanh Nguyen, Chong Ding, Ahmed Kamal, Chinya Ravishankar, Amit Roy-Chowdhury, Allen Ivers, and Brenda Varda. Videoweb dataset for multi-camera activities and non-verbal communication. In *Distributed Video Sensor Networks*, pages 335–347. Springer London, 2011. (page 14)

[38] Piotr Dollar, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *Proc. PETS Workshop*, 2005. (page 18, 27)

[39] Michael Donoser and Dieter Schmalstieg. Discriminative feature-to-point matching in image-based localization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE, 2014. (page 56)

[40] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2003. (page 28)

[41] A. Elgammal, V. Shet, Y. Yacoob, and L. S. Davis. Learning dynamics for exemplar-based gesture recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2003. (page 27)

[42] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf, 2006. (page 96)

[43] Abdalrahman Eweiwi, Muhammad Shahzad Cheema, and Christian Bauckhage. Discriminative joint non-negative matrix factorization for human action classification. In *Proc. German Conference on Pattern Recognition*, 2013. (page 65, 66)

[44] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015. (page 13)

[45] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the Scandinavian Conference on Image Analysis*, Gothenburg, Sweden, 2003. (page 20)

[46] A. Fathi, Xiaofeng Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011. (page 14)

[47] Alireza Fathi and Greg Mori. Action recognition by learning mid-level motion features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008. (page 49)

[48] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. Dynamically encoded actions based on spacetime saliency. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015. (page 84, 91, 93)

[49] Basura Fernando, Efstratios Gavves, Jose M. Oramas, Amir Ghodrati, and Tinne Tuytelaars. Modeling video evolution for action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015. (page 114)

[50] Simon Fothergill, Helena M. Mentis , Pushmeet Kohli, and Sebastian Nowozin. Instructing people for training gestural interactive systems. In *SIGCHI Conference on Human Factors in Computing Systems*. ACM, May 2012. (page 16)

[51] Simone Frintrop, Thomas Werner, and Germán M. García. Traditional saliency reloaded: A good old model in new shape. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015. (page 91)

[52] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(11):2188–2202, 2011. (page 60)

[53] Juergen Gall and Victor Lempitsky. Class-specific hough forests for object detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. (page 26, 35, 38)

[54] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury. String of feature graphs model for recognition of complex activities in natural videos. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2011. (page 77, 78)

[55] Peter V. Gehler and Olivier Chapelle. Deterministic annealing for multiple-instance learning. In *In Proc. Artificial Intelligence and Statistics (AIStats)*, 2007. (page 59, 70, 72)

[56] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015. (page 113)

[57] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as Space-Time Shapes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007. (page 6, 26, 41, 49, 50, 91, 93, 94)

[58] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal Saliency detection using phase spectrum of quaternion fourier transform. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008. (page 83)

[59] Sunil Kumar Gupta, Dinh Phung, Brett Adams, Truyen Tran, and Svetha Venkatesh. Nonnegative shared subspace learning and its application to social media retrieval. In *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010. (page 58, 65)

[60] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based Visual Saliency. In *Advances in Neural Information Processing Systems*, 2006. (page 83, 84)

[61] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Proc. Alvey Vision Conference*, 1988. (page 18)

[62] Ngoc-Diep Ho. *Nonnegative matrix factorization algorithms and applications*. PhD thesis, Universite' Catholique de Louvain, 2008. (page 62)

[63] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence, 17*, pages 185–203, 1981. (page 20)

[64] Sayed Hossein Khatoonabadi, Nuno Vasconcelos, Ivan V. Bajic, and Yufeng Shan. How many bits does it take for a stimulus to be salient? In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015. (page 83)

[65] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936. (page 36)

[66] Yuxiao Hu, Liangliang Cao, Fengjun Lv, Shuicheng Yan, Yihong Gong, and Thomas S. Huang. Action detection in complex scenes with spatial and temporal ambiguities. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2009. (page 59, 60)

[67] Eyke Huellermeier, Johannes Fuernkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16-17):1897 – 1916, 2008. (page 64)

[68] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016. (page 14)

[69] Nazli Ikizler, R. Gokberk Cinbis, and Pinar Duygulu. Human action recognition with line and flow histograms. In *Proc. Intern. Conf. on Pattern Recognition*, 2008. (page 28, 46)

