Human Action Description

Javier Alexander Montoya Zegarra

HUMAN ACTION DESCRIPTION

BY

JAVIER ALEXANDER MONTOYA ZEGARRA

B.E., San Pablo Catholic University, Arequipa - Peru, 2005
M.S., University of Campinas, Sao Paulo - Brazil, 2007

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Informatique
at ENSIMAG
Institut National Polytechnique Grenoble, 2010

Supervisor

Alexander Kläser: INRIA Rhône-Alpes
Cordelia Schmid: INRIA Rhône-Alpes

Abstract

This master thesis describes a supervised approach to recognize human actions in video sequences. With human action we can understand *human action classification* and *human action localization*. In *human action classification*, the goal is to issue an action class label to sequences with pre-defined temporal extent. In human action localization, the objective is to localize human actions in space (the 2D image region) and in time (temporal extent). In this thesis, we are interested in action classification and localization tasks. The type of video data used are controlled and uncontrolled (realistic) video sequences. In realistic type of video data, human action recognition is especially difficult due to variations in motion, view-point, illumination, camera ego-motion, and partial occlusion of humans. In addition to that, action classes are subject to large intra-class variabilities due to the anthropometric differences across humans. In the case of action localization, the search is computationally expensive due to the large volume of video data. Recently, global representations have shown impressive results for action localization in realistic videos. Nevertheless, these representations have not been applied to repetitive actions, such as running, walking etc. In fact, they even seem not appropriate for this type of actions. Therefore, we propose and evaluate in this work a novel approach that addresses this problem by representing actions as loose collection of movement patterns.

To recognize actions in video, we first detect and track human positions throughout time. We then represent human tracks by a series of overlapping *chunks segments*. These track chunks are processed independently. For our action representation, we extract appearance and motion information with histograms of spatio-temporal gradient orientations. Since chunks with a high affinity to an action of interest will be classified with a larger score, the beginning and end of an action (i.e., action localization) can be determined with clustering approaches. For our method, we employ a variable band-with meanshift.

Our results on publicly available datasets demonstrate and advantage of our method, especially for repetitive type of actions. On realistic data and for non-repetitive actions, we are able to compete with the state-of-the-art results for some action classes and loose for other classes.

Table of Contents

List of Figures

Chapter 1

Introduction

Contents

This master thesis addresses the problem of recognizing and localizing human actions in video sequences. Section 1.1 introduces the problem context and defines the exact goal of this study. Then, section 1.2 gives an overview of related approaches found in the literature. The main contributions of our work are summarized in section 1.3, and section 1.4 gives the outline of this document.

1.1 Context and Goal

Human action recognition is one of the most promising topics in computer vision. In fact, the recognition of actions from video sequences has many applications ranging from animation and synthesis [FAI+05], human-computer interaction [Pen01], behavioral biometrics [SJZ+05] to video surveillance and retrieval tasks [CL00].

The recognition of human movements can be done at different levels of abstraction and different terminoligies have been adopted in the literature to refer to actions. In the present work, we adopt the terminology suggested by Moeslund et al. [BAV06]: action primitives, actions, and activities.

An *action primitive* is an atomic movement that can be described at the limb level and can be recognized without contextual or sequence knowledge. An *action* consists of repetitive action primitives. Finally, an *activity* is a large-scale event that besides being composed of actions also depends on the context of the environment, objects, and human interactions. For instance, "playing tennis" is an *activity*, whilst "run

left/right", "forehand", or "backhand" are examples of tennis *action primitives*. The term *action* is for example used to denote a sequence of action primitives needed to return a ball.

The focus of the present work is the automatic recognition of *human actions* in video sequences. With human action recognition we can understand *human action classification* and *human action localization*. In *human action classification*, a number of action classes is pre-defined and, for each class, training samples (positives and negatives) are given. A classifier is then learned from these training samples. Given a test input video sequence, the objective is to issue a corresponding action class label to the entire video sequence. In other words, the question to be answered here is *if* an action occurs. On the other hand, in *human action localization* the objective is to localize human actions both in space (the 2D image region) and in time (temporal extent). The question to be answered here is therefore *when* and *where* an action happens. However, *human action localization* is an even more difficult task. Especially due to the large volume of video data, an exhaustive search to localize an action is computationally expensive as both, spatial and temporal extents need to be scanned. The question of how to localize actions in an efficient manner is therefore of importance.

The type of video data that is used in this work are both, controlled and uncontrolled, i.e., realistic video sequences. In realistic type of video data, the task of action recognition is especially difficult due to variations in motion, view-point, illumination, camera ego-motion, and partial occlusion of humans induced respectively by camera/perspective and lighting effects. Therefore, robust features that take these variations into account are crucial.

In addition to this, action classes are subject to large intra-class variabilities due to anthropometric differences across humans induced by differences in size, shape, gender, etc. Variations in the execution style and speed of actions further complicate this task. Therefore, the amount of training samples plays an important role. At the same time, the manual annotation of training samples might be time-consuming and tedious due to the large amount of video data.

In this dissertation we focus on action *classification* and *localization* tasks. In the remainder of this work, we will introduce a novel method to represent actions temporally as a loose collection of their *movement patterns*.

1.2 Related Work

In the literature, two main categories for action representation can be distinguished [Ron10]:

1. *Global Representations* encode the visual observation of an action as a whole. First, the human

performing the action is localized. Then, the extracted region of interest is encoded as a whole to obtain the action descriptor.

2. *Local Representations* describe the observation as a set of independent and local features. Usually, spatio-temporal interest points are first detected. Then, local features are computed for a local neighborhood at those points. The action is finally represented as a loose combination of these local features.

1.2.1 Global Representations

Global representations encode body structure and its dynamics as a whole. In these type of approaches, a Region of Interest (ROI) is first obtained usually by background substraction or, human detection and tracking. To characterize global motion and appearance of actions, information derived from silhouettes, edges or optical flow is used. The main assumption of these type of approaches is that the global dynamics of the human body are discriminative enough to recognize human actions.

One of the earliest approaches that uses silhouette information to represent actions is the work of Bobick and Davis [FW01] (See Figure 1.1). They introduce two measurements to extract temporal information from video sequences: Motion Energy Image (MEI) and Motion History Image (MHI). In the binary MEI, silhouettes are extracted from a single view and differences between consecutive frames are aggregated. The MEI indicates therefore where motion occurs. At the same time, MHI are used in combination with MEIs to weight regions that occurred more recently in time such that recent motion has a stronger weight.

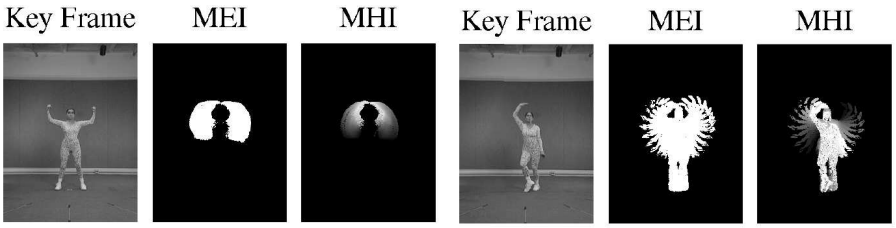


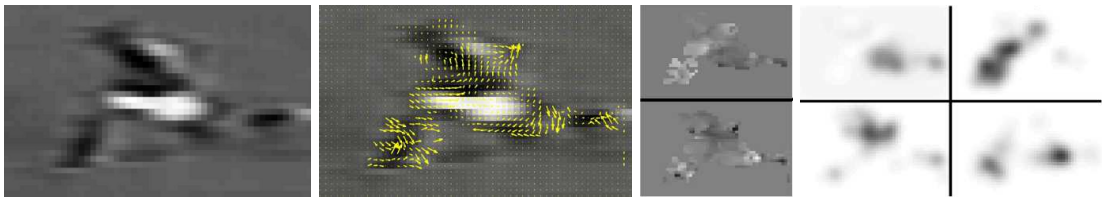
Figure 1.1: Examples of Motion Energy Image (MEI) and Motion History Image (MHI) signatures (courtesy of [FW01]).

Another way to model actions as human silhouettes are Space-time Shapes [BGS+05]. A space-time shape encodes both, spatial and dynamic information of a given human action. More precisely, the spatial information describes the location and orientation of the torso and limbs, whilst the dynamic information represents the global motion of the body and of the limbs relative to the body. Note that human silhouettes are in general computed using background substraction techniques. Blank et al. [BGS+05] propose to

compute local shape properties from the solution of its Poisson equation. Those properties include, for example, local saliency, action dynamics, shape structure, and orientation. A sliding temporal window is then used to extract space-time chunks of 10 frames long with an overlap of 5 frames. Those chunks are then described by a high-dimensional feature vector and are matched against space-time shapes of test-sequences. The authors employ a nearest-neighbor classifier to vote for the associated class. In their work, they introduce the Weizmann dataset for action classification tasks. Examples of Space-time Shapes are depicted in Figure 1.2.

Figure 1.2: Space-time Shapes for "jumping-jack", "walking", and "running" actions (courtesy of [BGS+05]).

Global action representations are often based on optical flow in combination with shape information. An earlier approach to use optical flow for action representation is the work of Efros et al. [EBMM03]. They proposed a system to recognize human actions in low-resolution videos. In a first step, human-centered tracks are obtained from sports footage. For human tracking, they used a correlation based approach. In a second step, motion information is encoded by optical flow. The motion descriptor blurs optical flow and separates horizontal and vertical as well as positive and negative components yielding four different motion channels (See figure 1.3). To classify a human action, the test sequence is aligned to a labeled dataset of example actions. Their method has shown promising results on different sport video datasets such as: ballet, tennis, and soccer video sequences. A drawback of this approach is that they only consider full actions of completely visible people in simple scenarios (i.e., no occlusion and simple backgrounds). In addition, their descriptor only considers motion information.



(a) (b) (c) (d)

Figure 1.3: Action representation using optical flow: (a) input image, (b) optical flow, (c) horizontal and vertical components of optical flow vectors, (d) half-wave rectification and blurred version of each component (courtesy of [EBMM03]).

Jhuang et al. [JSWP07] proposed a biologically-inspired system for action recognition. Their approach is based on a extension of the static object recognition method proposed by Serre et al. [SWB+07] to the spatial-temporal domain. The original form features are replaced by motion-direction features obtained from: gradient-based information, optical flow, and space-time oriented filters. A local max operation is first applied on the filter responses. Then, the responses are mapped to a higher level by comparing them to templates learnt from training videos, followed by a global max operation. These new responses are then fed into a multi-class SVM classifier to determine the action class.

Schindler et al. [SG08] proposed a method that combines both motion and appearance information to characterize human actions. For each frame, appearance information is extracted from the responses of Gabor filters, whilst motion information is extracted from optical flow filters. Both types of information are then weighted and concatenated. Principal Component Analysis is used to recude the dimensionality of the feature vectors. Finally, a linear multi-class SVM estimates the final class label for a given test sequence.

A different way to model human actions was proposed by Ali et al. [ABS07]. In their approach, they use concepts from the theory of chaotic systems to model and analyze non-linear dynamics of human actions. Basically, the main joints of 2D stick figures are connected as joint trajectories. Then, these trajectories are represented by chaotic invariants. Classification is done with k-nearest-neighbors based on a training set of trajectory templates. Figure 1.4 depicts two stick figure representations for two action classes: jump and sit. Stick figures with green joints correspond to the first frame of the sequence, whilst stick figures with blue joints correspond to the last sequence frame. In their approach, the joint positions are manually annotated.