[70] Nazli Ikizler-Cinbis, R. Gokberk Cinbis, and Stan Sclaroff. Learning actions from theweb. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2009. (page 60, 68, 76)

[71] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. of International Conference on Machine Learning*, volume 37, 2015. (page 13)

[72] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. (page 83)

[73] Mahdi Jampour, Vincent Lepetit, Thomas Mauthner, and Horst Bischof. Pose-specific non-linear mappings in feature space towards multiview facial expression recognition. *Image and Vision Computing*, 58:38 – 46, 2017. (page 79)

[74] Mahdi Jampour, Thomas Mauthner, and Horst Bischof. Pairwise linear regression: An efficient and fast multi-view facial expression recognition. In *Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015. (page 79)

[75] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2013. (page xix, 10, 11, 12, 16, 20)

[76] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2007. (page 49, 50)

[77] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Nanning Zheng, and Shipeng Li. Automatic Salient Object Segmentation Based on Context and Shape Prior. In *Proc. British Machine Vision Conf.*, 2011. (page 94, 100, 107, 108)

[78] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *arXiv*, 1502.07209, 2015. (page 13)

[79] Zhuolin Jiang, Zhe Lin, and L.S. Davis. Label consistent K-SVD: Learning a discriminative dictionary for recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664, Nov 2013. (page 60, 65, 66, 67)

[80] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, 1973. (page 82)

[81] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. (page 105, 114)

[82] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to Predict Where Humans Look. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2009. (page 82, 97)

[83] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014. (page 13)

[84] Yan Ke, Rahul Sukthankar, and Martial Hebert1. Event detection in crowded videos. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2007. (page 7, 27)

[85] Sameh Khamis, Vlad I. Morariu, and Larry S. Davis. A flow model for joint action recognition and identity maintenance. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012. (page 26)

[86] Kris M. Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011. (page 14)

[87] Alexander Kläser, Marcin Marszalek, and Cordelia Schmid. A spatio-temporal descriptor based on 3D-Gradients. In *Proc. British Machine Vision Conf.*, 2008. (page 22)

[88] Orit Kliper-Gross, Tal Hassner, and Lior Wolf. The action similarity labeling challenge. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(3):615–621, March 2012. (page 5, 12)

[89] Yu Kong, Yunde Jia, and Yun Fu. Learning human interaction by interactive phrases. In *Proc. European Conf. on Computer Vision*, 2012. (page 14)

[90] Adriana Kovashka and Kristen Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010. (page 26)

[91] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2011. (page 10)

[92] Tian Lan, Yang Wang, and Greg. Discriminative figure-centric models for joint action localization and recognition. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2011. (page 59, 60, 74, 84, 98, 99)

[93] Gert R. G. Lanckriet, Tijl de Bie, Nello Cristianini, Michael I. Jordan, and William S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635, 2004. (page 46)

[94] Ivan Laptev. On space-time interest points. *International Journal on Computer Vision*, 64:107–123, 2005. (page 18)

[95] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008. (page 21, 22, 59, 74)

[96] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, 2001. (page 58, 61)

[97] Jian Li, Levine, M.D., Xiangjing An, Xin Xu, and Hangen He. Visual Saliency Based on Scale-Space Analysis in the Frequency Domain. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(4):996–1010, 2013. (page 83, 101, 102)

[98] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2010. (page 16)

[99] Yin Li, Xiaodi Hou, Christof Koch, JamesM Rehg, and Alan L. Yuille. The secrets of Salient Object Segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014. (page 91, 103)

[100] Yin Li, Zhefan Ye, and James M. Rehg. Delving into egocentric actions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015. (page 14)

[101] Zhenyang Li, Efstratios Gavves, Mihir Jain, and Cees G. M. Snoek. VideoLSTM convolves, attends and flows for action recognition. *arXiv*, 1607.01794, 2016. (page 114)

[102] Zhe Lin, Zhuolin Jiang, and Larry S. Davis. Recognizing actions by shape-motion prototype trees. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2009. (page 26, 27, 28, 40, 49, 50)

[103] Jialu Liu, Chi Wang, J. Gao, Chi Wang, Jing Gao, and Jiawei Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proc. SIAM International Conference on Data Mining*, 2013. (page 65)

[104] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to Detect A Salient Object. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007. (page 83)

[105] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981. (page 21)

[106] V. Mahadevan and N. Vasconcelos. Spatiotemporal Saliency in Dynamic Scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(1):171–177, 2010. (page 85)

[107] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. (page 12)

[108] Stefan Mathe and Cristian Sminchisescu. Dynamic Eye Movement Dataset and Learnt Saliency Models for Visual Action Recognition. In *Proc. European Conf. on Computer Vision*, 2012. (page 84, 93, 95, 100, 107, 108)

[109] Thomas Melzer, Michael Reiter, and Horst Bischof. Appearance models based on kernel canonical correlation analysis. *Pattern Recognition*, 36(9):1961–1971, 2003. (page 36)

[110] Parag K Mital, Tim J Smith, Robin L Hill, and John M Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24, 2011. (page 82, 84)

[111] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database HDM05. Technical Report CG-2007-2, Universität Bonn, 2007. (page 16)

[112] Meinard Müller, Andreas Baak, and Hans-Peter Seidel. Efficient and robust annotation of motion capture data. In *Proc. of ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, 2009. (page 16)

[113] Markus Murschitz. Video-based human body action recognition for games. Master's thesis, Graz University of Technology, 2011. (page xx, 22, 23, 56)

[114] Markus Murschitz, Thomas Mauthner, Peter Roth, and Horst Bischof. Real-time activity repetition detection. In *Proceedings of Computer Vision Winter Workshop*, 2012. (page 56)

[115] Fiza Murtaza, Muhammad Haroon Yousaf, and Sergio A Velastin. Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description. *IET Computer Vision*, 10:758–767, 2016. (page 14)

[116] JuanCarlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *Proc. European Conf. on Computer Vision*, 2010. (page xix, 10, 91, 95)

[117] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006. (page 26, 28, 29, 32, 35, 38)

[118] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley MHAD: A comprehensive multimodal human action database. In *Proc. IEEE Workshop on Applications of Computer Vision (WACV)*, 2013. (page xix, 16, 17)

[119] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011. (page 8, 14)

[120] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51 – 59, 1996. (page 29, 31)

[121] Stephen E. Palmer. *Vision Science, Photons to Phenomenology.* MIT Press, 1999. (page 85)

[122] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Proc. European Conf. on Computer Vision*, 2010. (page 100)

[123] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990, 2010. (page 5)

[124] Qiang Qiu, Zhuolin Jiang, and R. Chellappa. Sparse dictionary-based representation and recognition of action attributes. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2011. (page 60, 63)

[125] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian. Histogram of oriented principal components for cross-view action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 38(12):2430–2443, Dec 2016. (page 17)

[126] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition. In *Proc. European Conf. on Computer Vision*, Cham, 2014. (page 17)

[127] Esa Rahtu, Juho Kannala, Mikko Salo, and Janne Heikkilä. Segmenting Salient Objects from Images and Videos. In *Proc. European Conf. on Computer Vision*, 2010. (page 82, 83, 84, 85, 86, 88, 93, 94, 95, 96, 100, 102, 107, 108, 109)

[128] Tom Rainforth and Frank Wood. Canonical correlation forests, 2015. (page 56)

[129] Alain Rakotomamonjy, Francis R. Bach, Stèphane Canu, and Yves Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9, 2008. (page 46, 47)

[130] Michalis Raptis and Leonid Sigal. Poselet key-framing: A model for human activity recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013. (page 60, 78)

[131] Kishore K. Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, Jul 2013. (page 10)

[132] Xiaofeng Ren and Deva Ramanan. Histograms of sparse codes for object detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013. (page 60, 76)

[133] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH - a spatio temporal maximum average correlation height filter for action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008. (page xix, xx, 9, 26, 31, 50, 91, 94)

[134] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *Proc. European Conf. on Computer Vision*. Springer, 2012. (page xix, 14, 15)

[135] E. Rosten and T. Drummond. Machine learning for high- speed corner detection. In *Proc. European Conf. on Computer Vision*, 2006. (page 19)

[136] Peter M. Roth, Thomas Mauthner, Inayatullah Khan, and Horst Bischof. Efficient human action recognition by cascaded linear classifcation. In *1st IEEE Workshop on Video-Oriented Object and Event Classification*, 2009. (page 43, 45, 46, 47, 48, 49, 56, 60, 68, 76, 79)