Two promising global approaches for action localization in realistic video have been proposed by Laptev et al. [LP07] and Kläser et al [KMSZ10]. Both methods propose an initial filtering to identify possible action localizations and to reduce the computational complexity. To avoid an exhaustive spatio-temporal search for

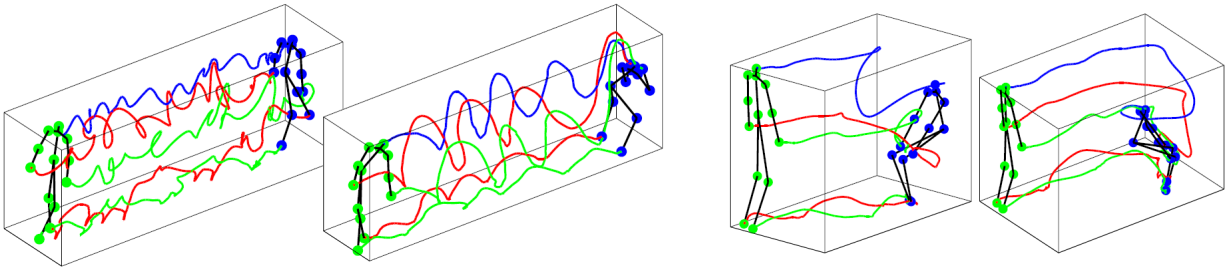


Figure 1.4: Stick figure representations for the *jump* and *sit* action classes. Stick figures with green joints denote the beginning of the action, whilst stick figures with blue joints denote the ending of the action. (courtesy of [ABS07]).

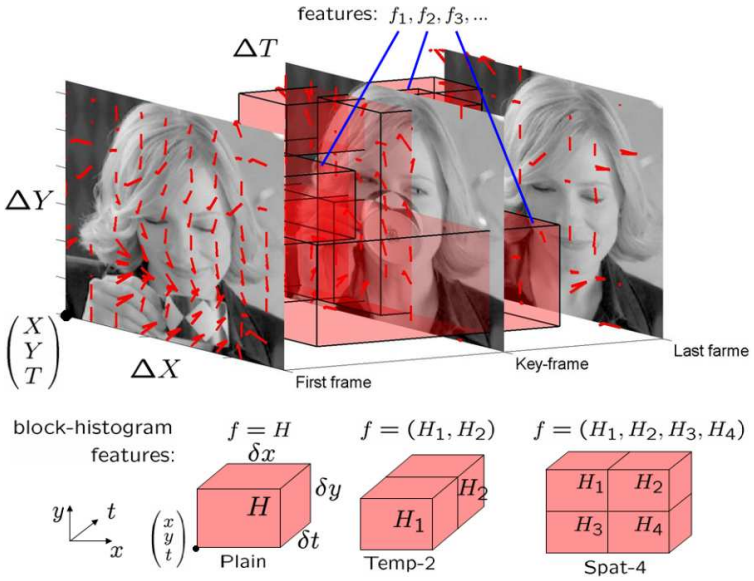


Figure 1.5: (top) Appearance and motion features for a sample of drinking action; (bottom) different spatial and temporal layouts for action representation (courtesy of [LP07]).

localizing actions, Laptev et al. [LP07] use a human key-pose detector trained on keyframes of an action. In a second step, action hypotheses are generated and represented as cuboids with different temporal extents and aligned to the detected keyframes. The cuboid region is represented by a set of appearance (histograms of oriented spatial gradients) and motion (histograms of optical flow) features which are learned in an AdaBoost classification scheme. These features can be organized in different spatial and temporal layouts within the cuboid search window. Figure 1.5 illustrates the appearance and motion features used to represent a *drinking* action sample.

Kläser et al. [KMSZ10] propose as generic pre-filtering approach to detect and track humans in video sequences. Action localization is done with a temporal sliding window classifier on the human tracks. For the desription of actions, the authors introduce a spatio-temporal extension of histograms of oriented gradients

(HOG) [DT05], which extracts appearance and motion information.

1.2.2 Local Features based Representations

Local features based representations encode a video as a collection of local feature descriptors or patches. For this type of representation, patches are first sampled either densely or at salient space-time positions. Local descriptors encode appearance and/or motion information in the local neighborhood of a given space-time position.

Feature detectors

Feature detectors localize salient points in video at which characteristic motion changes occur. The main assumption is that salient locations carry sufficient information to discriminate between different actions.

One of the earliest work was proposed by Laptev et al. [LL03]. In this approach, the authors extend the Harris corner detector to 3D. The space-time interest points correspond to local regions with a significant change in both, the spatial and the temporal domain. For this, the eigenvalues of a spatio-temporal second-moment matrix are computed at each space-time position. Then, local maxima over the combined eigenvalues determine the final points of interest. Figure 1.6 depicts interest points detected for a person walking.

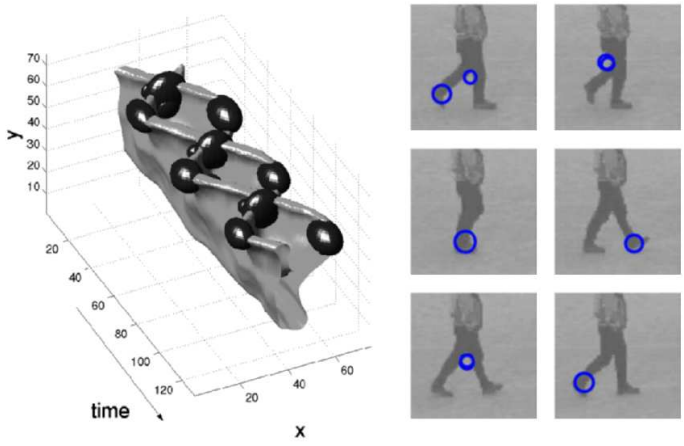


Figure 1.6: Spatio-temporal interest points from the motion of the legs of a walking person. 3D plot of a leg pattern (upside down) and the detected local interest points (left image). Spatio-temporal interest points overlayed on single frames in the original sequence (courtesy of [LL03]).

A different feature detector has been proposed by Dollar et al. [DRCB05]. The authors use spatially a Gaussian filter and temporally Gabor filters. As in the previous method, local maxima over the response function of Gaussian and Gabor filters determines the final interest points. In addition, the authors also

suggest that the number of interest points detected by Harris 3D can be insufficient. In their own approach, the number of interest points is adjusted by the spatial and temporal filter size.

Another extension from images to the video domain is the work proposed by Willems et al. [WTG08]. In their work, the authors define saliency as a scale-normalized determinant of the 3D Hessian Matrix. As a result, interest points are detected at different spatial and temporal scales. To speed up computations of scale-spaces, box-filters are used in combination with an integral video structure. Examples of interest points detected in human actions are depicted in Figure 1.7. By adjusting parameters, the density of the features can be controlled between sparse (first and third subimages) and dense detections (second and fourth subimages).

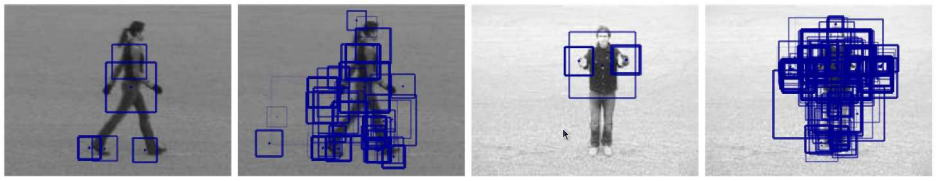


Figure 1.7: Spatio-temporal interest points when using the determinant of the Hessian as salient measure. Detections can be very sparse (first and third subimages) and very dense (second and fourth subimages) (courtesy of [WTG08]).

Feature descriptors

Once a set of sample locations has been determined for a video sequence, feature descriptors are used to capture discriminative information in their local neighborhood. At the same time, the spatial and temporal size of a patch is in general determined by the feature detector, as well.

In an earlier approach, Dollar et al. [DRCB05] evaluate different local space-time descriptors based on normalized brightness, gradient, or optical information. To account for variations in appearance-, motion-, and lighting-invariant representations, the authors extend their descriptors by considering: simple concatenation of pixel values (flattening), grid of local histograms, and a simple global histogram. To reduce the dimensionality of each descriptor, Principal Component Analysis (PCA [HTF01]) is used.

As a feature descriptor based on 3D gradients, Scovanner et al. [SAS07] proposed an extension of the SIFT descriptor [Low04] to 3D. A set of random sampled points at different locations, times, and scales are first selected. Then, spatio-temporal gradients are computed in the regions around the interest points. Each pixel in the spatio-temporal patch votes into a $M \times M \times M$ grid of histograms of oriented gradients (e.g. $2 \times 2 \times 2$ or $4 \times 4 \times 4$), where M corresponds to the number of bins in the spatial and temporal domain.

To quantize the orientation, gradients are represented in spherical coordinates ϕ, θ and divided into 8×4 histograms. Furthermore, the 3D patch surrounding an interest point is aligned to its dominant orientation ($\phi = \theta = 0$). The length of the descriptor depends on the number of bins used to represent ϕ and θ, and also on the number of sub-histograms considered.

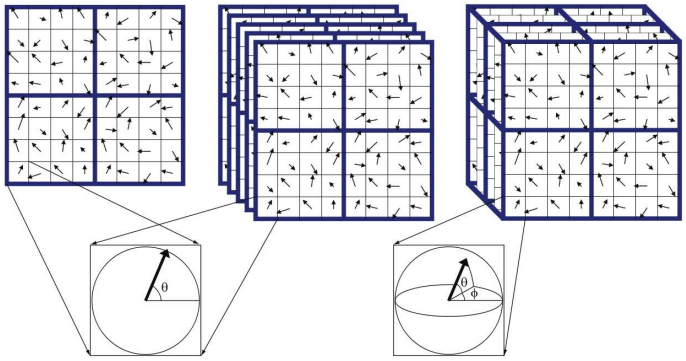


Figure 1.8: 2D Sift descriptors applied straightforward to video sequences (two images on the left). 3D Sift descriptor applied to sub-volumes, each sub-volume is accumulated into its own sub-histogram. The concatenation of such histograms generate the final descriptor (courtesy of [SAS07]).

A related descriptor that uses spatio-temporal Haar filter response was introduced by Willems et al. [WTG08]. The authors extended the SURF descriptor [BETG08] to 3D. Each temporal patch around an interest point is divided into $M \times M \times N$ bins, where M represents the number of bins in the spatial domain and N in the temporal domain. Each cell is then represented by the weighted sum of the responses of 3D Haar Wavelets. Rotation-invariance can also be achieved by considering the dominant orientation of the local neighborhood.

Another recent approach that proposes a video descriptor is the work of Kläser et al. [KMS08]. Their approach is based on Histograms of Oriented Gradients (HOG) [DT05], which have been succesfully used in images. A spatio-temporal patch is divided into a sef of $M \times M \times N$ cells (M and N denote respectively the number of bins in the spatial and temporal domain). For each cell, an orientation histogram is computed, where the orientation is quantized using regular polyhedrons. The final feature vector consists of all cell histograms normalized separately and then concatenated.

Bag of Features

A popular approach that uses local features is the Bag-of-Features (BoF) model. It was originally proposed for representing textual data [Sal68], but has been extended for visual recognition tasks [SZ03], [CDF+04]. For video data, feature detectors first localize salient spatio-temporal points. Then, feature descriptors

capture discriminative information in the local neighborhood of the salient points. For a compact representation, a *visual vocabulary* is computed through the clustering of feature descriptors of training sequences. Each cluster point is referred to as a *visual word*. To represent a video sequence, the corresponding feature descriptors are quantized to their closest visual word and a histogram of ocurrences of visual words is computed. Following a supervised learning approach (e.g., Support Vector Machines), the histograms are used to robustly classify the entire sequence. Since a video is represented as an orderless collection of local salient features, geometric and temporal information about the relations between descriptors are discarded.

An earlier approach that uses a BoF representation for human action recognition is the work of Niebles et al. [NL07]. In their approach, a hierarchical representation of human actions is proposed. At a top layer, a constellation model of P parts is defined. Each part is then associated to a BoF model in a lower layer and parts are related to each other by a distribution of their relative positions. Figure 1.9 ilustrates a four part model for an action sample of class *hand waving*. Parts are represented as colored ellipses and the associated features are colored according to its part membership. Static features are represented by crosses, whilst motion features by diamonds. They only use simplistic video data with static and homogeneous background information.