[137] Dmitry Rudoy, Dan B Goldman, Eli Shechtman, and Lihi Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013. (page 84, 85, 88)

[138] Michael S. Ryoo. Early recognition of ongoing activities from streaming videos. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2011. (page 78)

[139] Michael S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2009. (page 8, 14, 67, 76, 77, 78)

[140] Daniel Scharstein. Middlebury optical flow benchmark. http://vision.middlebury.edu/flow/, 2015. (page 88)

[141] Konrad Schindler and Luc van Gool. Action snippets: How many frames does human action recognition require? In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008. (page 27, 28, 43, 46, 48, 49, 50)

[142] Christian Schueldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local svm approach. In *Proc. Intern. Conf. on Pattern Recognition*, 2004. (page 6, 27, 41)

[143] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2013. (page 16)

[144] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016. (page 17)

[145] Nataliya Shapovalova, Michalis Raptis, Leonid Sigal, and Greg Mori. Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization. In *Advances in Neural Information Processing Systems*, 2013. (page 84, 85, 93, 95, 97, 100)

[146] Nataliya Shapovalova, Arash Vahdat, Kevin Cannons, Tian Lan, and Greg Mori. Similarity Constrained Latent Support Vector Machine: An Application to Weakly Supervised Action Classification. In *Proc. European Conf. on Computer Vision*, 2012. (page 59, 60, 74, 84, 93, 95)

[147] Wataru Shimoda and Keiji Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *Proc. European Conf. on Computer Vision*, 2016. (page 105, 114)

[148] Gunnar A. Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. (page 114)

[149] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proc. European Conf. on Computer Vision*, 2016. (page 15)

[150] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016. (page 8)

[151] Sanchit Singh, Sergio A Velastin, and Hossein Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010. (page 14)

[152] Parthipan Siva, Chris Russell, and Tao Xiang. In defence of negative mining for annotating weakly labelled data. In *Proc. European Conf. on Computer Vision*, 2012. (page 59, 60, 70, 72, 74)

[153] Sören Sonnenburg, Bernhard S. Bernhard, P. Bennett, and Emilio Parrado-Hernández. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006. (page 46)

[154] Khurram Soomro and Amir Roshan Zamir. Action recognition in realistic sports videos. In *Computer Vision in Sports*, pages 181–208. Springer, 2014. (page 9)

[155] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human action classes from videos in the wild. Technical Report CRCV-TR-12-01, Center for Research in Computer Vision, November 2012. (page xix, 10, 11)

[156] S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proc. of ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2013. (page 15)

[157] Waqas Sultani and Imran Saleemi. Human Action Recognition across Datasets by Foreground-weighted Histogram Decomposition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014. (page 84, 88, 90, 92, 93, 95, 96, 98, 100, 108, 109)

[158] Ju Sun, Xiao Wu, Shuicheng Yan, L. F. Cheong, T. S. Chua, and Jintao Li. Hierarchical spatio-temporal context modeling for action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. (page 21)

[159] Peter M. Roth Thomas Mauthner and Horst Bischof. Instant action recognition. In *Proc. Scandinavian Conference on Image Analysis*, 2009. (page 47, 56, 79)

[160] Peter M. Roth Thomas Mauthner and Horst Bischof. Temporal feature weighting for prototype-based action recognition. In *Proc. Asian Conf. on Computer Vision*, 2010. (page)

[161] Peter M. Roth Thomas Mauthner and Horst Bischof. Learn to move: Activity specific motion models for tracking by detection. In *Proc. European Conf. on Computer Vision*, 2012. (page 56)

[162] Ch. Thurau and V. Hlaváč. Pose primitive based human action recognition in videos or still images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008. (page 27, 28, 33, 43, 48, 49, 60, 65, 68, 76)

[163] J.A. Tropp and A.C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, Dec 2007. (page 66)

[164] Eleonora Vig, Micheal Dorr, and David Cox. Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014. (page 82, 83)

[165] Georg Waltner. Indoor activity detection and recognition for automated sport games analysis. Master's thesis, Graz University of Technology, 2014. (page 56)