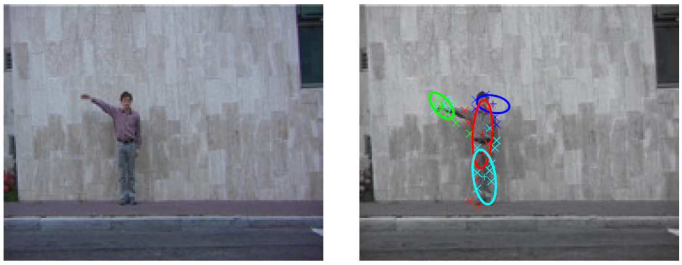


Figure 1.9: A four part constellation model for the *hand waving* action. Ellipses correspond to the parts of the constellation model and associated features are colored according ot its part membership (courtesy of [NL07]).

More recent approaches have used an action voting scheme to localize actions spatially and temporally. For instance, Willems et al. [WBTG09] consider local features assigned to the strongest visual codebook entries learned from training data to generate action hypotheses. To remove weak generated hypotheses, a prunning step is applied. The remaining hypotheses are evaluated by a non-linear X^2 SVM that assigns a confidence value to the action hypotheses. Non-maxima suppression is then applied in the voting space of the hypotheses to obtain final action hypotheses. In their work, local features are found using the Hessian 3D detector [WTG08]. As descriptors, a HOG3D [KMS08] and eSURF [WTG08] are employed. Examples

of action hypotheses and the final detection of drinking actions are show in figure 1.10.

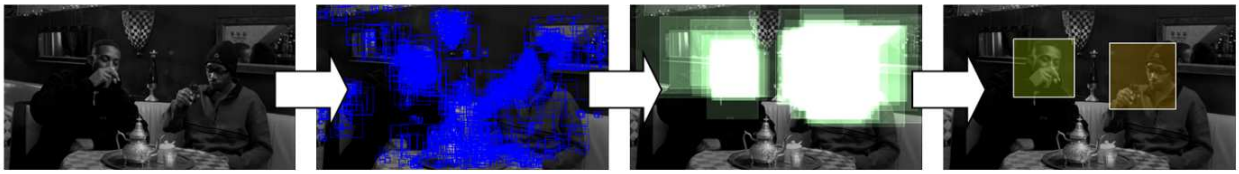


Figure 1.10: Action localization of drinking actions following a voting scheme (courtesy of [WTG08]).

A recent study on how well local space-time features (both localization and description) perform for action recognition tasks was carried out by Wang et al. [WUK+09]. In this study, four common feature detectors and six local feature descriptors have been used in combination with a bag-of-features SVM approach for action recognition. The authors concluded that the Harris 3D [LL03] and Cuboid [DRCB05] detectors are a good choice for interest point detectors. Furthermore, feature descriptors based on the combination of gradient and optical flow information achieved the best performance.

1.3 Contributions

The goal of this work is the classification and localization of human actions in video sequences. Contributions of our work are given for both, controlled and uncontrolled (realistic) video sequences. Summarizing, we provide the following main contributions:

- We introduce a novel way to represent actions temporally as a loose collection of movement patterns. To achieve this, we first detect and track humans throughout time. Then, we represent human tracks by a sequence of overlapping track segments, so-called *chunks*, which are extracted at fixed time steps. We process chunks independently and extract appearance and motion information by using a spatio-temporal descriptor adapted to human tracks.

- We evaluate the proposed method for action recognition tasks, investigate in depth its parameters, and compare to the current state-of-the-art. Experimental results over three publicly available datasets with varying difficulty and for a total of 14 different action classes are given.

The most related approaches to our work are the methods proposed by Laptev et al. [LP07] and Kläser et al. [KMSZ10] (see subsection 1.2.1). As in the work of Kläser et al., we first consider an initial filtering step based on human tracks to avoid an exhaustive spatio-temporal search of possible action localizations. One major benefit of human tracks is that they can be reused for any type of action. A main drawback

of both approaches is that actions are represented globally, i.e., as a fixed sequence of primitive motion events. However, actions can also consist of repetitive motion patterns which cannot be captured in a suitable manner with a global representation. In our approach, we address this shortcoming with an action representation based on a loose collection of movement patterns. As a consequence, we are able to: (*i*) better model re-ocurring action patterns and (*ii*) robustly represent actions that do not follow a clear sequential pattern, yet contain key movements.

1.4 Outline

This Master Thesis is organized as follows. Chapter 2 introduces the proposed system to classify and localize human actions in video sequences. Chapter 3 presents the experimental results obtained on standard datasets for human action classification and localization. Finally, chapter 4 concludes this work.

Chapter 2

Human Action Description

Contents

This chapter introduces our supervised approach to action recognition. It can be divided into three main parts. The first part (section 2.1) of our system detects and tracks humans in a given input video sequence.

In the second part (section 2.2), we represent each track by a series of overlapping *chunks*.

Finally, in the third part (sections 2.3, 2.4), a learned classifier evaluates each *chunk* on the track, and a voting scheme determines when an action is exactly happening.

2.1 Human Tracks

Our approach for detecting and tracking humans follows the work of Kläser et al. [KMSZ10]. We obtain human tracks by first detecting humans in every video frame. These detections are then linked together with a general purpose tracker. Since some human detections might be missing, detections are interpolated, and a final track classifier helps to improve the overall detection precision by eliminating false positive detections.

To automatically detect human beings in frames of video sequences, we use the pedestrian detector proposed by Dalal et al. [DT05]. The detector is applied to all frames of the video input sequence. The output is a set of temporal bounding boxes of detected humans. Depending on the type of video data, we employ in our work either a full-body pedestrian detector [DT05] or a detector specifically trained for upper bodies [KMSZ10].

Human detections of neighboring frames are linked together using a tracking-by-detection approach [ESZ09]. For this, interest points are localized in the region of interest of a given detection and are propagated to sucessive frames by using the Kanade-Lucas-Tomasi (KLT) tracker [ST94]. Points that cannot be tracked

reliably are discarded and replaced with new points. The number of point tracks passing between two detections determines in an agglomerative clustering scheme detections corresponding to the same person.

For action recognition, continuous tracks of human actions are necessary. Since some humans might not be detected in some frames of a sequence and in order to stabilize detections over time, we employ an interpolation and smoothing step. Estimated detection windows are optimized over track parameters $\{\mathbf{p}_t\}$ [KMSZ10]:

$$\min_{\{\mathbf{p}_t\}} \sum_{t \in T} \left(||\mathbf{p}_t - \bar{\mathbf{p}}_t||^2 + \lambda^2 ||\mathbf{p}_t - \mathbf{p}_{t+1}||^2 \right) \tag{2.1}$$

where the parameter $p_t = (x_t, y_t, w_t, h_t)$ denotes respectively the spatial position, width, and height of a bounding box at time t for a track T. In addition, $\bar{p}_t = (\bar{x}_t, \bar{y}_t, \bar{w}_t, \bar{h}_t)$ represents the detections and λ is a temporal smooting parameter. Whenever a detection $\bar{p}_t$ is missed, the related term is removed from the cost function for that given frame. Optimizing equation (2.1) results in a linear equation with a tri-diagonal matrix. This matrix can be solved efficiently and for each term x, y, w, h independently by Gaussian elimination with partial pivoting.

Human detections can be erroneous due to background clutter. Since the human detection is a crucial pre-filtering step for our action recognition system, a post-processing step helps to further reduce the number of false positive tracks [KMSZ10]. For this, a track is represented by a 12-dimensional feature vector based on: (*i*) human SVM detection score (false detections usually have lower score than true detections), (*ii*) track length (false detections are frequently shorter), (*iii*) variability on the scale and position of the detections (artificial detections are not stable), and (*iv*) occlusion by other tracks. For each of these four measures, statistics based on min, max, and average are computed to form the 12-dimensional feature vector. An SVM classifier is then able to significantly improve the precision rated of correctly detected human tracks.

2.2 Human Action Representation

Given human tracks for a video sequence, our objective is to determine which of those tracks contain an action and also to localize the action within the track. In addition, we seek an action representation that is able to cope with repetetive type of actions. Due to the tracks, the spatial extent of an action is already defined, and only its temporal extent (i.e., beginning and end) needs to be determined. This allow us to reduce significantly the the search complexity for actions. In our approach, we localize actions by representing each track as a series of overlapping track-chunks. Chunks have a pre-defined temporal extent (*length*) and are extracted at a fixed time interval (*stride*). To obtain a vector representation for each chunk, we apply

a spatio-temporal window descriptor denoted as HOG-Track descriptor [KMSZ10]. The chunk descriptors are then evaluated with a classifier which returns a probability score of a chunk belonging to the action of interest. To generate action hypotheses, we cluster chunks together according to their computed scores. The duration of an action is then delimited by the action chunks assigned to it. With this representation, we are able to effectively model an action temporally as a loose collection of movement patterns.

HOG-Track Descriptor The HOG-Track descriptor has been introduced by Kläser et al. [KMSZ10]. Given a track segment, the descriptor encodes appearance and motion information using histogram of oriented spatio-temporal gradients. The track segment is first divided into n_t equally long temporal segments that are aligned with a human track. The spatial region corresponding to each of the segments is given by the bounding box of their centre frame. Each segment is split into a spatial grid of $n_x \times n_x$ cuboid cells with a spatial overlap of 50% as illustrated in figure 2.2. A histogram of spatio-temporal (3D) gradient orientations encodes the content of each cell. Using an icosahedron, the orientation of the spatio-temporal gradients is quantized into a histogram of 20 orientations. Each side of the polygon corresponds to a histogram bin. Opposing directions are associated with the same orientation bin halfing the number of bins to 10. Finally, each gradient votes with its magnitude into its closest bins using interpolation. Besides icosahedrons, other regular polyhedrons with congruent faces can also be used to quantize the spatio-temporal (3D) gradient orientations. They are namely: tetrahedron (4-sides), cube (6-sides), octahedron (8-sides), dodecahedron (12-sides). In our experiments, we consider the icosahedron since it results in the largest number of orientation bins. Figure 2.1 illustrates the different existing regular polyhedrons.

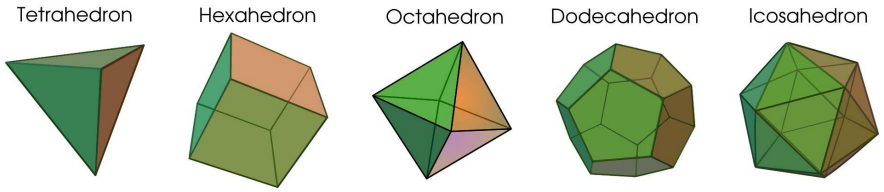


Figure 2.1: Illustration of the the different existing regular polyhedrons (courtesy of Wikipedia).

For the final descriptor, cell histograms are concatenated, and all cells of a temporal segment are normalized with their $L2$-norm. The dimensionality of the resulting descriptor is given by: 10 orientation bins $\times\ n_x^2$ cuboid cells $\times\ n_t$ temporal slices $= 10 \times n_x^2 \times n_t$.

In our approach, we employ the HOG-Track descriptor as representation for our defined action chunks. For each chunk, we extract appearance and motion information using the spatio-temporal HOG-Track descriptor. For the experiments, we fix the number of spatial cells of the HOG-Track descriptor to $n_x = 5$, the

value optimized by Kläser et al. [KMSZ10]. For the number of temporal division (n_t), we evaluate different values. Additional parameters of our approach include the *chunk length* and the *stride length* at which chunks are extracted. In our evaluations, we consider different values and investigate their influence on the final performance of our full approach.

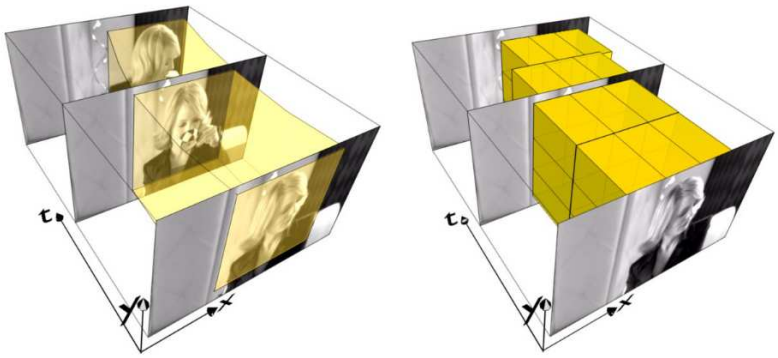


Figure 2.2: HOG-Track descriptor: (left) an example of an upper body detection; (right) *track* division into temporal slices. Each temporal slice is further divided into a spatial grid of cuboid cells (courtesy of Kläser et al. [KMSZ10]).

2.3 Action Classification

For action classification, video sequences are given as positive and negative samples which contain a particular action or not. For a positive sequence, we consider all chunks on a track as positive samples. Additional positive samples are obtained by jittering the initial positives in time and by jittering the bounding box size of the track. Chunk descriptors from sequences that do not contain the action of interest are considered as negative samples. All positive and negative chunk samples are encoded as feature vectors by computing their HOG-Track representation. Given a trained classifier, all chunk descriptors of a track are classified and their scores are summed up and averaged. The action label of the classifier that maximizes the average score is then assigned to the test video sequence.

To evaluate the discriminative properties of our method to classify actions, we conduct experiments on 9 different type of actions. Comparison to state-of-the art show that our loose action model proofs beneficial for repetitive types of actions.

2.4 Action Localization

Our loose action representation motion patterns that are typical for a given action. Since we model each movement part of the particular action independently, we can assume that chunks that overlap within the action will be classified with a higher score. Therefore, we can localize an action by grouping together chunks with high classification scores.

To localize actions, we first represent tracks as a sequence of chunks and encode each chunk with the HOG-Track descriptor. As in the case of classification, feature vectors are evaluated with a trained classifier and a score is obtained for each chunk vector. To generate final action hypotheses—the duration of an action will in general enclose several action chunks—, we cluster chunks according to their scores into a set of non overlapping action hypotheses. In this work, we consider three clustering strategies: a simplistic *Multi-level* and *Peak Threshold* clusterings as baseline and a more principled clustering approach using *Mean-shift*.

Multi-level clustering In the Multi-level clustering approach (MTC), several tresholds are applied iteratively over the chunk scores of a given track. All neighbor chunks having a score higher than γ are clustered together and a hypothesis is generated, where γ represents a treshold value. In our experiments we set the value of $\gamma = 1.0$ and iteratively decrease it at steps of 0.1. Processed chunks are still used from the further clustering process. The length of a generated hypothesis is found by considering the number of frames between the start frame of its first and the last frame of its last chunks. In addition, the final score of a detection is given by the average of the clustered chunks scores.

Peak Treshold clustering In the Peak Treshold clustering approach (PTC), hypotheses are generated by iteratively clustering neighbor chunks of score peaks. More precisely, the chunk p_i on a track with the highest global score s_i is first found. If its direct neighbors have scores higher than $s_i \cdot \gamma$, they are clustered together and a hypothesis is generated. The threshold γ controls the sensitivity of the clustering approach and is set to $\gamma = 0.75$ in our experiments. Processed chunks are then removed from the further clustering process. This process is done iteratively until all chunks have been processed. The length of a generated hypothesis is found by considering the number of frames between the start frame of its first and the last frame of its last chunks. The final score of the detection is then the average of the clustered chunk scores.

Mean-Shift clustering The Mean-Shift clustering [CM02] is a non-parametric clustering technique that automatically finds the number of significant clusters following a recursive mode detection procedure. Modes are detected by first finding stationary points of the underlying probability density function. Then, these points are pruned by retaining only the local maxima of the density. Points associated to the same modes

define a basin of attraction and points that lie in the same basin are clustered together. The clusters are therefore delimited by the boundaries of the basis.

To evaluate our action localization approach, we employ two different movie datasets with varying difficulties. For each dataset, we evaluate systematically the relevant parameters of our approach and compare our different localization strategies. Results show that *Mean-Shift* improves the localization of actions when compared to *Multi-level* and *Peak Treshold* clustering approaches. Comparisons with state-of-the-art show the promising properties of our method for localizing actions.

Chapter 3

Experimental Results

Contents

This chapter presents the experimental setup and results achieved in our study. First, the action recognition datasets for action classification and localization employed in this work are described (section 3.1). Along with dataset descriptions, we detail how ground-truth and automatic tracks are obtained. Second, we present experimental results for classification (section 3.2) and localization (section 3.3), and also compare with the current state-of-the-art.

3.1 Datasets

3.1.1 Weizmann

We employ the Weizmann dataset [BGS+05] to evaluate the classification performance of our approach. The dataset consists of nine different subjects performing nine types of human actions: *bending down*, *jumping jack*, *jumping*, *jumping in place*, *galloping sideways*, *running*, *walking*, *waving one hand*, and *waving with*

both hands. Some examples of these 9 classes are illustrated in figure 3.1. In total, there are 81 sequences, one sequence per action and person. In each sequence, only one particular action is performed. The average length of the sequences is 2.4 seconds. Furthermore, the video sequences have a spatial resolution of 180×144 pixels and were recorded at 25 fps with a static camera.

Figure 3.1: Sample frames from the Weizmann action dataset.

In order to obtain spatial annotations for the sequences, we use the foreground masks available with the dataset. We convert the masks to tracks by finding for each frame the minimum rectangle enclosing the foreground actor mask. Figure 3.2 shows some of the extracted ground trouth tracks as yellow bounding boxes.

Automatic tracks

In order to obtain automatic tracks, we employ the pedestrian detector of Dalal et al. [DT05] which has been trained on upright pedestrians. The detector is applied to each frame of a sequence, and its detections are linked smoothed for our final tracks in section 2.1. In order to identify correct and erroneous tracks during learning, the automatically computed tracks are matched to ground truth tracks obtained from foreground masks. We consider tracks that match a ground truth track with an overlap $O(X, Y) \geq 0.5$ as positive

training samples. Given an annotation Y and a track X, the overlap between them is given by [LP07]:

$$O(X, Y) = (X \cap Y)/(X \cup Y) \tag{3.1}$$

During action classification, we evaluate all the tracks extracted from a test video sequence, correct tracks as well as erroneous tracks (see figure 3.2). For each track, all scores of the corresponding chunks are summed up and averaged. The action label of the action classifier that maximizes the highest averaged track score is then assigned to the test video sequence. Results of automatically computed human tracks are shown as green bounding boxes in figure 3.2. Some of the tracks correspond to a human performing an action, whilst others are associated with background. Note that automatically computed tracks match well the ground truth tracks. However, note that human detections for actions like *bending down* are difficult, since the detector was trained for upright pedestrians.

Figure 3.2: Sample ground truth and automatically obtained tracks for the Weizmann action dataset.

3.1.2 Coffee and Cigarettes

The Coffee and Cigarettes (*C&C*) dataset comprises *drinking* and *smoking* actions from three different video sources, i.e., the *Coffee & Cigarettes* and *Sea Of Love* movies, as well as, manually recorded *drinking*

sequences. The *C&C* movie has 11 short stories, each with different scenes and actors.

Initially, the *C&C* dataset has been used for detecting *drinking* actions [LP07], yet recently, also results on *smoking* actions have been reported [KMSZ10]. The annotation data for *drinking* actions consist of 38 test samples from two *C&C* short stories. The training samples consist of 41 drinking sequences obtained from six short *C&C* stories. Additionally, 32 samples from the *Sea Of Love* movie and 33 drinking sequences recorded at a laboratory are also considered. This results into 106 training samples and 38 testing samples. The total time of test video data for *drinking* actions is about 24 minutes.

The annotation data for *smoking* actions consists of 42 test samples obtained from three *C&C* short stories. The *smoking* training set contains 78 samples: 70 from six *C&C* short stories (the same ones used for *drinking* training samples) and 8 samples from the Sea Of Love movie. The total time of test video data for *smoking* actions is about 21 minutes.

An annotated action is given by a spatio-temporal cuboid that determines its localization in space and time. The temporal extent of an action is given by the corresponding start and end frames. The spatial extent is given by a 2D bounding box at a predefined keyframe located in the middle of the action.

To determine if an action is correctly detected, we follow the protocol proposed by Laptev et al. [LP07]. Given a ground truth cuboid Y and a track segment X, the overlap between them is given by $O(X, Y) = (X \cap Y)/(X \cup Y)$. An action is then correctly detected, if the overlap is $O(X, Y) \geq 0.2$. Once an annotated sample has been detected, any further detection for that annotation is counted as a false positive.

Automatic tracks

On the *C&C* dataset, humans are usually visible only with their upper body part. Therefore, we employ the detector of Dalal et al. [DT05] but trained for upper bodies as proposed by Kläser et al. [KMSZ10]. The upper body detector is trained in *Hollywood-Localization* training movies by extracting positive and negative windows. Human heads are annotated in keyframes and are automatically extended to their upper bodies. Each annotated window is then jittered and flipped horizontally. The negative training windows are randomly sampled guaranteeing that there is no significant overlap with positive annotations. The training data consists of $30k$ positive windows and $55k$ negative windows. To improve the precision of the upper body detections, a retraining stage is followed as proposed by Dalal et al. [DT05] in which hard negatives are used for retraining the detector. The upper body detector is then applied on the *C&C* dataset. To obtain the final tracks, we employ the tracking and smoothing approach as presented in section 2.1. Figure 3.3 depicts some samples of *drinking* and *smoking* actions with their corresponding upper body detections.

Figure 3.3: Upper body detections for *drinking* (left column) and *smoking* (right column) actions.

3.1.3 Hollywood Localization

To evaluate the performance of our approach to localize actions on realistic video sequences, we use the Hollywood-Localization dataset introduced by Kläser et al. [KMSZ10]. The dataset contains *Answering Phone* and *Standing Up* actions extracted from Hollywood movies. The annotation data contains 130 clips for *Answering Phone* actions and 278 clips for *Standing Up* actions. In both cases, the same number of randomly selected clips not containing the action are used as negatives. Training and testing samples are

selected by roughly dividing the samples into two halves. In total, there is about 17.5 minutes of testing data for *Answering Phone* actions and 39 minutes for testing *Standing Up* actions.

In our experiments, we follow the evaluation protocol proposed by Kläser et al. [KMSZ10] to determine the overlap between an action hypothesis and a ground truth annotation. The ground truth data specifies an action by its temporal extent (i.e. start and end frames) and by its spatial localization (i.e. a rectangle for one of the intermediate frames). Given an annotation Y and a track X, the overlap in time between them is given by $O_t(X,Y) = O(X_t, Y_t)$, where Y_t and X_t denote their corresponding temporal extents. Furthermore, the overlap in space is given by $O_s(X,Y) = O(X_s, Y_s)$, where Y_s and X_s represent the corresponding spatial rectangles in the annotated action frame. The final overlap is then defined as $O^{'}(X,Y) = O_t(X,Y) \times O_s(X,Y)$. An action is detected correctly if there is an overlap between the ground truth annotation and an action hypothesis of $O^{'}(X,Y) \geq 0.2$.

Since in the Hollywood Localization dataset, humans are usually visible only with their upper body part as in the *C&C* dataset, we follow the same procedure to obtain human tracks. Some examples of upper body detections are ilustrated in figure 3.4.

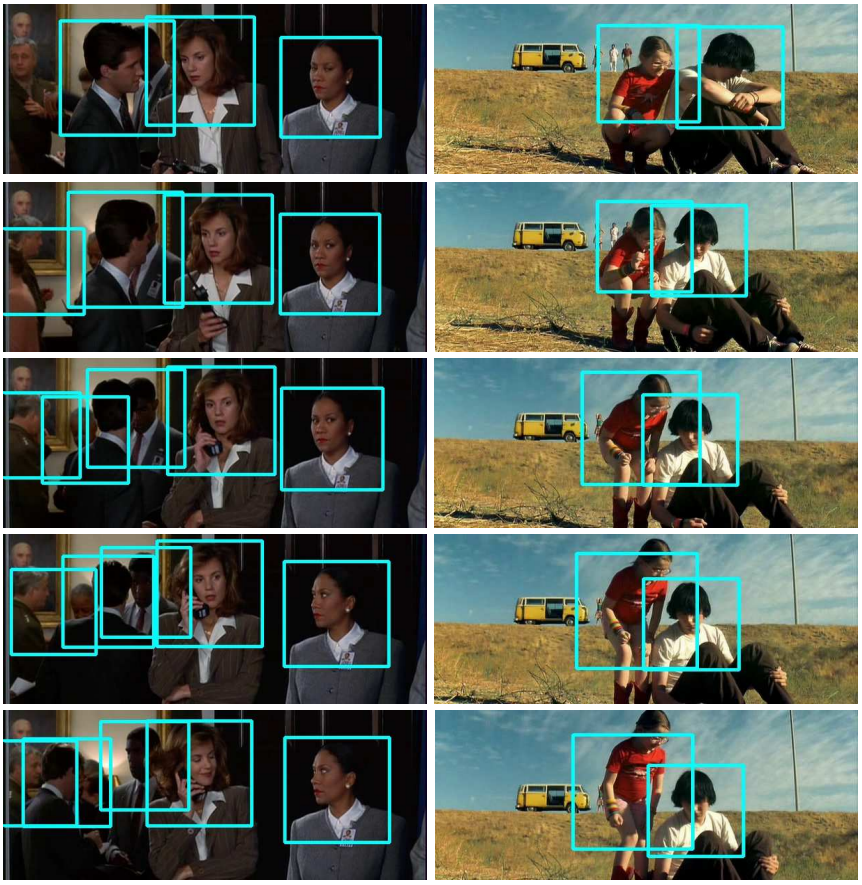


Figure 3.4: Upper body detections for *answer phone* (left column) and *stand up* (right column) actions.