[166] Georg Waltner, Thomas Mauthner, and Horst Bischof. Improved sport activity recognition using spatio-temporal context. In *Proc. DVS-Conference on Computer Science in Sport (DVS/GSSS)*, 2014. (page 56)

[167] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, May 2013. (page xx, 20, 21, 24)

[168] Heng Wang, Muhammad Muneeb Ullah, Alexander Kläser Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. British Machine Vision Conf.*, 2009. (page 21, 26, 50)

[169] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012. (page 16)

[170] J. Wang, X. Nie, Y. Xia, Y. Wu, and S. C. Zhu. Cross-view action modeling, learning, and recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014. (page 17)

[171] Xiaoyu Wang, Tony X. Han, and Shuicheng Yan. An HOG-LBP human detector with partial occlusion handling. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2009. (page 31)

[172] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic Saliency Using Background Priors. In *Proc. European Conf. on Computer Vision*, 2012. (page 83, 101, 103)

[173] Daniel Weinland and Edmond Boyer. Action recognition using exemplar-based embedding. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008. (page 27, 28, 33)

[174] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2):249–257, November 2006. (page 14)

[175] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224 – 241, 2011. (page 5)

[176] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 optical flow. In *Proc. British Machine Vision Conf.*, 2009. (page 20)

[177] Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, Christophe Garcia, and Bülent Sankur. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127(Supplement C):14 – 30, 2014. (page 16)

[178] L. Xia, C.C. Chen, and JK Aggarwal. View invariant human action recognition using histograms of 3D joints. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2012. UT-kinect. (page 16)

[179] C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso. Can humans fly? Action understanding with multiple classes of actors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2015. (page xix, 8, 9, 105)

[180] Qiong Yan, Li Xu, Jiangping Shi, and Jiaya Jia. Hierarchical Saliency Detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013. (page 83, 101, 102)

[181] Y. Yan, C. Xu, D. Cai, and J. J. Corso. Weakly supervised actor-action segmentation via robust multi-task ranking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. (page 8)

[182] Angela Yao, Juergen Gall, and Luc Van Gool. A hough transform-based voting framework for action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010. (page 26, 40, 50)

[183] Benjamin Yao and Song-Chun Zhu. Learning deformable action templates from cluttered videos. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2009. (page 50)

[184] Lahav Yeffet and Lior Wolf. Local trinary patterns for human action recognition. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2009. (page 31)

[185] Gang Yu, Junsong Yuan, and Zicheng Liu. Propagative hough voting for human activity recognition. In *Proc. European Conf. on Computer Vision*, 2012. (page 60, 78)

[186] Tsz-Ho Yu, Tae-Kyun Kim, and Roberto Cipolla. Real-time action recognition by spatiotemporal semantic and structural forests. In *Proc. British Machine Vision Conf.*, 2010. (page xx, 18, 19, 77, 78)

[187] Junsong Yuan, Zicheng Liu, and Ying Wu. Discriminative subvolume search for efficient action detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. (page 7)

[188] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2012. (page 14)

[189] Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Jin Young Choi. Action-decision networks for visual tracking with deep reinforcement learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, July 2017. (page 113)

[190] Yun Zhai and Mubarak Shah. Visual Attention Detection in Video Sequences Using Spatiotemporal Cues. In *Proc. ACM International Conference on Multimedia*, 2006. (page 83, 90, 102)

[191] Jiangming Zhang and Stan Sclaroff. Saliency Detection: A Boolean Map Approach. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2013. (page 82, 83, 101, 103, 105, 110, 111)

[192] Jing Zhang, Wanqing Li, Philip O. Ogunbona, Pichao Wang, and Chang Tang. RGB-D-based action recognition datasets: A survey. *Pattern Recognition*, 60(Supplement C):86 – 105, 2016. (page 5, 15)

[193] Jingjing Zheng and Zhuolin Jiang. Learning view-invariant sparse representations for cross-view action recognition. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2013. (page 60)

[194] Feng Zhou, Sing Bing Kang, and Michael F. Cohen. Time-mapping using space-time saliency. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014. (page 6, 84, 88, 93, 94, 95, 100, 107)

[195] Maryam Ziaeefard and Robert Bergevin. Semantic human activity recognition: A literature review. *Pattern Recognition*, 48(8):2329 – 2345, 2015. (page 5)