3.2 Action Classification

Our goal is to quantify the improvement in human action recognition when representing an action as a loose collection of movement patterns. For a first experiment, we apply our action representation to the task of action classification on the Weizmann dataset. The performance is evaluated in terms of classification accuracy, i.e., the number of correctly classified sample sequences divided by the total number of samples.

In the following, we first compare our loose approach to baseline using a global representation (section 3.2.1). In section 3.2.2, we conduct an extensive analysis of different configuration parameters for our approach. Finally, we compare results with other action classification approaches (section 3.2.3).

For all experiments presented in this section, we employ a linear SVM classifier for each action class following a leave-one-out evaluation protocol: 8 subjects are used for training and the remaining one for testing. The same procedure is repeated for all 9 permutations. Since experimental results using a non-linear

SVM classifier are comparable, we report results only for a linear SVM classifier.

3.2.1 Baseline

As a baseline, we employ a global action representation similar to Kläser et al. [KMSZ10] and extract a HOG-Track descriptor for the entire track. By doing so, we model actions as one single chunk where the chunk length is equal to the temporal extent of the track, i.e. the temporal extent of the action is delimited by the start and end frames of the track. The spatial extent of the action is already fixed by the human detections. In the HOG-Track descriptor, we fixed the number of spatial cells to the value optimized by Kläser et al. [KMSZ10] of $n_x = 5$. In contrast, for optimal performance, we investigate different numbers of temporal slices (n_t). Figure 3.5 ilustrates the average performance achieved over four independent runs. It can be seen that the baseline representation is able to classify actions up to an accuracy of 83.74%. The results show that the classification accuracy improves as the number n_t of temporal slices increase and saturate for a value of $n_t = 3$. This can be explained since more slices are able to capture more exactly the motion in a video sequence.

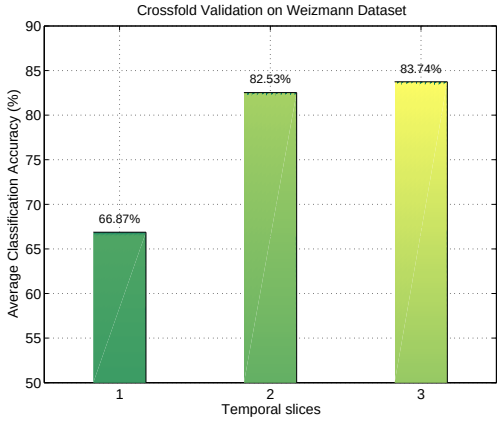


Figure 3.5: Classification accuracy on the Weizmann Dataset for the baseline approach. Actions are modeled with one global descriptor [KMSZ10]. Results using different number of temporal slices (n_t) are illustrated.

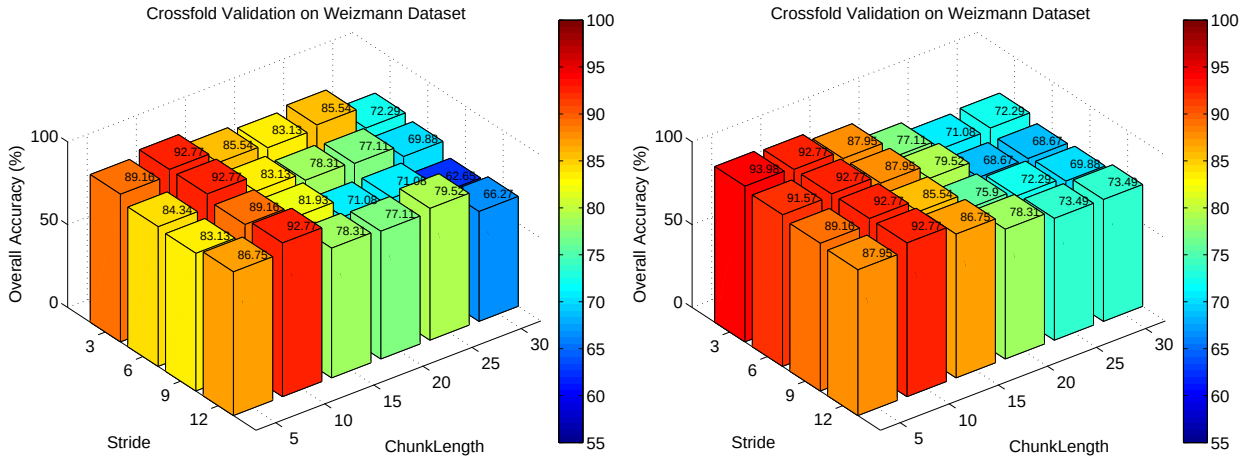
3.2.2 Parameter Evaluation

In this section, we evaluate the gain of modeling actions as a loose collection of movement patterns, i.e., tracks are processed by overlapping chunk descriptors which are classified independently. Since chunks have a pre-defined temporal extent (*length*) and are extracted at a fixed time interval (*stride*), we evaluate systematically different parameters. In order to focus the evaluation on the temopral parameters of our

system, we fix the spatial grid of the HOG-Track descriptor to the optimal value of $n_x = 5 \times 5$ found by Kläser et al [KMSZ10].

First, we fix the number of temporal slices for the HOG-Track descriptor. Then, we evaluate different combinations of *stride* and chunk *length* parameters. For each possible combination of those parameters, a full cross-fold validation on the Weizmann dataset is done. The average of the classification accuracy is plotted as a single 3D bar.

Figure 3.6(a) and 3.6(b) illustrate the performance of our system for different numbers of temporal slices, $n_t = 1$ and $n_t = 2$, respectively. Since results for both settings achieve a similar performance level, we do not consider other values for temporal slices. From both figures some important observations can be drawn. First, regardless of the number of temporal divisions (n_t), our method benefits from combining shorter chunk lengths with smaller stride values. This is mainly because short chunks are able to better capture re-occurring motion patterns. From the video samples in the dataset, it can be verified that the cycle length at which actions are repeated lies approximately within 5 to 12 frames. Second, the temporal divisions of chunks allows us to obtain a more robust action representation since more temporal information is encoded. Larger stride values in combination with longer chunks degrade the performance as can be seen in both figures. Nevertheless, the parameter setting that yields lowest performance is still able to classify about 70% of the sequences correctly.



(a) Temporal divisions ($n_t = 1$)

(b) Temporal divisions ($n_t = 2$)

Figure 3.6: Accuracy plots for the Weizmann action dataset: different *stride* and chunk *length* parameters are considered to represent actions as a loose collection of movement patterns.

Table 3.1, bottom row, presents the confusion matrix for our baseline representation, which models action

sequences with one global descriptor. We plot the results obtained from the best parameters obtained earlier. In the bottom row, we show the confusion matrix for our loose representation with parameters $n_t = 2$, stride set to 3 frames, and chunk length equal to 5 frames. Note that for both cases the main source of confusion is related to the action "jump" with "run" or "walk" actions. This is basically due to the visual similarity among these action classes. One can see that the overall confusion is reduced for our method.

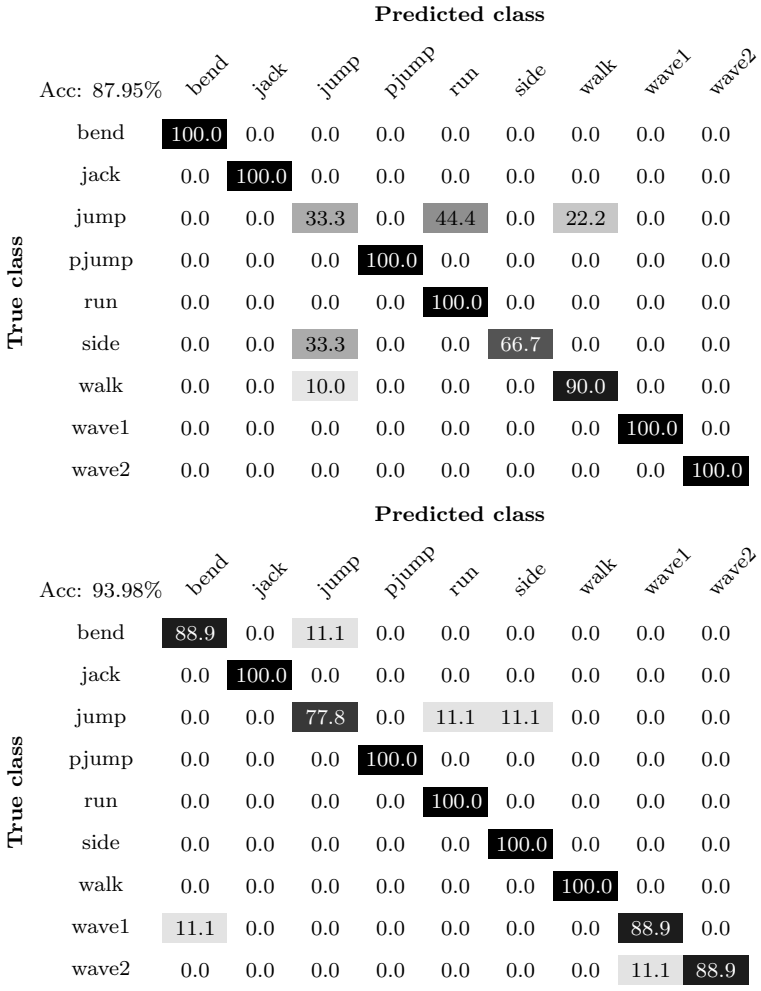


Table 3.1: Confusion matrices for the Weizmann dataset using our baseline with global representation (top) and using our proposed representation (bottom).

3.2.3 Comparison to the state-of-the-art

Table 3.2 shows how well the proposed approach performs in comparison to state-of-the-art. Since there is no standard testing protocol on the Weizmann dataset, researchers have followed their own evaluation

protocol. For instance, Niebles et al. [NL07] use 5 subjects for trainning and the remaining 4 subjects for testing. In the case of Dollar et al. [DRCB05] and Jhuang et al. [JSWP07], 6 random subjects were chosen for training and the other 3 for testing. Finally, in the work of Ali et al. [ABS07] and Schindler et al. [SG08], a leave-one-out protocol was followed. In our experiments, we employ the latter evaluation protocol. We cite the best results achieved for each method. From Table 3.2, its can be shown that our results are situated among the state-of-the-art results.

Method	Accuracy
Niebles [NL07]	72.8%
Dollar [DRCB05]	86.7%
Ali [ABS07]	92.6%
Our Approach	94.0%
Jhuang [JSWP07]	98.8%
Schindler [SG08]	100.0%

Table 3.2: Comparision of action classification results with state-of-the-art methods.

3.3 Action Localization

Given a set of human tracks, our goal is to determine which tracks contain an action and to localize the action within the track. To evaluate the performance of our approach for localizing human actions, we employ two different datasets: the *Coffee & Cigarettes* dataset (section 3.3.1) and the *Hollywood Localization* dataset (section 3.3.2).

The performance in both datasets is evaluated in terms of average precision and precision-recall curves. For each dataset, we evaluate systematically the relevant parameter of our approach. Then, we compare results our results to the state-of-the-art. Note that for all experiments a non-linear SVM with RBF kernel is used. The classifier is trained on the corresponding training part of the given dataset and evaluated on the test sets.

3.3.1 Coffee and Cigarettes

Our first experiment evaluates and compares the different localization strategies (as described in section 2.4): *Peak Tresholding*, *Multilevel Tresholding*, and *Mean-Shift*. In our experiments, we evaluate each localization strategy separately by considering different combinations of *stride* and chunk *length* parameters. In figure 3.7, we plot precision-recall curves for the overall best results per localization strategy. Figure 3.7 (left) illustrates the precision-recall curves for localizing *drinking* actions. It can be seen that the localization strategy based on *Mean-Shift* clustering outperforms the other two clustering approaches by at least in 4%. Figure 3.8 shows the top 12 *drinking* detections sorted by their final *Mean-Shift* scores. In general, we can observe that correct detections are found even by the presence of a wide range of camera viewpoints. At the same time, some of the false positives (e.g. rank 3) include vertical motion of the hand towards the head, but the person does not drink.

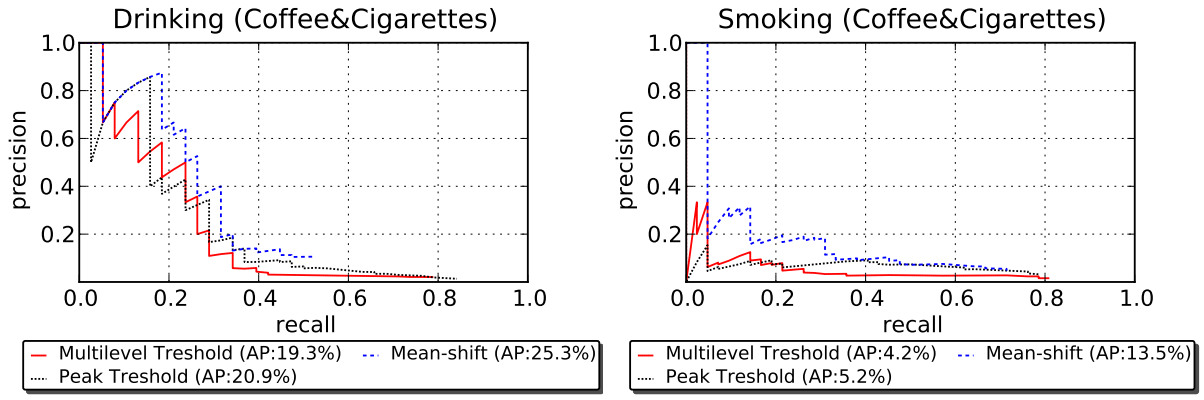
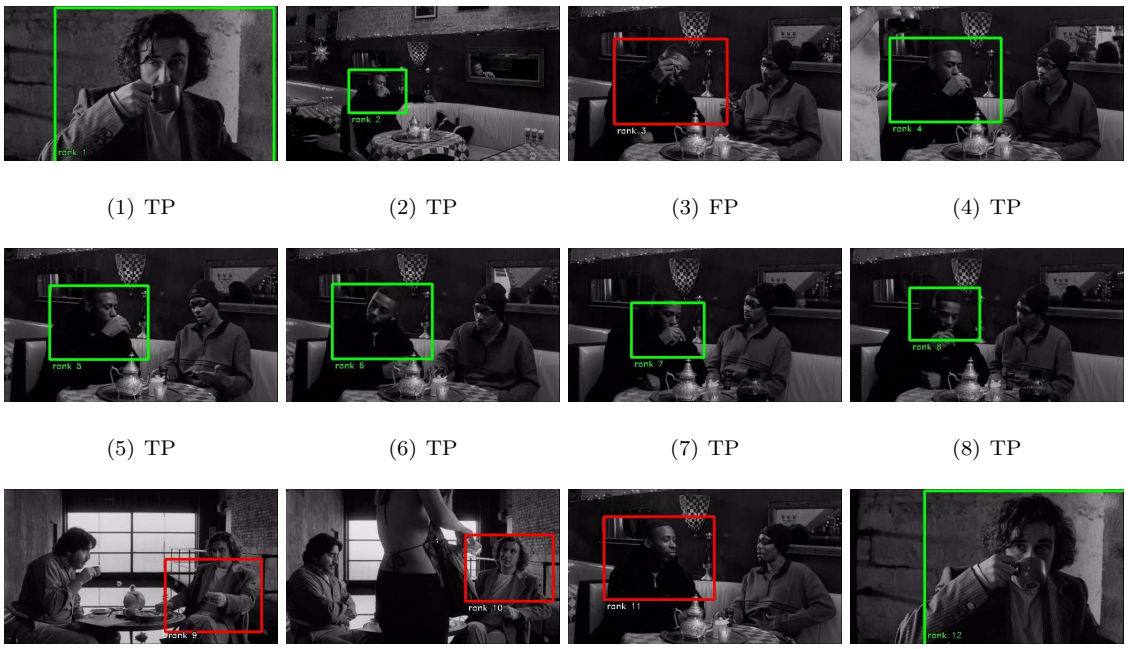


Figure 3.7: Precision-recall curves for our action localization approach on the *C&C* dataset. Human actions evaluated: drinking (left) and smoking (right). We compare our results to previously reported results in the literature.

Figure 3.7 (right) illustrates the precision-recall curves for localizing *smoking* actions. As in the case of *drinking* action detections, the localization with *Mean-Shift* achieves the highest performance and improves over *Multilevel Tresholding* by 9.3% and over *Peak Tresholding* by 8.3%. The top 12 *smoking* detections are depicted in figure 3.9. It is worth to mention that some of the false positives (e.g. rank 3, 6, 7) include vertical motions of hands towards the mouth. These motion patterns are similar to the ones exhibit by *smoking* actions. At the same time, the overall relatively low performance reflects the difficulty of the localization task due to large intra-class variability. In addition, actions like *talking* or even *drinking* can happen in parallel with *smoking* actions.



(1) TP (2) TP (3) FP (4) TP

(5) TP (6) TP (7) TP (8) TP

(9) FP (10) FP (11) FP (12) TP

Figure 3.8: The twelve highest ranked *drinking* detections on *C&C* dataset.

(1) TP (2) TP (3) FP (4) FP

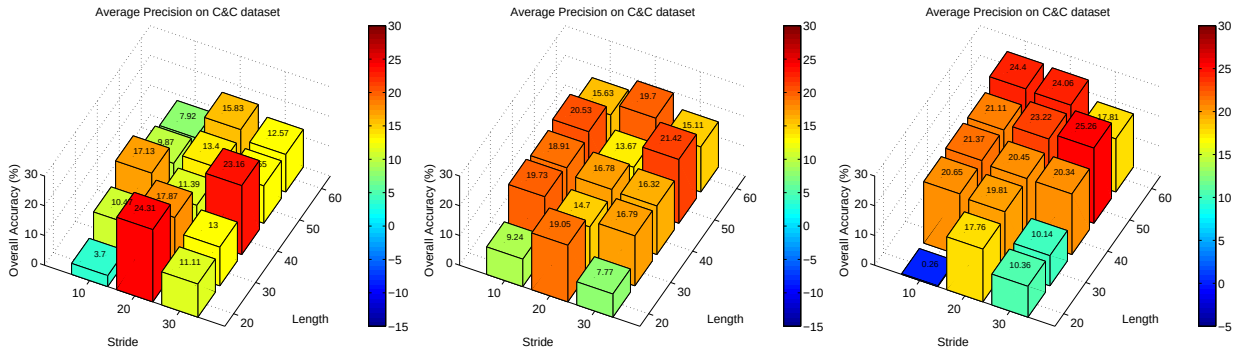
(5) FP (6) FP (7) FP (8) FP

(9) FP (10) FP (11) FP (12) TP

Figure 3.9: The twelve highest ranked *smoking* detections on *C&C* dataset.

Figures 3.10 and 3.11 summarize the performance for different parameter values of *stride*, *chunk length*, and *temporal divisions*. For each parameter combination, we perform a complete evaluation on the *C&C* dataset. The height of each 3D bar indicates the average precision. Figure 3.10 summarizes the results for *drinking* actions. It can be noticed that the performance improves as the number of temporal divisions increases. For instance, the performance of the localization improves significantly from 21.42% to 25.26% when the number of temporal slices is increased from 2 to 3. This can be explained due to additional temporal information that is encoded in the descriptor. Furthermore, larger chunks tend to improve the detection performances regardless of the number of temporal slices considered.

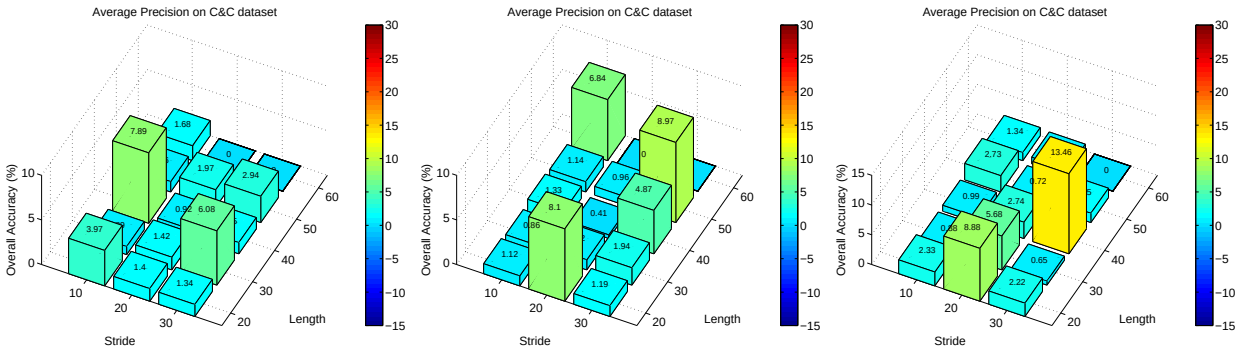
Localization results for the *smoking* action are shown in figure 3.11. As for the *drinking* action, better results are achieved for a larger number of temporal divisions. By increasing the number of temporal slices from 2 to 3, we can obtain a significant improvement of localizing *smoking* actions from 8.1% to 13.46%.



(a) Temporal divisions $n_t = 1$ (b) Temporal divisions $n_t = 2$ (c) Temporal divisions $n_t = 3$

Figure 3.10: Accuracy plots for *drinking* actions on the *C&C* dataset: different *stride* and chunk *length* parameters are considered to represent actions as a loose collection of movement patterns.

Comparison to the state-of-the-art Figure 3.12 shows precision-recall curves to assess the performance of our localization approach. We compare with previously reported results in the literature on the actions *drinking* and *smoking*. Figure 3.12 (left) shows localization results for *drinking* actions. Best results are reported by Kläser et al. [KMSZ10] (54.1%). Willems et al. [WBTG09] obtain in the same setup 45.2%, and Laptev et al. [LP07] 43.4%. It can be seen that for this dataset, our approach is not able to achieve state-of-the-art localization results. This is mainly because actions in this dataset can be appropriately characterized with a global action representation. Therefore there is no gain in modeling actions as a loose collection of movement patterns and a global action representation as suggested by Kläser et al. [KMSZ10]



(a) Temporal divisions $n_t = 1$ (b) Temporal divisions $n_t = 2$ (c) Temporal divisions $n_t = 3$

Figure 3.11: Accuracy plots for *smoking* actions on the *C& C* dataset: different *stride* and chunk *length* parameters are considered to represent actions as a loose collection of movement patterns.

is better suited for this task.

Figure 3.12 (right) shows detection results for localizing *smoking* actions. Since previous approaches in this dataset do not include results for *smoking* actions, we are only able to compare to Klaeser et al. [KMSZ10]. From the results we can notice again that there is no improvement in representing temporally *smoking* actions by a collection of movement patterns.

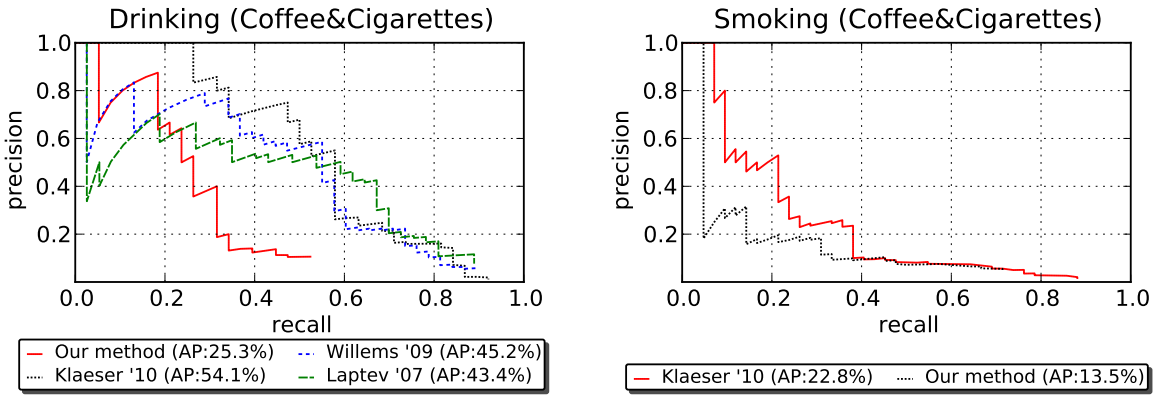


Figure 3.12: Precision-recall curves on the *C&C* test set. Human actions evaluated: drinking (left) and smoking (right). We compare different clustering strategies to generate action hypotheses.

3.3.2 Hollywood Localization

In this subsection, we evaluate how well our localization approach performs on Hollywood movies. First, we fix the number of temporal slices n_t for the HOG-Track descriptor. Then, for each temporal slice, we evaluate different combinations of *stride* and chunk *length* parameters. For each parameter combination, we

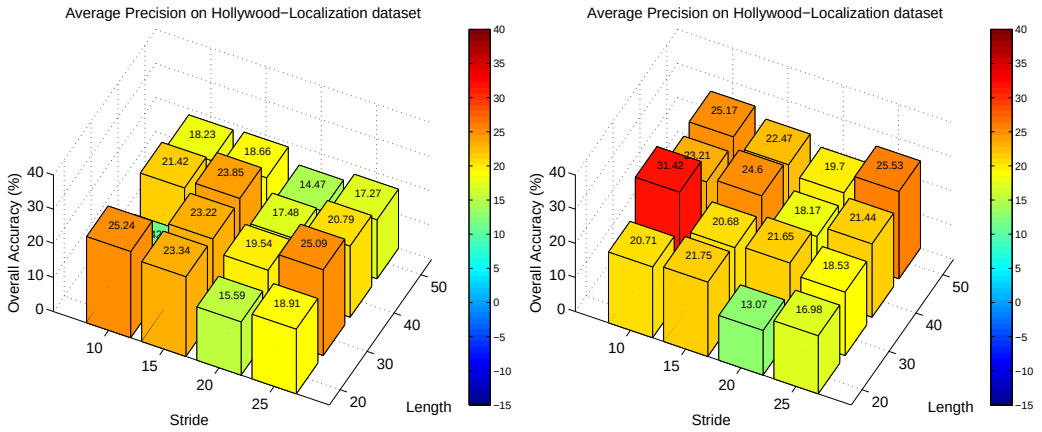
perform a complete evaluation on the Hollywood Localization dataset.

Figure 3.13 summarizes the results for localizing *Answer Phone* actions. We considered 5 different values for temporal slices ranging from $n_t = 1$ to $n_t = 5$. Since results with a $n_t = 5$ temporal slices are stable, we do not consider other values for temporal slices. From the figure, it can be seen that our method performs better the more temporal divisions (n_t) are used.

In general, there is an improvement between 1% and 3% as more of temporal divisions are considered. The largest improvement is achieved between $n_t = 2$ and $n_t = 3$. In this case, a performance of 31.4% is improved by 3% yielding to 34.5%. The gain for larger number of temporal divisions n_t is due to the additional temporal information that is encoded. Furthermore, larger chunk length benefits the localization of *Answering Phone* actions. This observation can be explained from the duration of the action samples in the dataset. In the Hollywood-Localization dataset, the shortest and longest *Answer Phone* actions are 30 and 296 frames long, respectively. From the results, we can see that by using larger chunk lengths of 40 or 50 the performance improves over shorter chunk lengths of 10 or 20 frames. Examples of the top 12 detections for *Answering Phone* are shown in figure 3.14. In general, we can observe that correct detections cover a large variety of poses, scenes, and camera viewpoints.

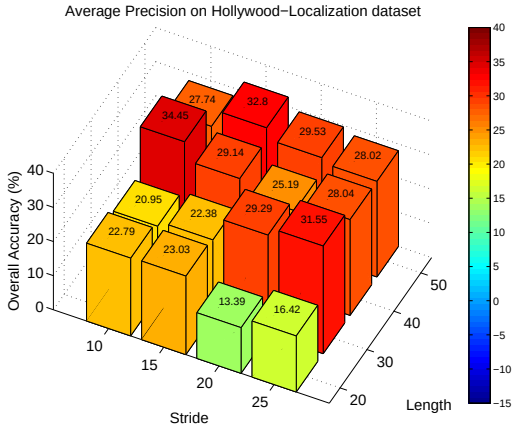
Figure 3.13 summarizes the results for localizing *Stand Up* actions. Similar to *Answer Phone* actions, we can see that the performance is improved as more temporal divisions n_t are used. By adding one temporal division at a time, the performance improves between 2% and 4%. Best results are obtained by setting the number of temporal divisions to $n_t = 4$. The performance obtained is 21.5%. Furthermore, shorter chunk length values improve results for the *Stand Up* action. If we reduce the length of the chunks from 40 to 20 and set $n_t = 3$, the performance increases from 14.2% to 21.5%. The same behavior is valid across the different number of temporal slices considered $n_t = 1 \dots 5$. It can be explained from the duration of *Stand Up* actions samples in the Hollywood Localization dataset. The shortest sequence is 30 frames and the largest 191 frames long. It seems that a representation with shorter chunks is more appropriate for this type of action. Finally, the overall relative low performance reflects the difficulty of localizing *Stand Up* actions. Indeed, other type of actions such as *Shaking Hands* or even *Answering Phone* can happen in parallel, therefore the performance is also compromized. Examples of the top 12 detections for *Stand Up* actions are shown in figure 3.16.

Comparison to the state-of-the-art It is worth to mention that the localization of actions in this dataset is more challenging due to the large variety of video types, which includes: camera ego motion, fast movements, occlusion, etc. In addition, negative samples contain other type of human actions that can share

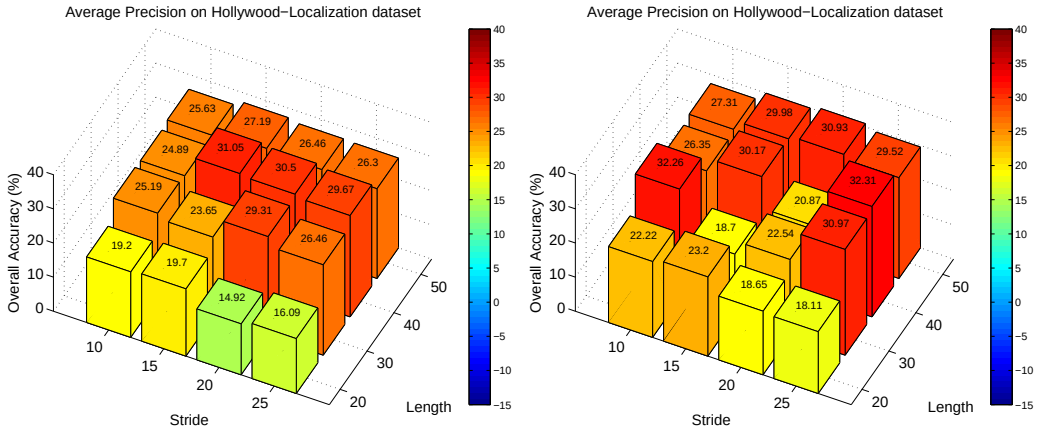


(a) Temporal divisions $n_t = 1$

(b) Temporal divisions $n_t = 2$



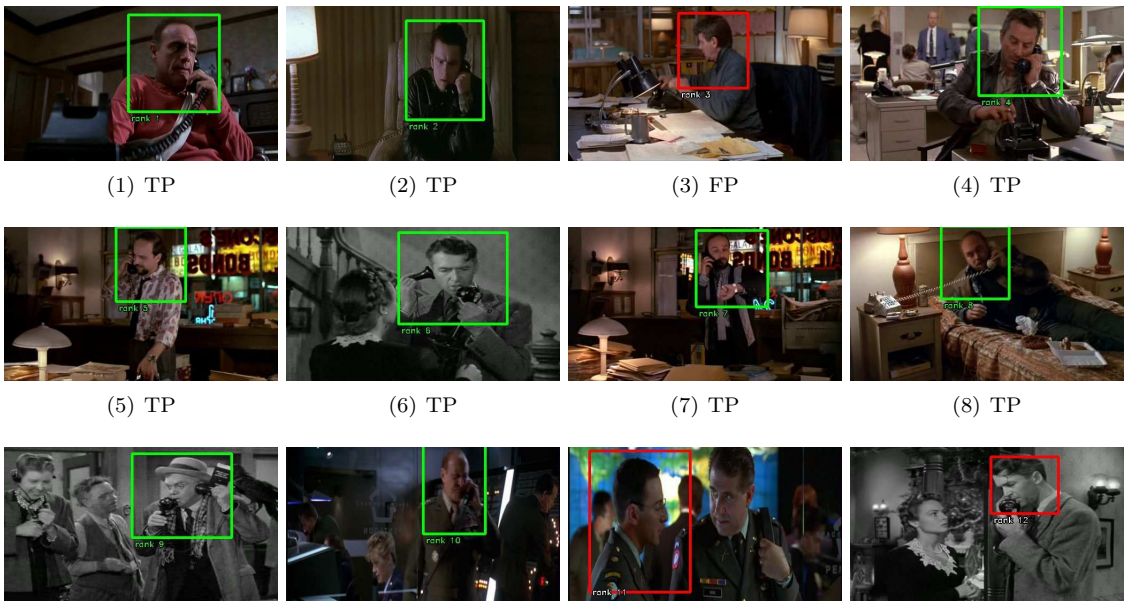
(c) Temporal divisions $n_t = 3$



(d) Temporal divisions $n_t = 4$

(e) Temporal divisions $n_t = 5$

Figure 3.13: Accuracy plots for *answering phone* actions on the Hollywood Localization dataset: different *stride* and chunk *length* parameters are considered to represent actions as a loose collection of movement patterns.



(1) TP (2) TP (3) FP (4) TP

(5) TP (6) TP (7) TP (8) TP

(9) TP (10) TP (11) FP (12) FP

Figure 3.14: The twelve highest ranked *answering phone* detections on *Hollywood-Localization* dataset.

similarities in motion or appearance with the the action of interest.

Figure 3.17 presents localization results for the actions *Answer Phone* and *Stand Up*. For the same experimental setup, results show that our approach compares favorably to the state-of-the-art by temporally representing actions as a loose collection of movement patterns. We observe that actions in this dataset are rather short and well localized in time. Due to this, chunk descriptors are able to capture key movements of these actions.

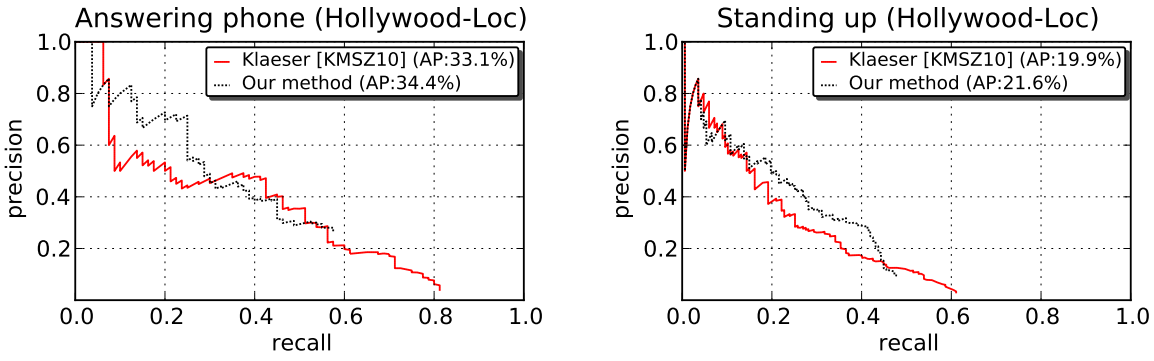
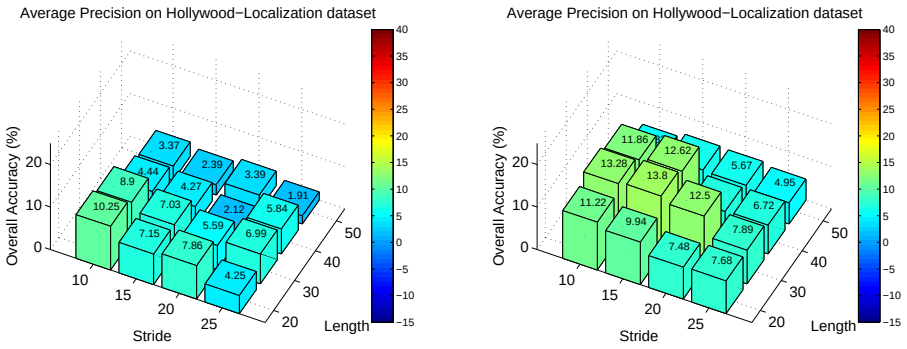
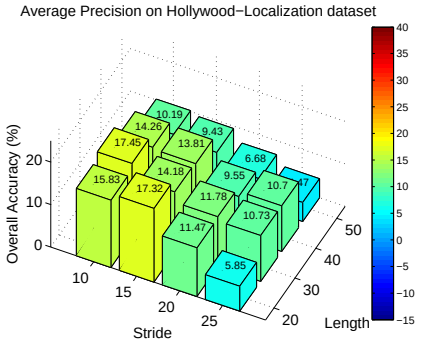


Figure 3.17: Precision-recall curves on the Hollywood-Localization test set. Human actions evaluated: answer phone (left) and stand up (right).

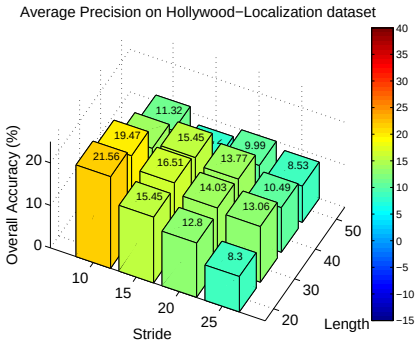


(a) Temporal divisions $n_t = 1$

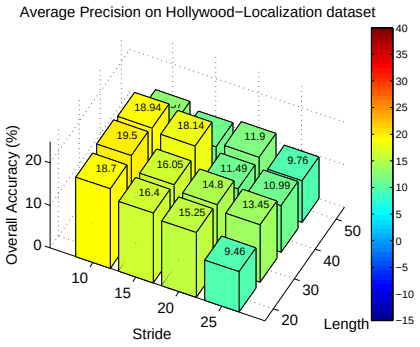
(b) Temporal divisions $n_t = 2$



(c) Temporal divisions $n_t = 3$

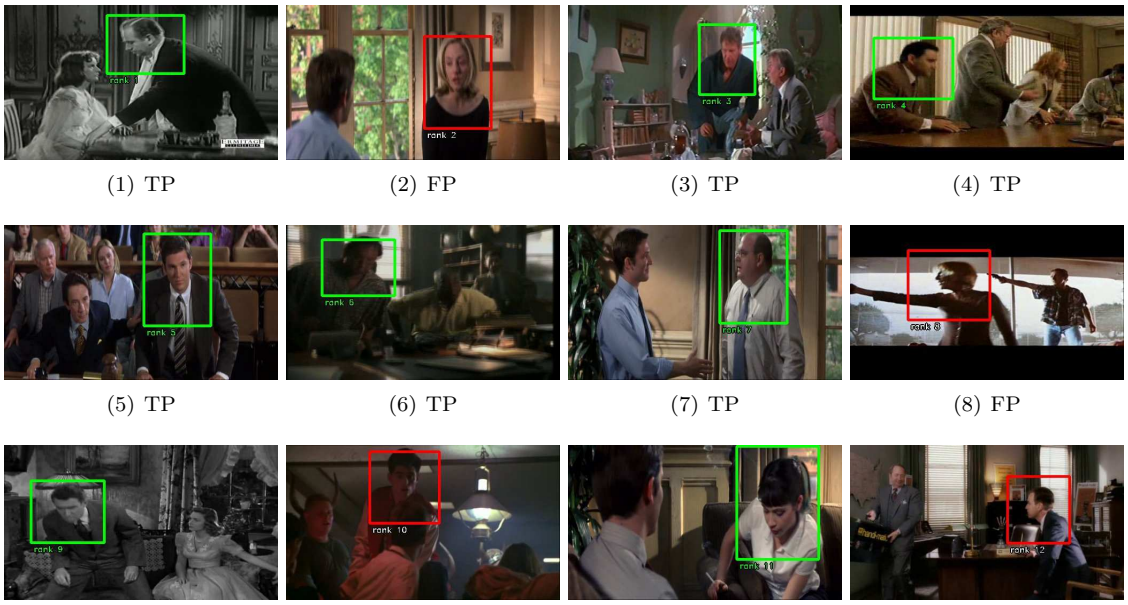


(d) Temporal divisions $n_t = 4$



(e) Temporal divisions $n_t = 5$

Figure 3.15: Accuracy plots for *standing up* actions on the Hollywood Localization dataset: different *stride* and chunk *length* parameters are considered to represent actions as a loose collection of movement patterns.



(1) TP (2) FP (3) TP (4) TP

(5) TP (6) TP (7) TP (8) FP

(9) TP (10) FP (11) TP (12) FP

Figure 3.16: The twelve highest ranked *standing up* detections on *Hollywood-Localization* dataset.

Chapter 4

Conclusions

In this master thesis, we proposed and evaluated an approach to represent actions temporally as a loose collection of movement patterns. Our supervised approach is divided into three main parts:

- First, our system detects and tracks humans throughout time. For human detection, we use the pedestrian detector of Dalal and Triggs [DT05] trained either for upright pedestrians or for upper bodies [KMSZ10]. Human detections are linked together using a tracking-by-detection approach. We interpolate and smooth the final human tracks and employ a classification approach on entire tracks to reduce the number of false positive detections.

- Second, we represent human tracks as a sequence of overlapping track segments, so-called action *chunks*. These are extracted at fixed time steps (*stride*), and are then processed independently. A spatio-temporal descriptor based on histograms of spatio-temporal gradient orientations [KMSZ10] captures appearance and motion information for each chunk.

- Finally, an SVM classifier evaluates the action chunks. The exact localization (i.e., beginning and end) of actions are then determined by a voting approach using a variable band-width mean-shift clustering method.

We evaluated our action represention system on human action classification and localization tasks. In our experiments, we considered different *temporal divisions* for our chunk descriptor in combination with varying values for chunk *length* and *stride*. Overall, we conclude that temporal divisions for the chunk descriptor help to improve action modeling since more temporal information is encoded. At the same time, we found it important to use sufficiently long chunks in combination with rather short stride to allow a large overlap.

In the case of action classification on the Weizmann dataset, we observed shorter chunks (5 frames) in combination with a small stride (3 frames) to work well. The optimal chunk length coincides with the cycle length of the repetitive action classes in this dataset. Similarly to results reported in state-of-the-art [NL07,DRCB05,ABS07], our approach showed promising results for action classification. For action localization, we found our localization strategy employing variable band-width Mean-Shift to work well. On

the Hollywood-Localization data set our approach was able to achieve very good results for both action classes (*answer phone* and *stand up*). This is certainly due to the nature of the action classes, both are rather short actions that are well localized in time and with a rather fixed length among all samples. This is why our loose action model is able to capture the key motion information. In contrast, on the Coffee & Cigarettes dataset, both action classes (*drink* and *smoke*) have a larger variability in their length as well as in the speed at which they are executed. These properties seemed to be more difficult to capture by our model. Therefore results are significantly below the state-of-the-art. Results on realistic movie data show that our approach is suitable for action localization in uncontrolled scenarios. For these, we have shown to compare favorably to state-of-the-art.

References

[ABS07] Saad Ali, Arslan Basharat, and Mubarak Shah. Chaotic invariants for human action recognition. pages 1–8, 2007.

[BAV06] Moeslund Thomas B., Hilton Adrian, and Krüger Volker. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, 2006.

[BETG08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[BGS+05] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1395–1402, Washington, DC, USA, 2005. IEEE Computer Society.

[CDF+04] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Brayy. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[CL00] Stauffer Chris and Grimson W. Eric L. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[CM02] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[DRCB05] Piotr Dollar, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *ICCCN '05: Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 65–72, Washington, DC, USA, 2005. IEEE Computer Society.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 886–893, 2005.

[EBMM03] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 726, Washington, DC, USA, 2003. IEEE Computer Society.

[ESZ09] Mark Everingham, Josef Sivic, and Andrew Zisserman. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5):545–559, 2009.

[FAI+05] David A. Forsyth, Okan Arikan, Leslie Ikemoto, James O'Brien, and Deva Ramanan. Computational studies of human motion: part 1, tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision*, 1(2-3):77–254, 2005.

[FW01] Bobick Aaron F. and Davis James W. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction.* Springer, 2001.

[JSWP07] Hueihan Jhuang, Thomas Serre, Lior Wolf, and Tomaso Poggio. A biologically inspired system for action recognition. In *ICCV '07: Proceedings of the Eleventh IEEE International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil, 2007.

[KMS08] Alexander Klaeser, Marcin Marszalek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *Proceedings of the British Machine Vision Conference*, pages 995–1004, 2008.

[KMSZ10] Alexander Klaeser, Marcin Marszalek, Cordelia Schmid, and Andrew Zisserman. Human focused action localization in video. *International Workshop on Sign, Gesture, and Activity (SGA) in Conjunction with ECCV*, pages xxxx–xxx, 2010.

[LL03] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 432, Washington, DC, USA, 2003. IEEE Computer Society.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LP07] Ivan Laptev and Patrick Pérez. Retrieving actions in movies. In *ICCV '07: Proceedings of the Eleventh IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[NL07] Juan Carlos Niebles and Fei-Fei Li. A hierarchical model of shape and appearance for human action classification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[Pen01] Alex Pentland. Smart clothes, smart rooms. In *WETICE '01: Proceedings of the 10th IEEE International Workshops on Enabling Technologies*, page 3, Washington, DC, USA, 2001. IEEE Computer Society.

[Ron10] Poppe Ronald. A survey on vision-based human action recognition. *Image Vision Computing*, 28(6):976–990, 2010.

[Sal68] G. Salton. *Automatic Information Organization and Retrieval.* McGraw Hill Text, 1968.

[SAS07] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 357–360, New York, NY, USA, 2007. ACM.

[SG08] Konrad Schindler and Luc J. Van Gool. Action snippets: How many frames does human action recognition require? In *CVPR '08: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, USA, 2008.

[SJZ+05] Sarkar Sudeep, Phillips P. Jonathon, Liu Zongyi, Vega Isidro Robledo, Grother Patrick, and Bowyer Kevin W. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[ST94] Jianbo Shi and Carlo Tomasi. Good features to track. In *CVPR'94: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[SWB+07] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

[SZ03] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, pages 127–144, 2003.

[WBTG09] Geert Willems, Jan Hendrik Becker, Tinne Tuytelaars, and Luc Van Gool. Exemplar-based action recognition in videos. In *bmvc*, pages x–x, 2009.

[WTG08] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 650–663, Berlin, Heidelberg, 2008. Springer-Verlag.

[WUK+09] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaeser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proceedings of the British Machine Vision Conference*, pages 127–138, 2009.